Channel Shortening in Wireless Communication

Sha Hu

LUND
UNIVERSITY

Doctoral Thesis
Faculty of Electrical Engineering
Lund, December 2017

Department of Electrical and Information Technology
Faculty of Electrical Engineering
Lund University
P.O. Box 118, SE-221 00, Lund, Sweden

Series of licentiate and doctoral theses
ISBN: 978-91-7753-517-1 (print)
ISBN: 978-91-7753-518-8 (pdf)
ISSN: <1654-790X; No. 110>

Typeset in Frutiger and Adobe Garamond Pro using LaTeX.
Printed in Sweden by Tryckeriet i E-huset, Lund University, Lund.

Dedicated to my parents

"The journey of a thousand miles begins beneath one's feet."

–Lao Zi, *Tao Te Ching*, 604 BC.

Abstract

The concept of Channel Shortening (CS) is a well-known technique that has a rich history over 40 years. CS transfers linear vector channels such as multi-input multi-output (MIMO) and intersymbol interference (ISI) channels into "shortened" versions, for the purpose of reducing demodulation-complexity and improving data-transmission performance. The original CS idea can trace back to the minimum-phase filtering on ISI channels to concentrate channel energy into a first few number of channel taps, and afterwards truncate the remaining channel tails or mitigate the ISI according to the tails with fed back hard-decisions. Since then, many other CS techniques have been extensively developed through various criteria. The CS demodulators investigated in this thesis are based on maximizing the achievable information rate (AIR), which is also referred to as the generalized mutual information (GMI).

This thesis comprises three parts, with AIR-maximization based CS extensively investigated both for reduced-complexity demodulation and precoding designs for wireless communication systems, and then followed by investigations on a newly envisioned system that is beyond traditional massive MIMO. In the first part, the designs of CS demodulators are considered for turbo equalization in linear vector channels with priori informations from outer decoder. Following that, a low-complexity reduced-state soft-output Viterbi equalizer (RS-SOVE) for ISI channels in a non-iterative receiver structure, and an AIR based partial marginalization (AIR-PM) detector for MIMO channels are introduced, respectively. In addition, a novel modulus operation based MIMO detection, namely, the modulus zero-forcing (MZF) detector, is proposed for boosting the detection performance of linear equalizers. In the second part, the CS idea is extended to precoder designs, and a generalized zero-forcing based dirty-paper (GZF-DP) precoder is developed for the broadcast channel (BC). Later, a linear precoder design is considered for MIMO-ISI channel, with priori information that receivers are using CS demodulation. In the last part, a new concept called "Large Intelligent Surface (LIS)" beyond a traditional massive MIMO concept is envisioned and its information-theoretical properties for both data-transmission and terminal-positioning are studied. LIS in its fundamental form uses the entire surface for transmission and reception of radiating signals, which provides ultimate limits

that a traditional large antenna-array system can possible achieve within the same deployed surface-area. In addition, as the effective channel after a matched-filtering (MF) process can be modeled as a sinc-function like linear vector channel, the CS techniques developed in the previous two parts can also be applied to the LIS system.

Popular Science Summary

Wireless communication systems has evolved for several generations, staring from Global System for Mobile communications (GSM), followed by Universal Mobile Telecommunications System (UMTS), and then the fourth generation (4G) Long-Term-Evolution (LTE) systems. Now it evolves toward the fifth generation (5G) with the concept of Internet-of-Things (IoT), device-to-device (D2D), and machine-to-machine (M2M) types of communications. Commercial 5G communication systems are expected to be deployed around year 2020 and featured to provide ubiquitous connectivity for various kinds of devices and applications. 5G enables the IoT, D2D, and M2M communications by providing massive number of connections with stringent energy and transmission constraints, as well as high data-throughput. The great demand for the number of connectivities and the ever increasing data-rate in wireless communication systems impose strong need on improving the spectral-efficiency (bits/s/Hz) and energy-efficiency (bits/J) compared to existing cellular networks. Moreover, in M2M communications, the built-in batteries are featured to last for more than ten years. All these aspects require more efficient physical layer algorithm designs that have better trade-offs between computational-cost and performance.

Massive multi-input multi-output (MIMO), as a promising and potential technology for 5G, can increases the throughput of cells by applying a large number of antenna-elements at the base-transceiver station (BTS). Not only that, when the number of antennas increases, channel vectors according to different antennas can become asymptotically orthogonal, which yields a favorable propagation condition for data transmission and reception. As a result, low-complexity physical layer algorithms such as linear precoders and detectors can perform close to optimal. However, massive MIMO cannot solve all issues. One potential problem is interference-mitigation, such as when users are located at cell-edges, the interference from neighboring cells can degrade the throughputs of them. Even within the same cell, closely located users such as in a stadium or a concert venue may still interfere with each other. Therefore, a receiver design needs to cope with such difficult cases to guarantee a satisfying service under all circumstances for the users. Further, in order to extend coverage and support mobility, a trend in modern wireless communication systems is heterogeneous cell-deployments, where the circuit power in small cells become dominant in addition to

transmission power. In these small cells, traditional MIMO with a few number of antennas may still be deployed to save cost. In all the mentioned scenarios, applying linear algorithms in physical layer design may yield suboptimal data-transmission performances, and better complexity-performance efficient algorithm designs are of interest.

One efficient and promising technique that satisfies such requirements is Channel Shortening (CS), which transfers original linear vector channels (such as MIMO and intersymbol interference (ISI) channels) into shortened versions, to simplify successive processes such as precoding and demodulation. There are many different ways of designing CS transceivers. In the first two parts of the thesis, we investigate an achievable information rate (AIR) based CS, and apply it to demodulator and precoder designs, respectively. The AIR is the highest information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability in a communication channel, which is essentially developed by Claude E. Shannon in 1948. Due to the AIR-maximization used as the optimization target, the obtained CS designs can guarantee that an optimal AIR is attained with the shortened channel. Not only that, another nice property with the AIR-maximization based CS is that, the optimization procedures can be solved in closed-forms with proper designs, and information-theoretical analysis can be carried out.

As massive MIMO is getting mature, new directions for future evolutions is of particular interest. In a third part of this thesis, we envision a new concept called Large Intelligent Surface (LIS), which is an extension of existing massive MIMO systems, but scales beyond a traditional antenna-array concept. In its fundamental form, we assume that a LIS uses its entire surface for transmitting and receiving radiating signals, which is an ultimate limit for what a traditional large antenna-array can possibly achieve within a given surface-area. With LIS, we carry out analysis both for data-transmission and terminal-positioning. We reveal that, the effective channel of the LIS is close to a sinc-like function (a function is central to the field of communications and has been studied at least since H. Nyquist's seminal 1928 paper), and it is efficient in interference suppression. To be specific, if two users are separated half a wavelength apart, there is almost no interference to the other. We also show that, the lower-bounds for positioning decreases not linearly in the surface-area as one may expect, but quadratically and even cubicly, which enlightens the potential of using LIS in future wireless communication systems.

Preface

This thesis summarizes my academic work carried out as a Ph.D. student since Mar. 2015 in the Communication Engineering group at the department of Electrical and Information Technology (EIT), Lund University, Sweden. It comprises two parts. The first part gives an overview of the research fields that I have been working on, while the second part contains the following papers that constitute my work on these fields:

1. S. Hu and F. Rusek, "On the design of channel shortening demodulators for iterative receivers in linear vector channels," submitted to *IEEE Trans. Inf. Theory* on May 2015, in the third round of review, revised on Nov. 2016 and Jun. 2017, respectively.

2. S. Hu, H. Kröll, Q. Huang, and Fredrik Rusek, "Optimal channel shortener design for reduced-state soft-output Viterbi equalizer in single-carrier systems," *IEEE Trans. Commun.,* vol. 65, no. 6, pp. 2568-2582, Jun. 2017.

3. S. Hu and F. Rusek, "A soft-output MIMO detector with achievable information rate based partial marginalization," *IEEE Trans. Signal Process.,* vol. 65, no. 6, pp. 1622-1637, Mar. 2017

4. S. Hu and F. Rusek, "Modulus zero-forcing detection for MIMO channels," submitted to *IEEE Access,* Nov. 2017.

5. S. Hu and F. Rusek, "A generalized zero-forcing precoder with successive dirty-paper coding in MISO broadcast channels," *IEEE Trans. Wireless Commun.,* vol. 16, no. 6, pp. 3632-3645, Jun. 2017

6. S. Hu, X. Gao, and F. Rusek, "Linear precoder design for MIMO-ISI broadcasting channels under channel shortening detection," *IEEE Trans. Signal Process. Lett.,* vol. 23, no. 9, pp. 1207-1211, Jul. 2016

7. S. Hu, F. Rusek, and O. Edfors, "Beyond massive-MIMO: The potential of data-transmission with large intelligent surfaces," submitted to *IEEE Trans. Signal Process.,* Nov. 2017.

8. S. Hu, F. Rusek, and O. Edfors, "Beyond massive-MIMO: The potential of positioning with large intelligent surfaces," accepted in *IEEE Trans. Signal Process.,* Dec. 2017.

As contributions are concerned, in all included papers I have done most of the work (including writings, algorithm implementations and simulations, theoretical analysis and derivations, and manuscript revisions), under the supervisions of my supervisors. All papers are reproduced with permissions of their respective publishers. Publications during my Ph.D. studies that are not included in this thesis are:

9. S. Hu, X. Li, and F. Rusek, "On time-of-arrival estimation in NB-IoT systems," submitted to *IEEE Signal Process. Lett.,* Nov. 2017.
10. S. Hu, K. Chitti, F. Rusek, and O. Edfors, "User assignment with distributed large intelligent surface (LIS) systems," accepted in IEEE Wireless Commun. and Netw. Conf. (WCNC), Barcelona, Spain, Apr. 2018.
11. S. Hu, A. Berg, X. Li, and F. Rusek, "Improving the performance of OTDOA based positioning in NB-IoT system," IEEE Global Commun. Conf. (GLOBECOM), Singapore, Dec. 2017.
12. F. Rusek and S. Hu, "Sequential channel estimation in the presence of random phase noise in NB-IoT systems," IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC), Montreal, Canada, Oct. 2017.
13. S. Hu, F. Rusek, and O. Edfors, "Cramér-Rao lower bounds for positioning with large intelligent surfaces," IEEE Veh. Technol. Conf. (VTC-Fall), Toronto, Canada, Sep. 2017.
14. S. Hu, F. Rusek, and O. Edfors, "A generalized zero-forcing precoder for multiple antenna Gaussian broadcast channels," IEEE Int. Symp. Inf. Theory (ISIT), Aachen, Germany, Jun. 2017, pp. 556-560.
15. S. Hu, F. Rusek, and O. Edfors, "The potential of using large antenna arrays on intelligent surfaces" IEEE Veh. Technol. Conf. (VTC-Spring), Sydney, Australia, Jun. 2017.
16. F. Rusek, K. Chitti, and S. Hu, "Methods and devices in a wireless communication system," PCT/EP2017/050200, patent application by Huawei Technologies Sweden AB, Jan. 2017.
17. S. Hu, F. Rusek, and O. Edfors, "Massive MIMO via cooperative users," Asilomar Conf. Signals, Syst. and Comput. (ACSSC), Pacific Grove, CA, USA, Dec. 2016, pp. 177-182.

18. J. Flordelis, S. Hu, F. Rusek, O Edfors, G. Dahman, X. Gao, and F. Tufvesson, "Exploiting antenna correlation in measured massive MIMO channels," IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC), Valencia, Spain, Sep. 2016.

19. S. Hu and F. Rusek, "Channel shortening algorithms for multiple intersymbol interference channels," IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC), Valencia, Spain, Sep. 2016.

20. S. Hu, H. Kröll, Q. Huang, and Fredrik Rusek, "A low-complexity channel shortening receiver with diversity support for evolved 2G devices," IEEE Int. Conf. Commun. (ICC), Kuala Lumpur, Malaysia, May 2016.

21. S. Hu, F. Rusek, and N. Al-Dhahir, "Comparison of two channel shortening approaches for MIMO-ISI channels," IEEE Wireless Commun. and Netw. Conf. (WCNC), Doha, Qatar, Apr. 2016.

22. S. Hu and F. Rusek, "On the design of reduced state demodulators with interference cancellation for iterative receivers, " IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC), Hong Kong, China, Sep. 2015, pp. 981-985.

Acknowledgments

During this course of Ph.D. study, I have received many helps from others that, without which I could not have come this far, and to whom I would like to take this opportunity to express my sincere thanks.

First and foremost, I thank my supervisor Prof. Fredrik Rusek for his consistent encouragements, inspirations, and guidances through the course of research, without which this thesis would not have been completed. His deep professional knowledge, rigorous scholarship attitude, vivid lecture slides, and easygoing life-style benefit me in a deep manner.

Secondly, I would like to thank my co-supervisor Prof. Ove Edfors for timely supports, especially for many long and constructive discussions in my paper reviews. I also thank the many other professors and lecturers in the department for impressive lectures and enlightening talks from time to time. I owe special thanks to Prof. Qiuting Huang and Dr. Harald Kröll at ETH Zürich, who kindly hosted my visit there during the autumn in 2015. I also thank my co-author Prof. Naofal Al-Dhahir from UT Dallas for cooperation.

Thirdly, I would also like to thank my friends around the corridors for time spent and enjoyed together in group lunches, coffee breaks, and technical discussions. I also thank the administration and IT groups for always being there to help me solving related issues.

Last but not the least, I thank my parents for always trusting in me, and for their selfless dedications to the entire family. I owe special thanks to Qinyan Ye for all supports that she gave as my colleague, friend, and wife. I also thank my little girl Ruizhen for her sweet smiles that always melt my heart, and dedicate this thesis to her one year-old birthday.

Sha Hu

Lund, Dec. 2017.

List of Acronyms and Abbreviations

4G	Fourth Generation
5G	Fifth Generation
AIR	Achievable Information Rate
APP	A Posteriori Probability
AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
BCJR	Bahl, Cocke, Jelinek, and Raviv
BER	Bit Error Rate
BF	Beamforming
BPSK	Binary Phase-Shift Keying
BTS	Base Transceiver Station
CDF	Cumulative Distribution Function
CIR	Channel Impulse Response
CP	Cyclic Prefix
CPL	Central Perpendicular Line
CRLB	Cramér–Rao lower bound
CS	Channel Shortening

CSI	Channel State Information
D2D	Device to Device
DDFSE	Delayed Decision Feedback Sequence Estimation
DL	Downlink
DMT	Diversity-Multiplexing Trade-Off
DoF	Degrees-of-Freedom
DPC	Dirty-Paper Coding
DTFT	Discrete-Time Fourier Transform
FDD	Frequency Division Duplex
FER	Frame Error Rate
FFT	Fast Fourier Transform
FIM	Fisher-Information Matrix
GMI	Generalized Mutual Information
GSM	Global System for Mobile Communications
GZF-DP	Generalized ZF with DPC
IC	Interference Cancellation
IDD	Iterative Detection and Decoding
IF	Integer-Forcing
IoT	Internet of Things
ISI	Intersymbol Interference
LDPC	Low-Density Parity-Check Code
LIS	Large Intelligent Surface
LLL	Lenstra-Lenstra-Lovász

LLR	Log-Likelihood Ratio
LMMSE	Linear MMSE
LMMSE-PIC	Linear MMSE with Parallel IC
LOS	Line-of-Sight
LR	Lattice Reduction
LS	Least Square
LTE	Long Term Evolution
LTE-A	LTE-Advanced
M2M	Machine to Machine
MAC	Multiple Access Channel
MAP	Max A Posteriori
MC	Multi-Carrier
MF	Matched-Filtering
MILB	Mutual Information Lower Bound
MIMO	Multi-Input Multi-Output
MISO	Multi-Input Single-Output
ML	Maximum Likelihood
MLM	ML with Max-Log Approximation
MLSE	ML Sequence Estimation
MMSE	Minimum MSE
MSE	Mean Square Error
MZF	Modulus ZF
NB	Narrowband

OFDM	Orthogonal Frequency Division Multiplexing
PAM	Pulse Amplitude Modulation
PAPR	Peak-to-Average Power Ratio
PDF	Probability Density Function
PM	Partial Marginalization
PR	Partial Response
PSD	Power Spectral Density
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase-Shift Keying
RS-SOVE	Reduced-State SOVE
SC	Single-Carrier
SC-FDMA	Frequency-Division Multiple Access
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
SOVA	Soft-Output Viterbi Algorithm
SOVE	Soft-Output Viterbi Equalizer
TDD	Time Division Duplex
THP	Tomlinson-Harashima Precoding
UL	Uplink
UMTS	Universal Mobile Telecommunications System
VP	Vector Perturbation
ZF	Zero-Forcing

List of Notations

$(\,)^*$	Complex Conjugate
$(\,)^\dagger$	Conjugate Transpose
$(\,)^{\mathrm{T}}$	Matrix Transpose
$(\,)^+$	Pseudo Inverse
$(\,)^{-1}$	Matrix Inverse
$(\,)^{1/2}$	Matrix Square-Root
$(n, K) \ominus \nu$	$\max(n, K-\nu)$
N_0	Noise Density
$[\,]^+$	Nonnegative Protection
$\lfloor\,\rfloor$	Floor Operation
$\mathbb{E}[\,]$	Expectation Operator
$\mathbb{Z}_p$	Ring of Integers modulo p
$\mathbb{Z}$	Integer Set
$\mathcal{A}$	PAM Alphabet
$\mathcal{CN}$	Complex Normal Distribution
$\mathcal{I}\{\,\}$	Taking the Imaginary Part
$\mathcal{Q}_{\mathcal{A}}(\,)$	Entrywise Quantization to Nearest Element in $\mathcal{A}$

$\mathcal{R}\{\,\}$	Taking the Real Part
$\mathcal{X}$	Constellation
$\mathrm{Tr}(\,)$	Trace Operator
argmin	Arguments of the Minima
det	Determinant Operator
exp	Exponential Function
ln	Natural Logarithm Function
log	Logarithm Function
max	Maximal Operation
min	Minimal Operation
vec	Vector Operation
$\mu(\boldsymbol{y}\|\boldsymbol{x})$	Path Metric in BCJR
$\odot$	Hadamard Product
$\otimes$	Kronecker Product
$\propto$	Proportional to
$\star$	Linear Convolution
$\boldsymbol{A}\succ\boldsymbol{B}$	$\boldsymbol{A}-\boldsymbol{B}$ is Positive Definite
$\boldsymbol{A}\succeq\boldsymbol{B}$	$\boldsymbol{A}-\boldsymbol{B}$ is Semi-Positive Definite
$\boldsymbol{I}$	Identity Matrix
$n\boxplus\nu$	$\min(n+\nu,N)$
$n\ominus\nu$	$\max(n-\nu,0)$
$\mathbb{C}^{N\times K}$	$N\times K$ Complex-Valued Matrix
$\mathbb{C}^{N}$	Length-N Complex-Valued Vector

Contents

Part I

Overview of the Research Field and the Thesis's Contributions

Chapter 1

Introduction

"Everything should be made as simple as possible, but not simpler."
– Albert Einstein

1.1 Linear Vector Channels in Wireless Communication

In wireless communication systems, information data-blocks are sent from a base transceiver station (BTS) to either a single or multiple receivers through wireless propagation channels [1–4], which can, in many cases of practical importance and interest, be modeled by linear vector channels as depicted in Fig. 1.1. Traditionally, there are two basic types of linear vector channels, the intersymbol interference (ISI) channel and the multi-input multi-output (MIMO) channel. In some systems, these two different types of channels coexist which yields the MIMO-ISI channel by applying MIMO in combination with single-carrier (SC) modulation. Examples of this include extended coverage Global System for Mobile Communications (GSM) (EC-GSM) aimed for narrowband IoT (NB-IoT) applications [5], and although not a wireless channel, the partial response (PR) channel in the field of multi-track magnetic recording [6]. Other frequently encountered channels are nonlinear, for example the underwater acoustic channel [7] and the satellite-communication channel [8], to name a few. The nonlinearity, which inarguably complicates the mathematical modeling of the channel, often has limited impact on signal processing after a number of countermeasures are taken. These include, e.g., pre-distortion at the transmitter, Volterra series expansion, and linear approximations of various kinds [7, 9]. Therefore, in this thesis the research area is limited to linear vector channels.

The ISI channel is due to the dispersive nature of propagation channels, through which the transmitted signals are scattered and reflected by objects and surroundings in the en-

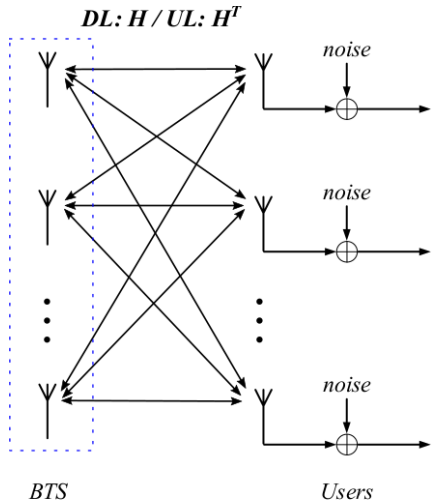


Figure 1.1: A BTS with multiple antennas is communicating to a number of users through DL channel $\boldsymbol{H}$ and UL channel $\boldsymbol{H}^{\mathrm{T}}$ (although the additive noise is only shown for users, it presents at both sides).

vironment. This yields multiple paths arriving at the BTS in the uplink (UL) or the receiver(s) in the downlink (DL) at different time-delays, attenuations, and phases [10]. ISI channels are often encountered in SC modulated systems such as GSM and Universal Mobile Telecommunications System (UMTS) [11], which causes performance degradation of data-transmission when not properly handled. In the fourth generation (4G) long-term evolution (LTE) [12] systems, orthogonal frequency-division multiplexing (OFDM) modulation is used in the DL which inserts a cyclic-prefix (CP) in the transmitted symbols to mitigate ISI, at a cost of a spectral-efficiency loss corresponding to the relative length of the ISI duration compared to the symbol duration. Due to a better peak-to-average power ratio (PAPR), SC frequency-division multiple access (SC-FDMA) is still used in the UL of LTE which leads to ISI.

Different from the ISI channel, a MIMO channel arises from spatial multiplexing techniques, which range from 2×2 MIMO in UMTS to 8×8 in LTE-advanced (LTE-A) systems [13]. In the featured fifth generation (5G) systems [14, 15], it is further increased following the concept of massive MIMO, where hundreds of antennas are deployed at the BTS. Massive MIMO is a key feature for achieving high data-throughputs and providing massive connectivities for various kinds of devices. Time division duplex (TDD) mode based massive MIMO exploits the UL-DL channel reciprocity for DL data-transmission, and relieves the overheads of channel state information (CSI) feedback in UL and acquiring in DL in a frequency division duplex (FDD) mode [16, 17]. There are a number of advantages with massive MIMO systems. Firstly, the spatial multiplexing gain, i.e., the pre-log factor of the channel capacity, linearly increases in the minimum number of transmit and receive antennas. Secondly, the beamforming (BF) efficiency is increased with more antennas deployed which yields highly power-efficient BF and low interference-levels

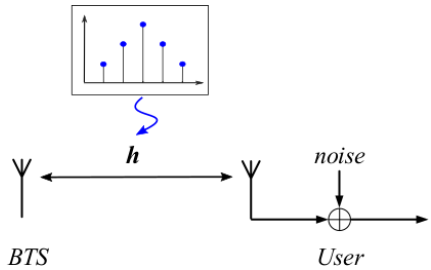


Figure 1.2: A single-antenna BTS is communicating to a single-antenna user through an ISI channel $\boldsymbol{h}$.

at the receiver(s), due to increased capability of spatial focusing of energy. Thirdly, with increased number of antennas, the channel conditions become better, and the channel vectors are asymptotically orthogonal to each other. Such a nice property facilitates close to optimal performance with simple linear signal processing such as matched-filtering (MF), zero-forcing (ZF), and linear minimum-mean-square-error (LMMSE), both for precoding in DL and demodulation in UL.

Mathematically, with K transmit-antennas and N receive-antennas in either DL or UL, a complex-valued $N \times K$ MIMO channel $\boldsymbol{H} \in \mathbb{C}^{N \times K}$ (for a narrowband system without ISI) is modeled as

$$\boldsymbol{H} = \begin{bmatrix} h_{0,0} & h_{0,1} & \cdots & h_{0,K-1} \\ h_{1,0} & h_{1,1} & \cdots & h_{1,K-1} \\ \vdots & \vdots & \vdots & \vdots \\ h_{N-1,0} & h_{N-1,1} & \cdots & h_{N-1,K-1} \end{bmatrix}, \tag{1.1}$$

where each element $h_{n,k}$ corresponds to one transmit and receive antenna pair, and the received signal $\boldsymbol{y} \in \mathbb{C}^N$ equals

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}. \tag{1.2}$$

In this thesis, the transmitted signal $\boldsymbol{x} \in \mathbb{C}^K$ comprises unit average energy information symbols that belong to a constellation $\mathcal{X}$, and $\boldsymbol{n} \sim \mathcal{CN}(\boldsymbol{0}, N_0\boldsymbol{I})$ represents additive white Gaussian noise (AWGN) with zero-mean and a covariance matrix $N_0\boldsymbol{I}$.

With data transmission over a single-input single-output (SISO) dispersive ISI channel with an L-tap channel impulse response (CIR)

$$\boldsymbol{h} = (h_0 \; h_1 \; \ldots \; h_{L-1}), \tag{1.3}$$

such as depicted in Fig. 1.2, the notation is slightly different. The received signal in this case is modeled as

$$\boldsymbol{y} = \boldsymbol{h} \star \boldsymbol{x} + \boldsymbol{n}. \tag{1.4}$$

Encapsulated into a matrix form, (1.4) can be written in the same form as (1.2), but $\boldsymbol{H}$ now is a $N \times K$ Toeplitz matrix representing the linear convolution and generated from $\boldsymbol{h}$ as[1]

$$\boldsymbol{H} = \begin{bmatrix} h_0 & & & \\ h_1 & h_0 & & \\ \vdots & h_1 & \ddots & \\ h_{L-1} & \vdots & \ddots & h_0 \\ & h_{L-1} & \ddots & h_1 \\ & & \ddots & \vdots \\ & & & h_{L-1} \end{bmatrix}, \tag{1.5}$$

where in this case K denotes the size of the transmit-block of a single-antenna transmitter and $N = K+L-1$ is the number of information-carrying observations at a single-antenna receiver, respectively.

Note that, Fig. 1.1 can also represent a MIMO-ISI channel with each link representing an individual ISI channel. That is, with the same received signal model (1.2), each entry $h_{n,k}$ in matrix $\boldsymbol{H}$ in (1.1) now represents a convolution matrix with structure as in (1.5). In what follows, the signal model (1.2) is referred to as an arbitrary $N \times K$ linear vector channel, and when the processing based on (1.2) is tailored for a particular structure of the matrix $\boldsymbol{H}$, such as for MIMO, ISI, and MIMO-ISI channels, it will be explicitly stated.

1.2 Challenges in Designing Physical Layer Algorithms

A typical transmit and receive diagram with linear vector channels applying optional turbo equalization is depicted in Fig. 1.3. The challenges in physical layer algorithm design usually comprise many aspects such as good complexity-performance trade-off, high power and memory efficiency, hardware-friendliness, low processing-latency, small buffer and die-size. Among them, good complexity-performance trade-off is essential and of particular interest, as a high computational cost typically consumes high power by the circuits and introduces a high processing-latency, as well as requiring large buffers. In many algorithm designs, such as precoding and demodulation, the channel $\boldsymbol{H}$ plays a critical role in deciding which algorithm to apply.

In ideal cases where the column or row vectors of $\boldsymbol{H}$ are orthogonal to each other (also known as a favorable propagation condition [16, 17]), linear precoders and demodulators are lossless when applied at both the transmitting and the receiving sides. Under the cases that

[1] The effective channel may look a bit different due to operations such as inserting dummy symbols in GSM or a CP in LTE. For instance, $\boldsymbol{H}$ becomes circular with a CP inserted in LTE.

the vectors of $\boldsymbol{H}$ are correlated, the power-efficiency and performance of linear precoders and demodulators may incur a penalty compared to optimal performance such as with the dirty-paper-coding (DPC) precoder in DL and maximum likelihood (ML) detection in UL, respectively. Even with massive MIMO, there are circumstances where $\boldsymbol{H}$ does not fulfill the favorable propagation condition, even asymptotically, due to a number of reasons.

Firstly, the number of antennas at a BTS cannot grow indefinitely [18] due to the limited physical-size of the transmitting device. When the number of served users is comparable to the number of transmit antennas such as in IoT applications, $\boldsymbol{H}$ can be ill-conditioned. Secondly, for closely located users such as in a stadium or a concert venue, the channel vectors in large antenna-array systems can still be correlated [19]. Thirdly, even though massive MIMO can effectively eliminate intra-cell interferences, there can still be interferers from neighboring cells, especially in dense cellular network deployments [20, 21]. Fourthly, the micro and pico cells in heterogeneous networks may still use traditional MIMO systems that apply only a few number of antennas for cost-savings or due to size-limitations. Lastly, hardware impairments, e.g., radio-frequency (RF) and analog circuits, can introduce unwanted distortions to $\boldsymbol{H}$ that impacts its properties.

For the reasons mentioned above, nonlinear precoder and demodulator designs are still of importance to boost data-transmission with affordable complexity to cope with difficult situations such as mitigating interference, in both traditional and massive MIMO systems. Channel shortening (CS) is one promising technique that can be utilized to design effective precoders and demodulators, which is extensively investigated in this thesis.

1.3 CS Technique and Related Work

Over the past 40 years, substantial amount of literature have been published in developing demodulators that have low complexity yet good performance. Among those works, one promising concept is called "CS" [22–33], which can be traced back to Falconer and Magee in 1973 [22], with the idea to filter the received signal with a prefilter such that the effective channel after filtering has a much shorter duration than the original one. Then, the Viterbi Algorithm (VA) [34, 35] is applied for detection with affordable complexity. Since then, CS demodulators have been developed and optimized from various perspectives such as minimum-phase filtering [36], Kalman-filtering [29], minimum mean-square-error (MMSE) [26], signal-to-noise ratio (SNR) [33, 37], minimum-mean-output-energy (MMOE) [38], and achievable information rate (AIR) [39–41]. In [39] the authors firstly consider the AIR [42–44] that a transceiver system can achieve if a CS demodulator is adopted for ISI channels, and later in [41], the authors extended the concept to any linear vector channel.

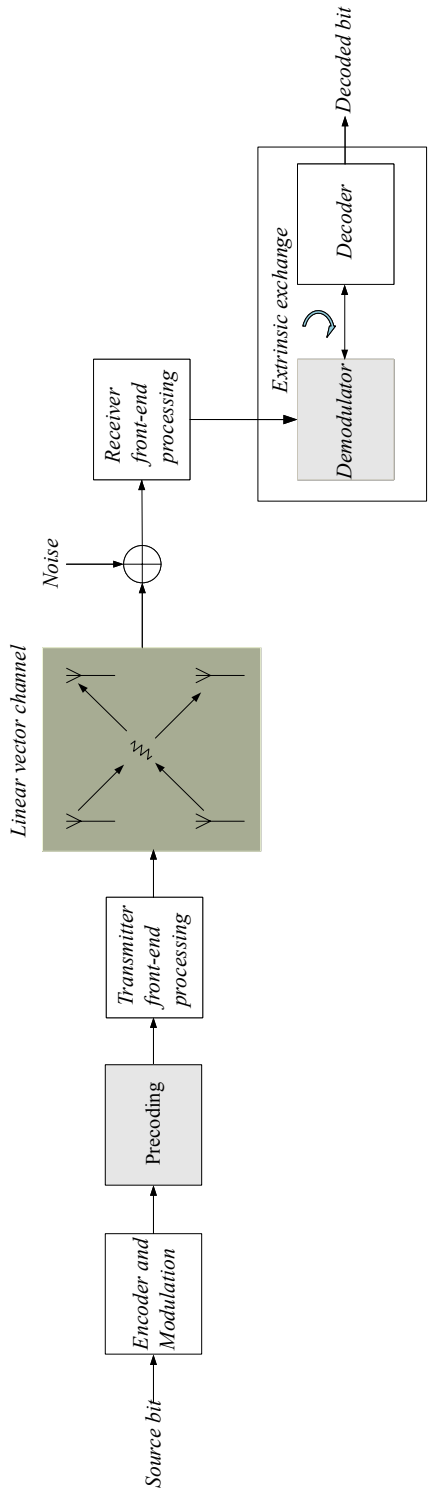
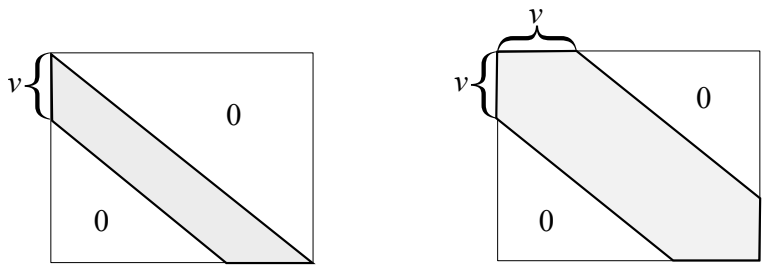


Figure 1.3: A simplified transmit and receive diagram through linear vector channels with iterative detection and decoding (IDD), e.g., turbo equalization. The CS techniques investigated in this thesis are applied to the demodulation (Paper I, II, III, and IV) and precoding (Paper V and VI) designs (gray blocks) under different linear vector channels. A new transmitting and receiving concept is also envisioned (Paper VII and VIII) for possible evolution of wireless communication systems.



Shortened channel matrix F *Shortened gram matrix G*

Figure 1.4: Illustrations of the effective channel matrix $\boldsymbol{F}$ and the gram matrix $\boldsymbol{G}$, respectively. Note that, the main diagonal is not included in the brackets in both cases.

The main idea proposed in [41] is to transfer the linear vector channel $\boldsymbol{H}$ into an effective channel $\boldsymbol{F}$ such that only the main diagonal and the first ν lower (or upper) diagonals can take nonzero elements, followed by a demodulation upon the Ungerboeck detection model [45–48]. Since with the Ungerboeck model, it is not $\boldsymbol{F}$ but $\boldsymbol{G}=\boldsymbol{F}^{\dagger}\boldsymbol{F}$ that matters in the detection process, the constraint is thereby put on $\boldsymbol{G}$ that, it is Hermitian and has a banded shape where only the main diagonal and the first ν lower and upper diagonals can take nonzero elements (a more elaborate discussion is given in Sec. 2.1). Such a generalization by replacing $\boldsymbol{F}^{\dagger}\boldsymbol{F}$ with $\boldsymbol{G}$ is important and meaningful in the sense that, the optimization procedure over $\boldsymbol{F}$ is relaxed to one over $\boldsymbol{G}$ (which is not necessarily positive-definite, but $\boldsymbol{I}+\boldsymbol{G}$ is) and this can be solved for in closed-form [41].

An illustration of $\boldsymbol{F}$ and $\boldsymbol{G}$ is depicted in Fig. 1.4, where the choice of ν in CS demodulator implies a trade-off between demodulation complexity and performance. For MIMO channels (ISI channels are similar), by setting $\nu=0$, $\boldsymbol{F}$ and $\boldsymbol{G}$ are diagonal and the CS demodulator is identical to LMMSE; while by setting $\nu=K-1$, the channel $\boldsymbol{H}$ is not shortened and the CS demodulator is identical to ML. Compared to earlier works, the CS technique in [41] has an advantage that, it is based on the AIR which guarantees that a certain information rate is attained for a given parameter ν in a coded system. In addition, the optimization procedures are in closed-form which allows for an information-theoretical analysis.

1.4 Motivation for CS

Abstractly, the intention of using CS can be interpreted from Fig. 1.5, where we use a sigmoid curve to illustrate a representative relationship between the complexity (in terms of measurable quantities such as processing time, computational cost, memory, etc) and performance (in terms of measurable quantities such as frame-error-rate (FER), bit-error-rate (BER), information-rates, etc). To pursue optimal performance and theoretical limits, the complexity can be prohibitive. What is of major interest from an engineering point of view

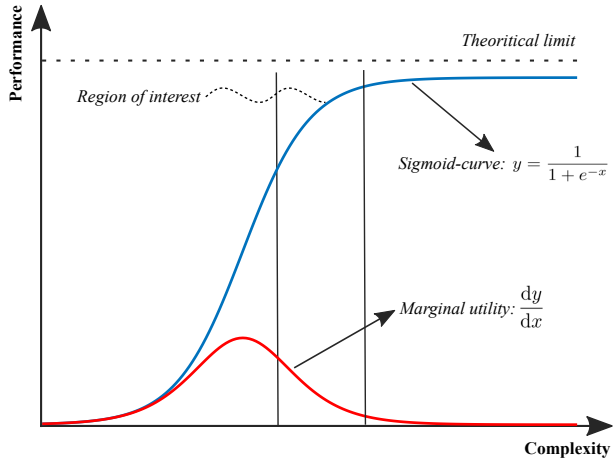


Figure 1.5: An illustration of the complexity-performance relationship in physical algorithm design through an analogous sigmoid curve.

is to optimize the complexity-performance trade-off [49] so that one operates in a region with sufficiently high marginal utility.

A nice property of the considered AIR-maximization based CS technique and what makes it useful is that, a small ν can yield reasonably good performance compared to the optimal algorithms in many cases of interest, as shown in the contributions. In other words, the sigmoid curves which reflect the complexity-performance trade-off in Fig. 1.5 for CS based precoding and demodulation are rather steep. This justifies the effectiveness of using CS in designing demodulator and precoder in wireless communication systems.

1.5 Looking One-Step Ahead

As of today, wireless communication systems have evolved toward the concept of massive MIMO [50]. A question of research interest is: What will be the next big thing? Going one-step beyond massive MIMO, this thesis envisions the concept of Large Intelligent Surface (LIS), which origins from the idea of using the entire deployed large surface for transmitting and receiving radio signals. LIS is a natural evolution of massive MIMO; by (theoretically) packing more and more antenna-elements within a given surface-area, the ultimate outcome is the LIS. Therefore, LIS corresponds to fundamental limits that a traditional large antenna-array system can possibly achieve.

Although LIS shares similarities with massive MIMO, it also brings new features and properties that are not revealed by traditional large antenna-array systems such as the fundamental limits for data-transmission and terminal-positioning with a certain deployed surface-area. Moreover, with LIS we usually consider a large surface system (either centralized or distributed) such that users are in the near-field. Typical examples can be using

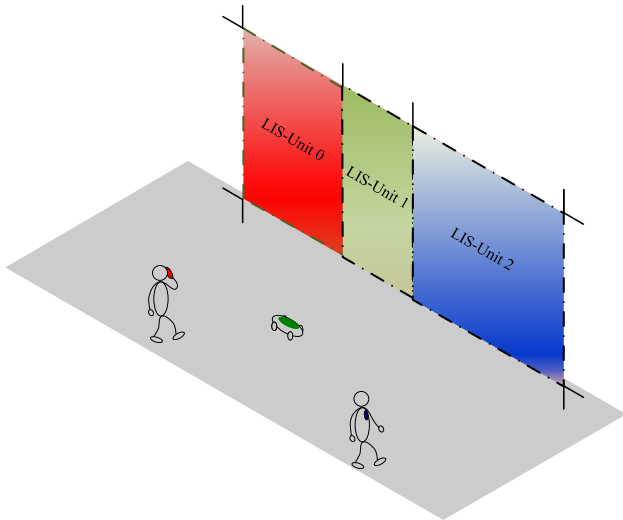


Figure 1.6: An illustration of a typical LIS application, where three devices are communicating to three separate units of a LIS system.

walls in the departure halls at airports or facades of tall buildings for deploying (or being) the LIS, which is illustrated in Fig. 1.6. In addition, artificial-intelligence (AI) technologies, such as machine-learning and convolutional neural networks [51], can be incorporated to the LIS system, with the abundant data and signal dimensions that are provided by the LIS, to make the whole system intelligent. This may facilitate a provision for applications such as terminal-positioning, remote-sensing, motion-detection, load-prediction, wireless-charging, etc.

Chapter 2

Channel Shortening based Demodulator Designs

We start with introducing the optimal maximum *a posteriori* (MAP) detection for MIMO channels in Sec. 2.1, followed by a review of the original AIR based CS demodulator [41] design for non-iterative receivers in Sec. 2.2. Across the iterations in a turbo equalization [52–57] process, the CS design in [41] is static, i.e., the parameters of the CS demodulator are not changing. In Sec. 2.3, we extend the design of the CS demodulator into a dynamic version where the parameters are updated across iterations in turbo equalization. This involves an interference cancellation (IC) step that was not present in [41]. Further, we extensively analyze the CS designs based on three different detection models, which are the main contributions of Paper I [58] in the included papers of this thesis. As turbo equalization suffers from high computational-cost and processing-latency, a new design of CS demodulator in a non-iterative receiver for ISI channels is introduced in Sec. 2.4 and dealt with in detail in Paper II [59]. This new CS demodulator cooperates with hard-decision feedbacks from a reduced-state soft-output equalizer (RS-SOVE). In Sec. 2.5, an AIR-maximization based partial marginalization (AIR-PM) MIMO detector, in Paper III [60], is introduced which has low detection-complexity and processing-latency since all layers can be processed in parallel. Lastly in Sec. 2.6, a modulus ZF (MZF) MIMO detector is also introduced, which is the main contribution of Paper IV [61]. MZF perfectly fits into the framework of CS from a broadened perspective that, CS applies a prefilter that modifies the channel matrix into one that allows for simpler detection.

2.1 Optimal MAP Demodulation

With linear vector channels, the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ according to model (1.2) equals

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(\pi N_0)^N} \exp\left(-\frac{1}{N_0}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2\right). \tag{2.1}$$

Denoting x_n^m as the mth bit of the nth symbol x_n in $\boldsymbol{x}$, and given the observable $\boldsymbol{y}$ and prior distribution $p(\boldsymbol{x})$, the MAP detector generates the posterior probability distribution, commonly in the form of a log-likelihood ratio (LLR), of bit x_n^m as

$$L(x_n^m|\boldsymbol{y}) = \ln\frac{p(x_n^m=1|\boldsymbol{y})}{p(x_n^m=-1|\boldsymbol{y})} = \ln\frac{\sum\limits_{\boldsymbol{x}:x_n^m=1}\exp\left(\mu(\boldsymbol{y}|\boldsymbol{x})+\ln p(\boldsymbol{x})\right)}{\sum\limits_{\boldsymbol{x}:x_n^m=-1}\exp\left(\mu(\boldsymbol{y}|\boldsymbol{x})+\ln p(\boldsymbol{x})\right)}, \tag{2.2}$$

with the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ being

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = -\frac{1}{N_0}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2, \tag{2.3}$$

and the priori probability $p(\boldsymbol{x})$ can be computed from soft information, such as the extrinsic output from the outer decoder. To reduce the computational complexity in (2.2), using Jacobian approximation [62]

$$\ln(e^a+e^b) \approx \max(a,b) \tag{2.4}$$

yields the ML with Max-Log (MLM) approximation of (2.2) as

$$L(x_n^m|\boldsymbol{y}) = \max_{\boldsymbol{x}:x_n^m=1}\left(\mu(\boldsymbol{y}|\boldsymbol{x})+\ln p(\boldsymbol{x})\right) - \max_{\boldsymbol{x}:x_n^m=-1}\left(\mu(\boldsymbol{y}|\boldsymbol{x})+\ln p(\boldsymbol{x})\right). \tag{2.5}$$

The evaluations of (2.2) and (2.5) can be based on the BCJR algorithm [63]. For a MIMO channel, a tree-search process (after a proper QR-decomposition on $\boldsymbol{H}$) is needed which is illustrated with an example in the left part of Fig. 2.1, where the number of states at the last detection-stage equals $|\mathcal{X}|^K$ (assuming $N \geq K$), with $|\mathcal{X}|$ being the cardinality of the constellation which is assumed identical for all layers. For an ISI channel, $\boldsymbol{H}$ has a banded structure and the band-size is limited to the length of the ISI response L, and the number of states at each detection-stage is $|\mathcal{X}|^{L-1}$. When K (in MIMO) and L (in ISI) are relatively large, the computational costs and processing latencies of MAP and MLM can be prohibitive. Therefore, in order to reduce the detection complexity, CS is applied to $\boldsymbol{H}$ to force it to have a banded shape (for an ISI channel, since it is already banded, the purpose is to reduce the band-size). Then, the BCJR is implemented using a simplified trellis-search (instead of tree-search) such as in the right part of Fig. 2.1, and importantly, without compromising much of the performance.

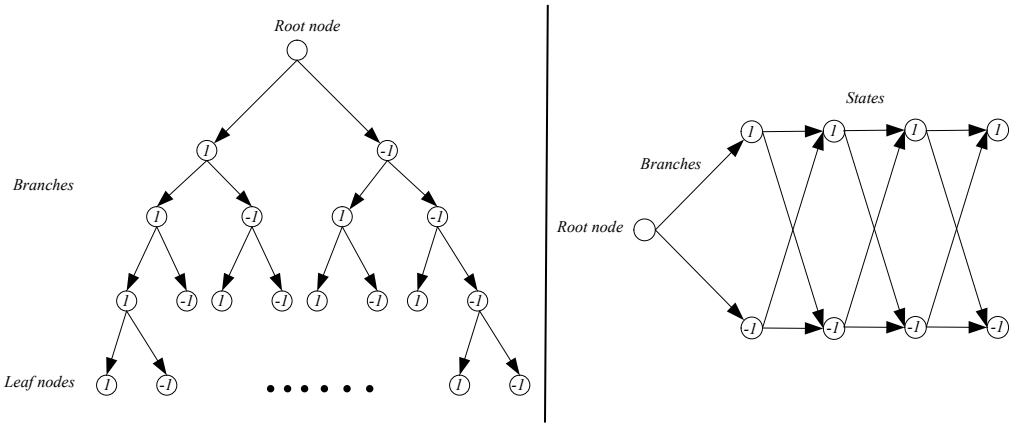


Figure 2.1: A tree-search example (left) for a 4×4 MIMO channel with binary phase shift keying (BPSK) modulation with an optimal demodulation. Applying CS, we can shorten the channel to have nonzero elements only along the main diagonal and the first lower diagonal. With that, the demodulation only requires a trellis-search (right) with 2 states at each stage in the BCJR. Note that, alternatives to CS for detection of MIMO channels via trellis-searches exist. In [64] the authors demonstrate that, there is always a lattice with a trellis-description that is close to the lattice generated by the MIMO channel.

2.2 Previous CS Design without Priori Information

Note that, $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (2.3) can be expressed as

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{N_0}\left(2\mathcal{R}\{\boldsymbol{x}^\dagger \boldsymbol{H}^\dagger \boldsymbol{y}\} - \boldsymbol{x}^\dagger \boldsymbol{H}^\dagger \boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}^\dagger \boldsymbol{y}\right). \tag{2.6}$$

As the last term $\boldsymbol{y}^\dagger \boldsymbol{y}$ is irrelevant for detection, it can be removed in (2.6). Further, by generalizing $\boldsymbol{H}$ and $\boldsymbol{H}^\dagger \boldsymbol{H}$ as $\boldsymbol{H}_{\mathrm{r}}$ and $\boldsymbol{G}$, respectively, we can get

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = 2\mathcal{R}\{\boldsymbol{x}^\dagger \boldsymbol{H}_{\mathrm{r}}^\dagger \boldsymbol{y}\} - \boldsymbol{x}^\dagger \boldsymbol{G}\boldsymbol{x}, \tag{2.7}$$

which corresponds to a mismatched detection model[2]

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\left(2\mathcal{R}\{\boldsymbol{x}^\dagger \boldsymbol{H}_{\mathrm{r}}^\dagger \boldsymbol{y}\} - \boldsymbol{x}^\dagger \boldsymbol{G}\boldsymbol{x}\right). \tag{2.8}$$

Without loss of generality, the noise power N_0 is absorbed into both $\boldsymbol{H}_{\mathrm{r}}$ and $\boldsymbol{G}$. In order to obtain a trellis-diagram in the demodulation [45, 46], constraints are put on $\boldsymbol{G}$ such that, it is Hermitian and only the main diagonal and the first ν upper and lower diagonals can take nonzero values ($\nu \leq K$).

With model (2.8), the AIR is defined as [42, 43]

$$I_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\ln \tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{y})}\left[\ln \tilde{p}(\boldsymbol{y})\right], \tag{2.9}$$

[2]Such a mismatched detection model does not necessarily correspond to a probability-density-function (pdf), but this is irrelevant for detection.

where the expectations are taken over the true pdf $p(\boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y})$, respectively. Following the approach in [41], and under the assumption that $\boldsymbol{x}$ is complex Gaussian distributed, a closed form for $I_{\mathrm{AIR}}(\boldsymbol{y}; \boldsymbol{x})$ can be reached. Optimizing (2.9) over the $N \times K$ prefilter matrix $\boldsymbol{H}_{\mathrm{r}}$ yields

$$\boldsymbol{H}_{\mathrm{r}} = \boldsymbol{W}^{\dagger}(\boldsymbol{I} + \boldsymbol{G}), \tag{2.10}$$

where $\boldsymbol{W}$ is the LMMSE filter[3]

$$\boldsymbol{W} = \boldsymbol{H}^{\dagger}\Big(\boldsymbol{H}\boldsymbol{H}^{\dagger} + N_0\boldsymbol{I}\Big)^{-1}. \tag{2.11}$$

Denote the mean-square-error (MSE) matrix of the LMMSE estimator as

$$\boldsymbol{C} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{H}. \tag{2.12}$$

The resulting $I_{\mathrm{AIR}}(\boldsymbol{y}; \boldsymbol{x})$ can be shown to be [41, 58]

$$I_{\mathrm{AIR}}(\boldsymbol{y}; \boldsymbol{x}) = K + \log\det(\boldsymbol{I} + \boldsymbol{G}) - \mathrm{Tr}\big(\boldsymbol{C}(\boldsymbol{I} + \boldsymbol{G})\big), \tag{2.13}$$

where $\boldsymbol{G}$ is chosen such that $(\boldsymbol{I}+\boldsymbol{G}) \succ \boldsymbol{0}$, and obtained through maximizing (2.13) under the constraints stated earlier. This optimization is treated in [41], and the optimal AIR is

$$I_{\mathrm{AIR}}(\boldsymbol{G}_{\mathrm{opt}}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})\big).$$

For ISI channels, since I_{AIR} is then dependent on the block length K, the asymptotic rate is considered which is defined as

$$\bar{I} = \lim_{K\to\infty} \frac{1}{K} I_{\mathrm{AIR}}. \tag{2.14}$$

As the dimensions of involved matrices are rather large, Szegő's eigenvalue distribution theorem [65, 66] can be applied to simplify the calculations. Moreover, for an ISI channel, the operation $\boldsymbol{H}_{\mathrm{r}}^{\dagger}\boldsymbol{y}$ becomes a filter operation and $\boldsymbol{G}$ has a Toeplitz structure.

2.3 Design of CS Demodulator in Turbo Equalization

Turbo equalization [52–57], is a powerful scheme to improve data detection performance through IDD. Typically, suboptimal demodulators such as dimension-reduction [67], subspace based detections [68], and LMMSE based approaches [53, 56] are used to replace optimal demodulators in turbo equalization for complexity savings. Therefore, it is of interest to evaluate the AIR-maximization based CS demodulator in combination with turbo

[3]As can be seen from here, the AIR-maximization based CS is closely connected to the LMMSE.

equalization, which takes soft information provided by an outer decoder into consideration such that, the parameters of the CS demodulator are designed for a particular level of prior knowledge. There are several possible ways of designing such a CS demodulator and we research three of them that are based on Ungerboeck model, Forney model, and LMMSE with parallel IC (LMMSE-PIC), respectively.

2.3.1 Ungerboeck Model based CS Demodulator

With soft information $\hat{\boldsymbol{x}}$, the CS demodulator based on the Ungerboeck detection model [45] is modified from (2.8) to

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{V}\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{x}})\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big), \tag{2.15}$$

where CS matrices $\boldsymbol{V}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ are denoted as the front-end filter, IC matrix, and trellis-representation matrix, respectively. Following the definition of AIR in (2.9), the AIR can be derived in closed-from, and based on that, the optimal CS parameters which depend on the selected structure of $\boldsymbol{R}$ and the quality of soft information, can be found.

Note that, although other approaches than the model (2.15) would be possible for designing CS with soft information, (2.15) follows a natural extension of the normal CS model (2.8) with the additional information that $\boldsymbol{x}\sim\mathcal{CN}(\hat{\boldsymbol{x}},\boldsymbol{C})$. In this case, the posterior probability $p(\boldsymbol{x}|\boldsymbol{y})$ becomes

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \\ &\propto \exp\left(-\frac{1}{N_0}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2\right)\exp\left(-(\boldsymbol{x}-\hat{\boldsymbol{x}})^\dagger\boldsymbol{C}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{x}})\right) \\ &\propto \exp\left(2\mathcal{R}\left\{\boldsymbol{x}^\dagger\left(\frac{1}{N_0}\boldsymbol{H}^\dagger\boldsymbol{y}+\boldsymbol{C}^{-1}\hat{\boldsymbol{x}}\right)\right\}-\boldsymbol{x}^\dagger\left(\frac{1}{N_0}\boldsymbol{H}^\dagger\boldsymbol{H}+\boldsymbol{C}^{-1}\right)\boldsymbol{x}\right). \end{aligned} \tag{2.16}$$

As can be seen, (2.15) is a generalization of model (2.16), since by substituting

$$\boldsymbol{V}=\frac{1}{N_0}\boldsymbol{H}^\dagger,\ \boldsymbol{R}=-\boldsymbol{C}^{-1},\ \text{and}\ \boldsymbol{G}=\left(\frac{1}{N_0}\boldsymbol{H}^\dagger\boldsymbol{H}+\boldsymbol{C}^{-1}\right) \tag{2.17}$$

into (2.15) can yield (2.16).

2.3.2 Forney Model based CS Demodulator

Instead of using Ungerboeck's model, the Forney detection model [34] can also be used for designing a CS demodulator by identifying

$$\boldsymbol{V}=\boldsymbol{F}^\dagger\boldsymbol{W},\ \boldsymbol{R}=\boldsymbol{F}^\dagger\boldsymbol{T},\ \text{and}\ \boldsymbol{G}=\boldsymbol{F}^\dagger\boldsymbol{F}, \tag{2.18}$$

and then inserting (2.18) into (2.15), which yields a detection model

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(-\|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{T}\hat{\boldsymbol{x}} - \boldsymbol{F}\boldsymbol{x}\|^2\big). \tag{2.19}$$

In this case, $\boldsymbol{F}$ is lower-triangular, where only elements on the main diagonal and the first ν lower diagonals are nonzero. Further, $\boldsymbol{T}$ is constrained to be zero wherever $\boldsymbol{F}$ can take nonzero values. The intention is that, the constraint on $\boldsymbol{F}$ is to shorten the memory for the trellis-search in BCJR, while the constraint on $\boldsymbol{T}$ is to cancel the signal part that $\boldsymbol{F}$ cannot handle with feedbacks.

Similarly as with the Ungerboeck model, the AIR and the CS parameters can be optimized. But unlike the optimization over $\boldsymbol{G}$ which is convex, the optimization over $\boldsymbol{F}$ is not. Moreover, in general the detection performance of the Forney model is inferior to that obtained with the Ungerboeck model, due to less degrees-of-freedom (DoF) in designing the CS parameters. However, the Forney model has the advantage that the branch metric defined in BCJR has a probabilistic meaning, which favors its application in many cases, for example in the CS with a reduced-state equalizer that will be introduced in Sec. 2.4.

2.3.3 LMMSE-PIC based CS Demodulator

Except for the Ungerboeck and Forney detection models, a suboptimal (in the sense of AIR) but simpler CS demodulator, which is closely related to LMMSE-PIC [52, 53, 55], is also proposed in Paper I [58]. This version has, unlike the two versions in Sec. 2.3.1 and 2.3.2, explicit constructions of the CS parameters.

By inserting the optimal $\boldsymbol{V}$ obtained with Ungerboeck model into (2.15) and setting $\boldsymbol{R}=\boldsymbol{0}$, the demodulator actually operates on the mismatched model

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) \;=\; \exp\big(2\mathcal{R}\big\{\boldsymbol{x}^\dagger(\boldsymbol{I}+\boldsymbol{G})\check{\boldsymbol{x}}\big\} - \boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big), \tag{2.20}$$

where

$$\check{\boldsymbol{x}} = \boldsymbol{H}^\dagger(\boldsymbol{H}\boldsymbol{H}^\dagger + N_0\boldsymbol{I})^{-1}\boldsymbol{y}, \tag{2.21}$$

is the LMMSE estimate, which can analogously be replaced by LMMSE-PIC estimates $\tilde{\boldsymbol{x}}$ over iterations in turbo equalization. That is, instead of (2.20) we operate on

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(2\mathcal{R}\Big\{\boldsymbol{x}^\dagger(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{x}}\Big\} - \boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big), \tag{2.22}$$

where only one parameter $\boldsymbol{G}$, which has the same constraints as the first two methods, needs to be optimized. A remark is that, as we prefer to handle the interference through the trellis-search process, the IC should not be present within the memory size ν, which is perfectly aligned with the LMMSE-PIC process.

2.4 CS based RS-SOVE for ISI Channel

Although turbo equalization provides significant gains through IDD, the latency and the computational complexity are too high, which limit its applications in many practical systems. Therefore, it is of interest to develop a similar CS demodulator cooperating with hard feedbacks rather than soft symbols for a non-iterative receiver, in particular for ISI channels.

In [69], Koch and Baier proposed the reduced-state soft output Viterbi equalizer (RS-SOVE) for ISI channels, in which the BCJR trellis only spans the first $\nu+1$ taps, and the signal part corresponding to the tail $L-\nu-1$ channel taps is canceled by a state-dependent decision-feedback mechanism along the detection. The CS demodulator can also be designed to cooperate with RS-SOVE in a similar way, with taking into consideration the quality of the fed back hard symbols when designing the CS parameters. From an information-theoretical perspective, as the traditional RS-SOVE is a special case of the proposed CS based RS-SOVE as shown in Paper II [70], the latter one is superior for data-detection. Such a CS demodulator design also outperforms the original CS [41], as no feedback is utilized in the latter one.

Note that, due to the lack of a probabilistic meaning of the branch metric [47, 71], the Ungerboeck model is not applicable for RS-SOVE and the CS demodulator design uses the Forney model in (2.19), where $\boldsymbol{W}$, $\boldsymbol{T}$ and $\boldsymbol{F}$ are $K\times K$ convolution matrices generated from filters $\boldsymbol{w}$, $\boldsymbol{t}$ and $\boldsymbol{f}$, respectively, and $\hat{\boldsymbol{x}}$ is the fed-back hard symbols. There is no constraint on $\boldsymbol{w}$, while $\boldsymbol{f}$ and $\boldsymbol{t}$ have the below shapes,

$$\boldsymbol{f} = \big(f_0, f_1, \cdots, f_\nu\big), \tag{2.23}$$

$$\boldsymbol{t} = \big(\underbrace{0, \cdots, 0}_{\nu+1}, t_0, t_1, \cdots, t_{L-\nu-2}\big). \tag{2.24}$$

Based on Szegő's theorem [66], the AIR corresponding to (2.19) equals [70]

$$\begin{aligned}\bar{I} &= \lim_{K\to\infty}\frac{1}{K} I_{\mathrm{AIR}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F}) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log\big(1+|F(\omega)|^2\big)-|F(\omega)|^2-\frac{L(\omega)}{1+|F(\omega)|^2}\right)\mathrm{d}\omega \\ &\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)\big(W(\omega)H(\omega)-\sigma T(\omega)\big)\big\}\mathrm{d}\omega,\end{aligned} \tag{2.25}$$

where $W(\omega)$, $T(\omega)$, and $F(\omega)$ are the discrete time Fourier transforms (DTFT) of $\boldsymbol{w}$, $\boldsymbol{t}$ and $\boldsymbol{f}$, respectively, and

$$\begin{aligned}L(\omega) = {} & |F(\omega)W(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\sigma|F(\omega)T(\omega)|^2 \\ & -2\sigma|F(\omega)|^2\mathcal{R}\{H(\omega)W(\omega)T^*(\omega)\}\,.\end{aligned}$$

The same optimization procedure can be taken to optimize the CS filters as with turbo equalization. However, one difference is that, with RS-SOVE the quality of feedback, reflected by the parameter σ, is unknown and an estimate of it has to be used for designing the optimal CS parameters.

2.5 AIR-PM Detection for MIMO Channel

Among many different kinds of MIMO detectors, one interesting approach to reduce complexity is via partial marginalization (PM) [72, 73], which carefully selects ν parent layers out of K layers, and marginalizes over the remaining $K-\nu$ child layers using ZF with decision feedback (ZF-DF) estimates. With PM, $\boldsymbol{x}$ (possibly after reordering) is split into two parts as

$$\begin{aligned}\boldsymbol{x}_a &= (x_{K-\nu+1}, x_{K-\nu+2}, \cdots, x_K)\,,\\ \boldsymbol{x}_b &= (x_1, x_2, \cdots, x_{K-\nu})\,.\end{aligned}$$

The posterior density is marginalized exactly over $\boldsymbol{x}_a$, while the density is marginalized approximately over $\boldsymbol{x}_b$. The metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (2.3) can be equivalently written as

$$\mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a) = -\frac{1}{N_0}\|\boldsymbol{y} - \boldsymbol{H}_b\boldsymbol{x}_b - \boldsymbol{H}_a\boldsymbol{x}_a\|^2, \tag{2.26}$$

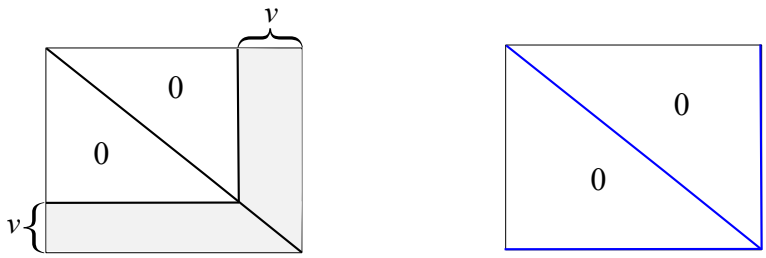
where $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$ are the sub-channels corresponding to the signals parts $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, respectively. With marginalization over $\boldsymbol{x}_b$, the LLRs in (2.2) corresponding to a bit x_n^m in $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are approximated [72] as

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}_a: x_n^m=1} \exp\left(\max\limits_{\boldsymbol{x}_b} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}{\sum\limits_{\boldsymbol{x}_a: x_n^m=-1} \exp\left(\max\limits_{\boldsymbol{x}_b} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}, \tag{2.27}$$

and

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}_a} \exp\left(\max\limits_{\boldsymbol{x}_b: x_n^m=1} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}{\sum\limits_{\boldsymbol{x}_a} \exp\left(\max\limits_{\boldsymbol{x}_b: x_n^m=-1} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}, \tag{2.28}$$

respectively. To further reduce the complexity and improve the performance, it is of interest to improve the ZF-DF process in PM through AIR-maximization based CS. The proposed AIR-PM detector in Paper III [60] provides a simpler and fully parallel hardware structure



The shape of matrix G in AIR-PM *A special case with v=1*

Figure 2.2: Design of $\boldsymbol{G}$ with the AIR-PM detector, where the gray parts in the left figure corresponds to the signal part $\boldsymbol{x}_a$, and the right figure is a special case of $\nu = 1$ when $\boldsymbol{x}_a$ only contains a single signal.

both for the parent and the child layers. With AIR-PM, which is a special case of CS, the shape of the matrix $\boldsymbol{G}$ changes to a special shape for the purpose of PM as depicted in Fig. 2.2; compare to the CS in Fig. 1.4.

With detection model (2.8) and the shape of $\boldsymbol{G}$ in the left part of Fig. 2.2, the metric in (2.26) can be rewritten as

$$\mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a) = \sum_{n=1}^{K-\nu} \mu_1^n \left(\tilde{y}_n | x_n, \boldsymbol{x}_a\right) + \sum_{n=K-\nu+1}^{K} \mu_2^n \left(\tilde{y}_n | x_n, x_{n+1}, \cdots, x_K\right), \quad (2.29)$$

where $\mu_1^n \left(\tilde{y}_n | x_n, \boldsymbol{x}_a\right)$ and $\mu_2^n \left(\tilde{y}_n | x_n, x_{n+1}, \cdots, x_K\right)$ are defined as

$$\mu_1^n \left(\tilde{y}_n | x_n, \boldsymbol{x}_a\right) = 2\mathcal{R}\left\{\left(\tilde{y}_n - \sum_{k=K-\nu+1}^{K} g_{n,k} x_k\right) x_n^*\right\} - g_{n,n}|x_n|^2, \quad (2.30)$$

$$\mu_2^n \left(\tilde{y}_n | x_n, x_{n+1}, \cdots, x_K\right) = 2\mathcal{R}\left\{\left(\tilde{y}_n - \sum_{k=n+1}^{K} g_{n,k} x_k\right) x_n^*\right\} - g_{n,n}|x_n|^2. \quad (2.31)$$

Inserting (2.29) back into the LLR calculations in (2.27) and (2.28), the processes are greatly simplified with the AIR-PM detector. As from (2.30), under each assumption of parent layers $\boldsymbol{x}_a$, the optimization over each child layer x_n is separate. Further, as $\mu_1^n \left(\tilde{y}_n | x_n, \boldsymbol{x}_a\right)$ is quadratic in x_n, a least-square (LS) estimate can be easily obtained. Moreover, utilizing (2.29) it can be shown that the chain rule of AIR holds as

$$I_{\text{AIR}}\left(\boldsymbol{y}; \boldsymbol{x}\right) = I\left(\boldsymbol{y}; \boldsymbol{x}_a\right) + \sum_{n=1}^{K-\nu} I\left(\boldsymbol{y}; x_n | \boldsymbol{x}_a\right), \quad (2.32)$$

which facilities an information-theoretical analysis.

Another favorable property with AIR-PM is that, the orderings inside parent and child layers have no impact on the AIR (which is not the case for the original PM due to ZF-DF processes), which saves complexity required for finding optimal orderings. However, selecting ν best parent layers that maximizes the AIR is still needed.

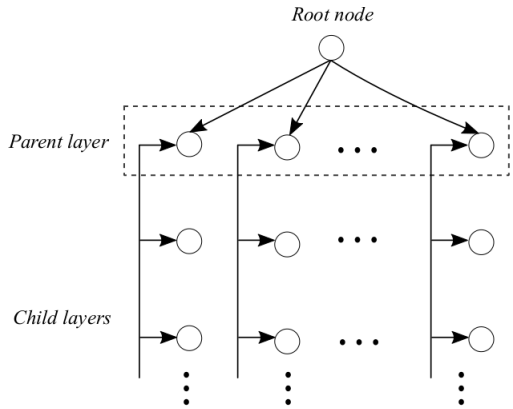


Figure 2.3: A simplified trellis-search with AIR-PM detector when selecting only a single parent layer, and all child layers are detected in parallel.

2.6 Modulus ZF based MIMO Detection

Tomlinson-Harashima precoding (THP) [74] and vector perturbation (VP) [75, 76] are well acknowledged techniques for approximating DPC through nonlinear modulus arithmetic operations. Using the underlying ideas of them, in Paper IV [61], we propose a modulus based zero-forcing (MZF) detector for MIMO channels, which falls in the framework of CS by filtering the channel into a simpler version that allows for mitigating the interference with modulus operation.

Using modulus operation for MIMO detection has been proposed earlier as integer-forcing (IF) receivers [77–79]. In [77], the authors show that transmitted messages can be recovered by decoding integer combinations of codewords according to an effective channel matrix comprising integer-valued entries. Further, IF significantly outperforms conventional linear architectures such as ZF and LMMSE and attains the optimal diversity-multiplexing trade-off (DMT) in the high SNR regime. Despite promising theoretical evidences, the IF receiver in [77] require each transmit-antenna to employ the same lattice code [80], which is challenging for higher-order modulations. A simpler IF receiver using binary linear codes (such as turbo codes and low-density parity-check code (LDPC) codes) is proposed in [78, 79]. Although the encoding/decoding process is simplified, the IF designs in [78, 79] still need to detect the linear combinations of different codewords first, followed by a matrix inversion process to recover the codeword on each transmit-antenna.

In our proposed MZF detector, the integer matrix is optimized according to each specific modulation-order such that, the symbol detection is separate and in parallel for all layers, and no encoding/decoding process is required for each transmit-antenna. Moreover, there is no need to invert an integer-valued effective matrix over a finite-field.

Without loss of generality, the matrix $\boldsymbol{H}$ can be assumed to be a square matrix, and with

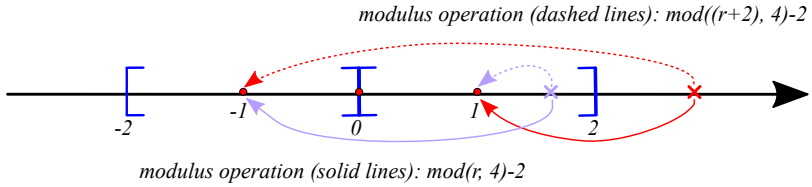


Figure 2.4: The decision regions and modulus operations with MZF detection for a 2-PAM (BPSK) modulation.

the following definitions,

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \mathcal{R}\{\boldsymbol{y}\} \\ \mathcal{I}\{\boldsymbol{y}\} \end{bmatrix}, \ \tilde{\boldsymbol{x}} = \begin{bmatrix} \mathcal{R}\{\boldsymbol{x}\} \\ \mathcal{I}\{\boldsymbol{x}\} \end{bmatrix}, \ \tilde{\boldsymbol{n}} = \begin{bmatrix} \mathcal{R}\{\boldsymbol{n}\} \\ \mathcal{I}\{\boldsymbol{n}\} \end{bmatrix}, \ \tilde{\boldsymbol{H}} = \begin{bmatrix} \mathcal{R}\{\boldsymbol{H}\} & -\mathcal{I}\{\boldsymbol{H}\} \\ \mathcal{I}\{\boldsymbol{H}\} & \mathcal{R}\{\boldsymbol{H}\} \end{bmatrix}, \tag{2.33}$$

we can rewrite (1.2) as a real-valued model

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{n}}, \tag{2.34}$$

where the $K \times K$ channel matrix[4] $\tilde{\boldsymbol{H}}$ is known, and $\tilde{\boldsymbol{x}} = [x_1 \ \dots \ x_K]^{\mathrm{T}}$ contains pulse-amplitude-modulation (PAM) symbols from an alphabet $\mathcal{A} = \{\pm 1, \pm 3, \dots, \pm(\sqrt{M} - 1)\}$ (that is, complex-valued $\boldsymbol{x}$ is M-QAM modulated), and $\tilde{\boldsymbol{n}}$ is random Gaussian noise with a covariance matrix $(N_0/2)\boldsymbol{I}$.

A traditional ZF detector is given by

$$\hat{x}_k = \mathcal{Q}_{\mathcal{A}}(r_k), \quad 1 \le k \le K, \tag{2.35}$$

where $\mathcal{Q}_{\mathcal{A}}(\cdot)$ denotes entrywise quantization to the nearest point in $\mathcal{A}$, and

$$r_k = \boldsymbol{\delta}_k \tilde{\boldsymbol{H}}^{+} \boldsymbol{y}, \tag{2.36}$$

with $\tilde{\boldsymbol{H}}^{+}$ being the pseudo-inverse[5] of $\tilde{\boldsymbol{H}}$ and the vector

$$\boldsymbol{\delta}_k = [\underbrace{0 \ \dots \ 0}_{k-1} \ 1 \ \underbrace{0 \ \dots \ 0}_{K-k}]. \tag{2.37}$$

Since the ZF nulls out all interferences from other layers, it suffers from noise-enhancement and we improve it by replacing (2.36) with

$$r_k = (\tau \boldsymbol{\delta}_k + \boldsymbol{q}_k)\tilde{\boldsymbol{H}}^{+} \boldsymbol{y}, \tag{2.38}$$

where $\boldsymbol{q}_k = [q_{k1}, \ q_{k2}, \ \dots \ q_{kK}]$ and $q_{k\ell} \in 2\mathbb{Z}$, i.e., the even integers, and τ is set to

$$\tau = 2^{(1-\log_2 \sqrt{M})}, \tag{2.39}$$

[4]Note that, although for simplicity we still use K to represent the dimensions of $\tilde{\boldsymbol{H}}$, its dimensions should be twice those of the complex-valued $\boldsymbol{H}$.

[5]In fact, $\tilde{\boldsymbol{H}}^{+}$ can be replaced by other linear equalizers and the remaining processes remain similar.

Then, for each layer

$$r_k = \tau x_k + \sum_{\ell=1}^{K} q_{k\ell} x_\ell + w_k, \tag{2.40}$$

and x_k can be recovered with a modulus operation (that depends on a property of $\boldsymbol{q}_k$) with one example shown in Fig. 2.4. The target of designing $\boldsymbol{q_k}$ is to optimize the post-processing SNR as

$$\boldsymbol{q}_k^{\text{opt}} = \arg\min_{\boldsymbol{q_k}} \|(\tau \boldsymbol{\delta}_k + \boldsymbol{q}_k)\tilde{\boldsymbol{H}}^{+}\|^2, \tag{2.41}$$

which is an instance of sphere detection with integers [81], but lattice-reduction (LR) based suboptimal approaches can also be used. Note that, here we optimize $\boldsymbol{q_k}$ for each layer separately, whereas the IF receiver designs in [77, 78] optimize all $\boldsymbol{q_k}$ jointly since they require the integer-valued effective channel to be full-rank over the reals.

Chapter 3

Channel Shortening based Precoder Designs

The same idea, using CS to strike a complexity-performance trade-off in demodulation, can also be applied in precoder design, and we consider two different cases. The first case is to apply CS in a nonlinear precoder design in combination with successive DPC [82] for multi-input single-output (MISO) BCs. The traditional DPC and its suboptimal variant, the ZF based DPC (ZF-DP) precoder, are briefly introduced in Sec. 3.1. In Sec. 3.2, a generalized ZF based DPC (GZF-DP) precoder is proposed to generalize the ZF-DP precoder for complexity savings. Such a precoder design is optimized via two different approaches[6], maximizing the sum-rate, and maximizing the minimum user-rate, respectively, which are the main contributions of Paper V [83]. In Sec. 3.3, a linear precoder design is optimized by taking into account that the receivers are using CS demodulators for MIMO-ISI channels, which is the main contribution of Paper VI [84].

3.1 Traditional ZF-DP Precoder Designs

Gelfand and Pinsker [85] derived the capacity of a single-user memoryless channel with an additive interference signal known to the transmitter but not the receiver, and show that the channel capacity $\mathcal{C}$ is the same as if the interference is not present. Based on that, Costa proposed the DPC [82] precoder which achieves the capacity. Practical DPC designs based on finite-alphabets have been extensively developed and are based on techniques such as

[6]When applying the CS in precoder designs, the AIR is considered as the sum-rate or the user-rate that can be achieved (with the optimal or the CS receivers) under a total transmit-power constraint and with an effective channel matrix (after precoding) $\boldsymbol{F}$, which has the same banded-shape as in the design of CS demodulators.

THP, lattice precoding [86], and trellis coded quantization and modulation [87, 88].

Consider a MISO-BC with an N-antenna transmitter and K single-antenna users and under the assumption $N \geq K$. Denote by x_n the DPC-encoded symbol of the nth user that cancels the non-causal interference from the other users, and y_n, z_n are the received sample and the noise term corresponding to the nth user, respectively. With an $N \times K$ precoding matrix $\boldsymbol{P}$ applied at the transmitter, the received signals at the K autonomous users can be written as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{z}, \tag{3.1}$$

where $K \times N$ matrix $\boldsymbol{H}$ represents the MISO-BC channel, and the noise term $\boldsymbol{z}$ comprises complex Gaussian variables with zero-mean and a covariance matrix $N_0\boldsymbol{I}$. The transmit symbols x_n are uncorrelated due to DPC encoding and have unit-transmit power, i.e., $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\dagger] = \boldsymbol{I}$. In addition, the transmit node is subject to a total transmit-power constraint P_{T} such that

$$\mathrm{Tr}\left(\boldsymbol{P}\boldsymbol{P}^\dagger\right) \leq P_{\mathrm{T}}. \tag{3.2}$$

Denote the $K \times K$ effective channel $\boldsymbol{F} = \boldsymbol{H}\boldsymbol{P}$, the interference channel corresponding to each of the K users from (3.1) can be written as

$$y_n = f_{n,n}x_n + \sum_{k=1}^{n-1} f_{n,k}x_k + \sum_{k=n+1}^{K} f_{n,k}x_k + z_n. \tag{3.3}$$

With a successive DPC [82] encoding scheme, the non-causal interference is canceled, while the causal interference is regarded as additive noise.

Caire and Shamai in [89] propose a ZF-DP design for MISO-BC, which is suboptimal but only exacts a small penalty compared to the optimal DPC. Assuming the channel decomposition $\boldsymbol{H} = \boldsymbol{R}\boldsymbol{U}$, where $\boldsymbol{R}$ is a $K \times K$ lower-triangular matrix and $\boldsymbol{U}$ is a $K \times N$ unitary matrix, the ZF-DP precoder is set to $\boldsymbol{P} = \boldsymbol{U}^\dagger\boldsymbol{D}$, where the $K \times K$ diagonal matrix $\boldsymbol{D}$ represents the power allocation. The effective channel with the ZF-DP precoder equals $\boldsymbol{F} = \boldsymbol{R}\boldsymbol{D}$, and the received sample y_n reads

$$y_n = f_{n,n}x_n + \sum_{k=1}^{n-1} f_{n,k}x_k + z_n. \tag{3.4}$$

Through successive DPC encoding, the non-casual interference is nulled out for each user, and the sum-rate maximization problem degrades to an optimal power allocation problem that can be solved through standard water-filling over $\boldsymbol{D}$.

3.2 GZF-DP Precoder Design

To generalize the ZF-DP precoder, we assume that the effective channel $\boldsymbol{F}$ is a band-shaped and lower-triangular matrix (the same shape as for CS demodulation) which yields the GZF-DP precoder design, and the received sample y_n of the nth user reads

$$y_n = f_{n,n}x_n + \sum_{k=\max(n-\nu,1)}^{n-1} f_{n,k}x_k + z_n, \tag{3.5}$$

where ν denotes the interfering depth of the effective channel $\boldsymbol{F}$. Under the case $\nu = 0$, the GZF-DP precoder degrades to the linear ZF precoder and no DPC is needed, while with $\nu = K-1$, the GZF-DP precoder is identical to the ZF-DP. As the interference is non-causally known at the transmit node, we can apply the same successive DPC encoding as the ZF-DP precoder to cancel it. For each of the K users, as there are at most ν users to be considered in the DPC and $\nu \ll K-1$, the GZF-DP precoder renders much lower complexity of the successive DPC operations than the ZF-DP.

The remaining problem is now to design the matrix $\boldsymbol{F}$, which we deal with in two different ways in Paper v [83], outlined in the following two subsections.

3.2.1 Sum-rate Maximization

Denoting $\boldsymbol{\Lambda} = (\boldsymbol{H}\boldsymbol{H}^{\dagger})^{-1}$, and for GZF-DP precoder with a fixed ν, the problem for the sum-rate maximization subject to the transmit-power constraint can be formulated as

$$\begin{aligned} &\underset{\boldsymbol{F}}{\text{maximize}} \quad \sum_{n=1}^{K} \log\left(1 + \frac{|f_{n,n}|^2}{N_0}\right) \\ &\text{subject to} \quad \mathrm{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{\Lambda}\boldsymbol{F}\right) \leq P_{\mathrm{T}}. \end{aligned} \tag{3.6}$$

As can be shown, the GZF-DP precoder actually sacrifices the user-rate of the last ν users to increase the sum-rate. Since ZF-DP is a special case of GZF-DP with $\nu = K-1$, the user-rates of the last users are sacrificed to maximize the sum-rate with the ZF-DP precoder. Hence, it is of interest to consider the minimum user-rate maximization with the proposed GZF-DP precoder for improving the information rate for each of the users.

3.2.2 Minimum User-rate Maximization

To maximize the minimum user-rate R, the design of GZF-DP precoder is formulated as

$$\begin{aligned} &\underset{\boldsymbol{F},\, R}{\text{maximize}} \;\; R \\ &\text{subject to} \;\; R \leq \log\left(1+\frac{|f_{n,n}|^2}{N_0}\right), 1 \leq n \leq K, \\ &\qquad\qquad \mathrm{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{\Lambda}\boldsymbol{F}\right) \leq P_{\mathrm{T}}. \end{aligned} \tag{3.7}$$

Again, it is implicit in (3.7) that a given ν is used. With GZF-DP, both optimizations in (3.6) and (3.7) can be solved in closed-forms, and we can guarantee a certain sum-rate and a minimum user-rate to all users, respectively, while at the same time reducing the interference depth from $K-1$ to ν, thereby lowering the complexity of the DPC encoding/decoding procedure. Moreover, with a small ν, the performance of GZF-DP is close to that of the ZF-DP as shown in Paper v [83].

Note that, the GZF-DP precoder can also be extended to MIMO-BC in similar ways, which we have dealt with in [90].

3.3 Linear Precoder Design for MIMO-ISI Channels

Conventionally, precoder design at the BTS assumes that either a linear or the ML detection is utilized at the receivers. This is because the precoder optimization for any other type of receiver operation is complicated [91, 92]. However, in practice the actual receiver operation typically falls between these two extremes. Within a CS framework, it is possible to optimize the precoder for a given level of receiver complexity and this is what we deal with in Paper vi [84]. We provide a brief overview next. With the priori information that CS demodulators are applied at receivers, the linear precoding can be optimized accordingly to improve its performance.

Consider a MIMO-ISI effective channel

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}_{1,1} & \boldsymbol{F}_{1,2} & \cdots & \boldsymbol{F}_{1,K} \\ \boldsymbol{F}_{2,1} & \boldsymbol{F}_{2,2} & \cdots & \boldsymbol{F}_{2,K} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{F}_{K,1} & \boldsymbol{F}_{K,2} & \cdots & \boldsymbol{F}_{K,K} \end{bmatrix}, \tag{3.8}$$

where each block $\boldsymbol{F}_{n,k}$ is an $N \times N$ circular convolution matrix generated from an effective

ISI channel after precoding. The received signal vector of the nth user reads

$$\boldsymbol{y}_n = \boldsymbol{F}_{n,n}\boldsymbol{d}_n + \sum_{k=1,k\neq n}^{K} \boldsymbol{F}_{n,k}\boldsymbol{d}_k + \boldsymbol{w}_n, \tag{3.9}$$

where $\boldsymbol{d}_n$ and $\boldsymbol{w}_n$ are the transmitted symbols and received noise vectors, respectively. The CS target on model (3.9), similar to (2.8), is to maximize the AIR with the detection model

$$\tilde{p}(\boldsymbol{y}_n|\boldsymbol{d}_n) = \exp\Big(-2\mathcal{R}\left\{\boldsymbol{d}_n^{\dagger}\boldsymbol{V}_n\boldsymbol{y}_n\right\} + \boldsymbol{d}_n^{\dagger}\boldsymbol{G}_n\boldsymbol{d}_n\Big), \tag{3.10}$$

where the circular convolution matrix $\boldsymbol{V}_n$ performs a filtering of the received samples, and $\boldsymbol{G}_n$ is a Hermitian Toeplitz matrix with only the middle $2\nu+1$ diagonals allowed to be nonzero. The detection of $\boldsymbol{d}_n$ can then be carried out over a trellis with $|\mathcal{X}|^{\nu}$ states.

In order to optimize $(\boldsymbol{V}_n, \boldsymbol{G}_n)$, the AIR corresponding to detection model (3.10) for each user is adopted as the objective function. With optimal $(\boldsymbol{V}_n, \boldsymbol{G}_n)$ in [41, Proposition 2], the AIR equals

$$I_n = -\log\det\left(\boldsymbol{B}_n^{\nu}\right), \tag{3.11}$$

where the $(\nu+1)\times(\nu+1)$ Hermitian matrix $\boldsymbol{B}_n^{\nu}$ is the principal submatrix formed by any contiguous $(\nu+1)$ rows and the corresponding columns of the MSE matrix

$$\boldsymbol{B}_n = \boldsymbol{I} - \boldsymbol{F}_{n,n}^{\dagger}\left(\sum_{k=1}^{K}\boldsymbol{F}_{n,k}\boldsymbol{F}_{n,k}^{\dagger} + N_0\boldsymbol{I}\right)^{-1}\boldsymbol{F}_{n,n}. \tag{3.12}$$

Moreover, the total transmit-power P_{T} equals

$$\begin{aligned} P_{\mathrm{T}} &= \frac{1}{N}\mathrm{Tr}\left\{\boldsymbol{F}^{\dagger}\boldsymbol{\Lambda}\boldsymbol{F}\right\} \\ &\approx \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{n=1}^{K}\sum_{k=1}^{K}\left(\sum_{t=1}^{K}F_{n,t}(\omega)F_{k,t}^{*}(\omega)\right)\Lambda_{k,n}(\omega)\mathrm{d}\omega, \end{aligned} \tag{3.13}$$

where the approximation (3.13) holds when N is sufficiently large, and $F_{k,n}(\omega)$ and $\Lambda_{k,n}(\omega)$ are the DTFTs of the (k,n)th block entry in $\boldsymbol{F}$ and $\boldsymbol{\Lambda} = (\boldsymbol{H}\boldsymbol{H}^{\dagger})^{-1}$, respectively. Then, the sum-AIR optimization is formulated as

$$\begin{aligned} &\max_{\left\{F_{n,k}(\omega)\right\}} \sum_{n=1}^{K} I_n \\ &\text{subject to } P_{\mathrm{T}} \leq 1, \end{aligned} \tag{3.14}$$

according to which the optimal set of $\{F_{n,k}(\omega)\}$ for $1\leq n,k\leq K$ can be found.

Chapter 4

Beyond Massive-MIMO: Large Intelligent Surface

A Large Intelligent Surface (LIS) is a new concept in wireless communication, envisioned in Paper VII [93] and VIII [94]. LIS can be seen as an extension of traditional massive MIMO [14, 15, 50] systems, however, it scales up beyond the traditional antenna-array concept, and the whole contiguous surface is used as transmitting and receiving device. LIS makes new and disruptive applications that require high energy-efficiency and transmission-reliability, low-latency, and ability to interact with the environment, possible such as in the IoT systems [95–97]. In Sec. 4.1, we briefly introduce the narrowband received signal model of the LIS under perfect line-of-sight (LOS) condition. Based on the received signal model, the capabilities of data-transmission with LIS is summarized in Sec. 4.2, which are the main contributions of Paper VII [93]. Since the effective channel with LIS after MF, mathematically, is a linear vector channel, the CS demodulators can therefore be applied, which is introduced in Sec. 4.3. In Sec. 4.4, the fundamental limits of positioning a device with LIS are summarized, which are the main contributions of Paper VIII [94].

4.1 Narrowband Signal Model with Perfect LOS

Consider a transmission from K autonomous single-antenna terminals located in a three-dimensional space to a two-dimensional LIS deployed on a plane as shown in Fig. 1.6. Expressed in Cartesian coordinates, the LIS center is located at $x = y = z = 0$, while terminals are located at $z > 0$ and arbitrary x, y coordinates. For analytical tractability, we assume perfect LOS propagation. The kth terminal located at (x_k, y_k, z_k) transmits data symbols u_k with power P_k, and all u_k are assumed to be independent Gaussian variables

with zero-mean and unit-variance.

Denote λ as the wavelength and N_0 as the spatial power spectral density (PSD) of AWGN. Considering a narrowband system, the received signal after the MF can be modeled as (with details in Paper VII [93]),

$$\boldsymbol{r} = \boldsymbol{G}\boldsymbol{u} + \boldsymbol{w}, \tag{4.1}$$

where the (ℓ, k)th element of matrix $\boldsymbol{G}$ equals

$$g_{k,\ell} = \sqrt{P_\ell P_k}\phi_{k,\ell}, \tag{4.2}$$

and $\boldsymbol{w}$ is the effective discrete noise after MF with zero-mean, and colored according to

$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}] = N_0\boldsymbol{G}. \tag{4.3}$$

Further, we have that

$$\phi_{k,\ell} = \iint\limits_{(x,y)\in\mathcal{S}} s_{x_\ell,y_\ell,z_\ell}(x,y)s^*_{x_k,y_k,z_k}(x,y)\mathrm{d}x\mathrm{d}y, \tag{4.4}$$

with $\mathcal{S}$ being the surface-area spanned by the two-dimensional LIS, and $s_{x_k,\,y_k,\,z_k}(x,y)$ is the effective channel at the LIS location $(x,y,0)$ corresponding to the kth user as

$$s_{x_k,\,y_k,\,z_k}(x,y) = \frac{\sqrt{z_k}}{2\sqrt{\pi}\eta_k^{3/4}}\exp\left(-\frac{2\pi j\sqrt{\eta_k}}{\lambda}\right), \tag{4.5}$$

where $\eta_k = (x-x_k)^2+(y-y_k)^2+(z-z_k)^2$.

4.2 Data-Transmission Capabilities with LIS

With the received signal model (4.1), the channel capacity $\mathcal{C}$ [98, 99] averaged by the number of terminals, in nats/s/Hz, equals

$$\mathcal{C} = \frac{1}{K}\log\left(\boldsymbol{I} + \frac{\boldsymbol{G}}{N_0}\right). \tag{4.6}$$

In Paper VIII [94], a special interest is put on the number of independent signal dimensions per deployed area-unit of the LIS that is possible to harvest, and is derived based on the capacity normalized with the total deployed surface-area; a quantity we refer to as $\hat{\mathcal{C}}$ with unit [nats/s/Hz/area-unit]. It is shown that, the limit of $\hat{\mathcal{C}}$, achieved when wavelength λ approaches zero, is $\hat{P}/(2N_0)$ [nats/s/Hz/volume-unit], where $\hat{P}$ is the transmit-power per

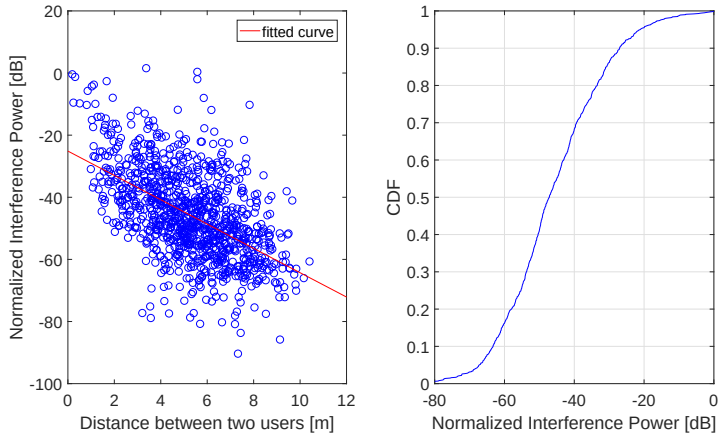


Figure 4.1: The interference powers (normalized by the signal powers) are measured for two users in front of the LIS.

volume-unit of the terminal-deployment. In particular, we show that for an infinitely large LIS, $2/\lambda$ terminals can be spatially multiplexed per meter (m) deployed surface for one-dimensional terminal-deployment, and π/λ^2 terminals can be spatially multiplexed per m^2 deployed surface-area for two and three dimensional terminal-deployments, respectively.

In Fig. 4.1, the interference power for a square LIS with length[7] 1 and centered at $x=y=z=0$ is measured. The wavelength is $\lambda=0.125$ (corresponding to a carrier-frequency 2.4 GHz) and two users that are located in front of the LIS, with coordinates $-4\leq x,y\leq 4$ and $0<z\leq 8$ for both users. The empirical cumulative density function (CDF) is measured for 1000 realizations of random uniform user locations. As can be seen, in most cases the interference power normalized by the received signal power is below -20 dB, which shows that the interference from the other user is significantly suppressed [100].

4.3 Utilizing CS Demodulation in LIS

Since the received signal model (4.1) represents a linear vector channel model, CS demodulators can be directly applied, while noticing that the noise in this case is colored. As explained earlier, CS demodulators can provide a complexity-performance trade-off between LMMSE and the optimal ML demodulator via selecting different interference depth ν. A nice property of a LIS system is that, even the LMMSE demodulator performs close to the optimal MAP, due to the strong interference-suppressing properties of the LIS illustrated in Sec. 4.2. However, as can be seen from Fig. 4.2, there is still a gap between LMMSE and MAP, which quickly reduces as ν increases.

The sum-rate is evaluated with the same LIS as in Fig. 4.2, but now the LIS is deployed

[7] In this thesis, the unit of all lengths is meter (m).

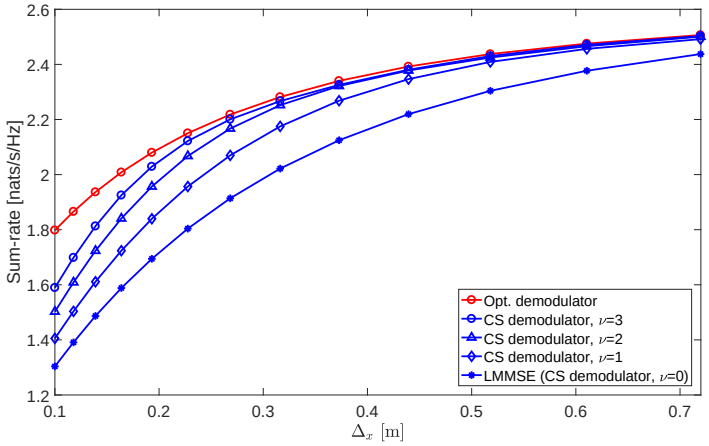


Figure 4.2: The sum-rate achieved with CS demodulators with a considered LIS system.

on the roof of a room of length 8, width 8, and height 4, and with $N_0 = 1$, $\lambda = 0.5$ and $\hat{P} = 10$ dB. The channel capacity is compared to those achieved with the CS demodulators where 15 terminals are located on the ground in a line and with spacing Δ_x, while the LIS is deployed in the center of the roof. As can be seen in Fig. 4.2, when Δ_x is larger than $\lambda/2 = 0.25$, the CS demodulator with $\nu = 1$ converges fast to the channel capacity, but has much less demodulation complexity.

4.4 CRLB for Positioning with LIS

With the effective channel model (4.5) with a LIS, the fundamental limits of terminal-positioning can also be analyzed. Denote the first-order derivatives of $s_{x_0,y_0,z_0}(x,y)$ with respect to variables x_0, y_0, and z_0 as Δs_1, Δs_2, and Δs_3, respectively. From [3, Chapter 15], the elements of the Fisher-information matrix (FIM) are given by the double integrals across the LIS,

$$I_{ij} = \frac{2}{N_0} \iint_{(x,y)\in\mathcal{S}} \mathcal{R}\{\Delta s_j \, (\Delta s_i)^*\} \, \mathrm{d}x\mathrm{d}y. \tag{4.7}$$

For terminals along a central-perpendicular-line (CPL), i.e., with coordinates $(0, 0, z_0)$, the CRLBs of all three Cartesian coordinates can be derived in closed-form. Based on that, the CRLB for arbitrary positions can be obtained with good approximations (see Paper VIII [94] for details), and we show that in general the CRLB decreases quadratically in the surface-area. The same derivations can be applied to practical LIS systems where an unknown phase can be present in the received signal model. In such a case, the CRLB for x and y dimensions remains the same as without an unknown phase. However, the CRLB for z-dimension in relation to the surface-area (in logarithmic domain) varies between 1 and 3, which shows the potential gains of going from massive MIMO to LIS systems.

Chapter 5

Summary of Specific Contributions of the Thesis

The main results and contributions of all the included papers are summarized in this chapter.

Paper 1: On the Design of Channel Shortening Demodulators for Iterative Receivers in Linear Vector Channels

We consider the problem of designing demodulators for linear vector channels with memory that use reduced-size trellis descriptions for the received signal. We assume an overall iterative receiver, and use interference cancellation (IC) based on the soft information provided by the outer decoder, to mitigate the parts of the signal that are not covered by the reduced-size trellis description. In order to reach a trellis description, a linear filter is applied as front-end to compress the signal structure into a small trellis. This process requires three parameters to be designed: (i) the front-end filter, (ii) the feedback filter through which the IC is done, and (iii) a target response which specifies the trellis. Demodulators of this form have been studied before under then name *channel shortening* (CS), but the interplay between CS, IC and the trellis-search process has not been adequately addressed in the literature. In this paper, we analyze two types of CS demodulators that are based on the Forney and Ungerboeck detection models, respectively. The parameters are jointly optimized based on a generalized mutual information (GMI) function. We also introduce a third type of CS demodulator that is in general suboptimal, but has closed-form solutions. Moreover, signal to noise ratio (SNR) asymptotic properties are analyzed and we show that the third CS demodulator asymptotically converges to the optimal CS demodulator in the sense of GMI-maximization.

Paper II: Optimal Channel Shortener Design for Reduced-State Soft-Output Viterbi Equalizer in Single-Carrier Systems

We consider optimal channel shortener design for reduced-state soft-output Viterbi equalizer (RS-SOVE) in single-carrier (SC) systems. To use RS-SOVE, three receiver filters need to be designed: a prefilter, a target response and a feedback filter. The collection of these three filters are commonly referred to as the "channel shortener". Conventionally, the channel shortener is designed to transform an intersymbol interference (ISI) channel into an equivalent minimum-phase equivalent form. In this paper, we design the channel shortener to maximize a mutual information lower bound (MILB) based on a mismatched detection model. By taking the decision-feedback quality in the RS-SOVE into consideration, the prefilter and feedback filter are found in closed forms, while the target response is optimized via a gradient-ascending approach with the gradient explicitly derived. The information theoretical properties of the proposed channel shortener are analyzed. Moreover, we show through numerical results that, the proposed channel shortener design achieves superior detection performance compared to previous channel shortener designs at medium and high code-rates.

Paper III: A Soft-Output MIMO Detector with Achievable Information Rate based Partial Marginalization

We propose a soft-output detector for multi-input multi-output (MIMO) channels that utilizes achievable information rate (AIR) based partial marginalization (PM). The proposed AIR based PM (AIR-PM) detector has superior performance compared to previously proposed PM designs and other soft-output detectors such as K-best, while at the same time yielding lower computational complexity, a detection latency that is independent of the number of transmit layers, and straightforward inclusion of soft input information. Using a tree representation of the MIMO signal, the key property of the AIR-PM is that the connections among all child layers are broken. Therefore, least-square (LS) estimates used for marginalization are obtained independently and in parallel, which have better quality than the zero-forcing decision feedback (ZF-DF) estimates used in previous PM designs. Such a property of the AIR-PM detector is designed via a mismatched detection model that maximizes the AIR. Furthermore, we show that the chain rule holds for the AIR calculation, which facilitates an information theoretic characterization of the AIR-PM detector.

Paper IV: Modulus Zero-Forcing Detection for MIMO Channels

We propose a modulus based zero-forcing (MZF) detection for multi-input multi-output (MIMO) channels. Traditionally, a ZF detector nulls out all interferences from other layers

when detecting a current layer, which can yield suboptimal detection-performance due to the noise-enhancement issue. In many communication systems, finite alphabets such as M quadrature-amplitude-modulation (QAM) are widely used, which comprises $\sqrt{M}$ pulse-amplitude-modulation (PAM) symbols for the real and imaginary parts. With finite alphabets, one feasible way to improve ZF detection is to allow controllable interferences that can be removed away by modulus operations.

Paper v: A Generalized Zero-Forcing Precoder with Successive Dirty-Paper Coding in MISO Broadcast Channels

We consider precoder designs for multiuser multi-input single-output (MISO) broadcasting channels. Instead of using a traditional linear zero-forcing (ZF) precoder, we propose a generalized ZF (GZF) precoder in conjunction with successive dirty-paper coding (DPC) for data-transmissions, namely, the GZF-DP precoder, where the suffix 'DP' stands for 'dirty-paper'. The GZF-DP precoder is designed to generate a band-shaped and lower-triangular effective channel $\boldsymbol{F}$ such that only the entries along the main diagonal and the ν first lower-diagonals can take nonzero values. Utilizing the successive DPC, the known non-causal inter-user interferences from the other (up to) ν users are canceled through successive encoding. We analyze optimal GZF-DP precoder designs both for sum-rate and minimum user-rate maximizations. Utilizing Lagrange multipliers, the optimal precoders for both cases are solved in closed-forms in relation to optimal power allocations. For the sum-rate maximization, the optimal power allocation can be found through water-filling, but with modified water-levels depending on the parameter ν. While for the minimum user-rate maximization that measures the quality of the service (QoS), the optimal power allocation is directly solved in closed-form which also depends on ν. Moreover, we propose two low-complexity user-ordering algorithms for the GZF-DP precoder designs for both maximizations, respectively. We show through numerical results that, the proposed GZF-DP precoder with a small ν (≤ 3) renders significant rate increments compared to the previous precoder designs such as the linear ZF and user-grouping based DPC (UG-DP) precoders.

Paper vi: Linear Precoder Design for MIMO-ISI Broadcasting Channels under Channel Shortening Detection

We consider optimal precoder design for multi-user multi-input multi-output (MIMO) broadcasting channels in single-carrier (SC) systems. Instead of linear detection, we assume that advanced nonlinear channel shortening (CS) detectors are utilized at the receivers. Such a scenario is challenging for precoder design as the uplink-downlink duality is inapplicable. The target of our linear precoder design is to maximize the sum of the achiev-

able information rate (sum-AIR), with AIR of each user being explicitly derived. We analyze such a precoder design in general, and provide an efficient per-user based optimization algorithm for the design of block-diagonalization precoder.

Paper VII: Beyond Massive-MIMO: The Potential of Data-Transmission with Large Intelligent Surfaces

We consider the potential of data-transmission in a system with a massive number of radiating and sensing elements, thought of as a contiguous surface of electromagnetically active material. We refer to this as a large intelligent surface (LIS). We firstly consider capacities of single-antenna autonomous terminals communicating to the LIS where the entire surface is used as a receiving antenna array. Under the condition that the surface-area is sufficiently large, the received signal after a matched-filtering (MF) operation can be closely approximated by a sinc-function-like intersymbol interference (ISI) channel. Secondly, we analyze the capacity per square meter (m^2) deployed surface, $\hat{\mathcal{C}}$, that is achievable for a fixed transmit-power per volume-unit, $\hat{P}$; the volume-unit can be m, m^2, and m^3 depending on the scenario under investigation. As terminal-density increases, the limit of $\hat{\mathcal{C}}$ achieved when the wavelength λ approaches zero is $\hat{P}/(2N_0)$ [nats/s/Hz/volume-unit], where N_0 is the spatial power spectral density (PSD) of the additive white Gaussian noise (AWGN). Moreover, we also show that the number of independent signal dimensions per m deployed surface is $2/\lambda$ for one-dimensional terminal-deployment, and π/λ^2 per m^2 for two and three dimensional terminal-deployments. Thirdly, we consider implementations of the LIS in the form of a grid of conventional antenna elements and show that, the sampling lattice that minimizes the surface-area of the LIS and simultaneously obtains one signal space dimension for every spent antenna is the hexagonal lattice. Lastly, we extensively discuss the design of the state-of-the-art low-complexity channel shortening (CS) demodulator for data-transmission with the LIS.

Paper VIII: Beyond Massive-MIMO: The Potential of Positioning with Large Intelligent Surfaces

We consider the potential for positioning with the LIS systems as a following work of Paper VI. In a first step, we derive Fisher-information matrix (FIM) and Cramér-Rao lower bound (CRLB) in closed-form for positioning a terminal located perpendicular to the center of the LIS, whose location we refer to as being on the central perpendicular line (CPL) of the LIS. For a terminal that is not on the CPL, closed-form expressions of the FIM and CRLB seem out of reach, and we alternatively find approximations which are shown to be accurate. Under mild conditions, we show that the CRLB for all three Cartesian dimensions (x, y and z) decreases quadratically in the surface-area of the LIS, except for a terminal exactly on the

CPL where the CRLB for the z-dimension (distance from the LIS) decreases linearly in the same. In a second step, we analyze the CRLB for positioning when there is an unknown phase φ presented in the analog circuits of the LIS. We then show that the CRLBs are dramatically for all three dimensions but decrease in the third-order of the surface-area. Moreover, with an infinitely large LIS the CRLB for the z-dimension with an unknown φ is 6 dB higher than the case without phase uncertainty, and the CRLB for estimating φ converges to a constant that is independent of the wavelength λ. At last, we extensively discuss the impact of centralized and distributed deployments of LIS, and show that a distributed deployment of LIS can enlarge the coverage for positioning and improve the overall performance.

References

[1] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[2] J. Proakis, *Digital communications*, ser. Elect. and Computer Engi.: Commun. and Signal Process. McGraw-Hill, 2001.

[3] S. M. Kay, *Fundamentals of statistical signal processing, volume I: Estimation theory*. Prentice Hall signal processing series, 1993.

[4] A. V. Oppenheim and R. W. Schafer, *Digital signal processing*. Englewood Cliffs, Prentice-Hall, 1989.

[5] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "A low-complexity channel shortening receiver with diversity support for evolved 2G devices," in *IEEE Int. Conf. Commun.(ICC)*, May 2016, pp. 1–7.

[6] H. Kobayashi and D. Tang, "Application of partial-response channel coding to magnetic recording systems," *IBM J. of Res. and Develop.*, vol. 14, no. 4, pp. 368–375, Jul. 1970.

[7] L. Xu and T. Xu, *Digital underwater acoustic communications*. Elsevier Science, 2016.

[8] T. Bouilloc and G. Favier, "Nonlinear channel modeling and identification using baseband Volterra–Parafac models," *Signal Process.*, vol. 92, no. 6, pp. 1492–1498, Jun. 2012.

[9] G. Colavolpe, A. Modenini, and F. Rusek, "Channel shortening for nonlinear satellite channels," *IEEE Commun. Lett.*, vol. 16, no. 12, pp. 1929–1932, Dec. 2012.

[10] A. F. Molisch, *Wireless communications*. Wiley-IEEE Press, 2010, vol. the seond edition.

[11] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA evolution and LTE*. John Wiley & sons, 2007.

[12] S. Sesia, M. Baker, and I. Toufik, *LTE-the UMTS long term evolution: From theory to practice*. John Wiley & Sons, 2011.

[13] 3GPP TS 36.211, "Evolved universal terrestrial radio access (E-UTRA): Physical channels and modulation," *Release 14*, Dec. 2016.

[14] 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 Process. Mag.*, vol. 30, no. 100, pp. 40–60, Jan. 2013.

[15] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[16] T. Marzetta, E. Larsson, H. Yang, and H. Ngo, *Fundamentals of massive MIMO*. Cambridge University Press, 2016.

[17] H. Q. Ngo, *Massive MIMO: Fundamentals and system designs*. Linköping University Electronic Press, 2015, vol. 1642.

[18] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[19] J. Flordelis, S. Hu, F. Rusek, O. Edfors, G. Dahman, X. Gao, and F. Tufvesson, "Exploiting antenna correlation in measured massive MIMO channels," in *IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, 2016, pp. 1–6.

[20] Y. Kishiyama, A. Benjebbour, T. Nakamura, and H. Ishii, "Future steps of LTE-A: Evolution toward integration of local area and wide area systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 12–18, Mar, 2013.

[21] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb, 2014.

[22] D. D. Falconer and F. R. Magee, "Adaptive channel memory truncation for maximum likelihood sequence estimation," *The Bell Syst. Tech. J.*, vol. 52, no. 9, pp. 1541–1562, Nov. 1973.

[23] S. Fredricsson, "Joint optimization of transmitter and receiver filters in digital pam systems with a Viterbi detector," *IEEE Trans. Inf. Theory*, vol. 22, no. 2, pp. 200–210, Mar. 1976.

[24] C. Beare, "The choice of the desired impulse response in combined linear-Viterbi algorithm equalizers," *IEEE Trans. Commun.*, vol. 26, no. 8, pp. 1301–1307, Aug. 1978.

[25] N. Sundstrom, O. Edfors, P. Ödling, H. Eriksson, T. Koski, and P. O. Börjesson, "Combined linear-Viterbi equalizers-a comparative study and a minimax design," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, vol. 2, Jun. 1994, pp. 1263–1267.

[26] N. Al-Dhahir and J. M. Cioffi, "Efficiently computed reduced-parameter input-aided MMSE equalizers for ML detection: A unified approach," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 903–915, May 1996.

[27] M. A. Lagunas, A. Perez-Neia, and J. Vidal, "Joint beamforming and Viterbi equalizer in wireless communications," in *Conference Record of the 31st Asilomar Conf. Signals, Syst. and Comput.(ACSSC)*, vol. 1, Nov. 1997, pp. 915–919.

[28] S. A. Aldosari, S. A. Alshebeili, and A. M. Al-Sanie, "A new MSE approach for combined linear-Viterbi equalizers," in *Proc. IEEE Veh. Technol. Conf. (VTC), Tokyo*, vol. 3, May 2000, pp. 1707–1711.

[29] S. Badri-Höher and P. A. Höher, "Fast computation of a discrete-time whitened matched filter based on Kalman filtering," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2417–2424, Nov, 2004.

[30] P. A. Höher, S. Badri-Höher, S. Deng, C. Krakowski, and W. Xu, "Joint delayed-decision feedback sequence estimation with adaptive state allocation," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2004, p. 132.

[31] S. Badri-Höher, P. A. Höher, C. Krakowski, and W. Xu, "Impulse response shortening for multiple co-channels," in *IEEE Int. Conf. Commun. (ICC)*, vol. 3, May 2005, pp. 1896–1900.

[32] R. Venkataramani and S. Sankaranarayanan, "Optimal channel shortening equalization for MIMO ISI channels," in *IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2008, pp. 1–5.

[33] U. L. Dang, W. H. Gerstacker, and D. T. Slock, "Maximum SINR prefiltering for reduced-state trellis-based equalization," in *IEEE Int. Conf. Commun. (ICC)*, Jun. 2011, pp. 1–6.

[34] G. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 363–378, May 1972.

[35] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[36] W. H. Gerstacker, F. Obernosterer, R. Meyer, and J. B. Huber, "On prefilter computation for reduced-state equalization," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 793–800, Oct. 2002.

[37] J.-W. Liang, J.-T. Chen, and A. J. Paulraj, "A two-stage hybrid approach for CCI/ISI reduction with space-time processing," *IEEE Commun. Lett.*, vol. 1, no. 6, pp. 163–165, Nov. 1997.

[38] D. Darsena and F. Verde, "Minimum-mean-output-energy blind adaptive channel shortening for multicarrier SIMO transceivers," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5755–5771, Dec. 2007.

[39] I. Aboy-Faycal and A. Lapidoth, "On the capacity of reduced-complexity receivers for intersymbol interference channels," in *IEEE Conv. Elect. and Electron. Eng. in Israel (IEEEI)*, Apr. 2000, pp. 263–266.

[40] R. Venkataramani and M. F. Erden, "A posteriori equivalence: A new perspective for design of optimal channel shortening equalizers," *arXiv preprint: 0710.3802*, 2007.

[41] F. Rusek and A. Prlja, "Optimal channel shortening for MIMO and ISI channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 810–818, Feb. 2012.

[42] N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[43] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[44] M. R. McKay, I. B. Collings, and A. M. Tulino, "Achievable sum rate of MIMO MMSE receivers: A general analytic framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 396–410, Jan. 2010.

[45] G. Ungerboeck, "Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems," *IEEE Trans. Commun.*, vol. 22, no. 5, pp. 624–636, May 1974.

[46] G. Colavolpe and A. Barbieri, "On MAP symbol detection for ISI channels using the Ungerboeck observation model," *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 720–722, Aug 2005.

[47] F. Rusek, G. Colavolpe, and C.-E. W. Sundberg, "40 years with the Ungerboeck model: A look at its potentialities [lecture notes]," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 156–161, May 2015.

[48] F. Rusek, M. Lončar, and A. Prlja, "A comparison of Ungerboeck and Forney models for reduced-complexity ISI equalization," in *IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2007, pp. 1431–1436.

[49] P. Lettieri and M. B. Srivastava, "Advances in wireless terminals," *IEEE Personal Commun.*, vol. 6, no. 1, pp. 6–19, Feb. 1999.

[50] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Oct. 2010.

[51] T. J. O'Shea and J. Hoydis, "An introduction to machine learning communications systems," *arXiv preprint: 1702.00832*, 2017.

[52] M. Tuchler and A. C. Singer, "Turbo equalization: An overview," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 920–952, Feb. 2011.

[53] S.-J. Lee, A. C. Singer, and N. R. Shanbhag, "Linear turbo equalization analysis via BER transfer and EXIT charts," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2883–2897, Jul. 2005.

[54] A. Shaheem, H.-J. Zepernick, and M. Caldera, "Enhanced channel shortened turbo equalization," in *Int. Conf. Advanced Technol. for Commun. (ATC)*, Oct. 2008, pp. 8–11.

[55] S. Hu and F. Rusek, "On the design of reduced state demodulators with interference cancellation for iterative receivers," in *IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Aug. 2015, pp. 981–985.

[56] A. Glavieux, C. Laot, and J. Labat, "Turbo equalization over a frequency selective channel," in *Proc. 1st Symp. Turbo Codes*, Sep. 1997, pp. 96–102.

[57] R. R. Lopes and J. R. Barry, "The soft-feedback equalizer for turbo equalization of highly dispersive channels," *IEEE Trans. Commun.*, vol. 54, no. 5, pp. 783–788, May 2006.

[58] S. Hu and F. Rusek, "On the design of channel shortening demodulators for iterative receivers in linear vector channels," *submitted to IEEE Trans. Inf. Theory, arXiv preprint: 1506.07331*, May 2015.

[59] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "Optimal channel shortener design for reduced-state soft-output Viterbi equalizer in single-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2568–2582, Jun. 2017.

[60] S. Hu and F. Rusek, "A soft-output MIMO detector with achievable information rate based partial marginalization," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1622–1637, Mar. 2017.

[61] ——, "Modulus zero-forcing detection for MIMO channels," *submitted to IEEE Access*, Nov. 2017.

[62] T. K. Moon, *Error Correction Coding: Mathematical methods and algorithms*. Wiley Online Library, 2005.

[63] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

[64] F. A. Monteiro and F. R. Kschischang, "Trellis detection for random lattices," in *IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, Nov. 2011, pp. 755–759.

[65] U. Grenander and G. Szegő, *Toeplitz forms and their applications*. Berkeley: University of California Press, 1958.

[66] R. M. Gray *et al.*, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Commun. and Inf. Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[67] J. W. Choi, B. Shim, A. C. Singer, and N. I. Cho, "Low-complexity decoding via reduced dimension maximum-likelihood search," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1780–1793, Mar. 2010.

[68] M. Čirkić and E. G. Larsson, "SUMIS: Near-optimal soft-in soft-out MIMO detection with low and fixed complexity," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3084–3097, Jun. 2014.

[69] W. Koch and A. Baier, "Optimum and sub-optimum detection of coded data disturbed by time-varying intersymbol interference (applicable to digital mobile radio receivers)," in *IEEE Global Telecommun. Conf. and Exhibition*, Dec. 1990, pp. 1679–1684.

[70] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "Optimal channel shortener design for reduced-state soft-output Viterbi equalizer in single-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2568–2582, Jun. 2017.

[71] M. Lončar and F. Rusek, "On reduced-complexity equalization based on Ungerboeck and Forney observation models," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3784–3789, Jul. 2008.

[72] E. G. Larsson and J. Jaldén, "Fixed-complexity soft MIMO detection via partial marginalization," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3397–3407, Aug. 2008.

[73] D. Persson and E. G. Larsson, "Partial marginalization soft MIMO detection with higher order constellations," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 453–458, Jan. 2011.

[74] Y. Sun, Y. Yang, A. D. Liveris, V. Stankovic, and Z. Xiong, "Near-capacity dirty-paper code design: A source-channel coding approach," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3013–3031, Jun. 2009.

[75] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, Mar, 2005.

[76] J. Maurer, J. Jaldén, D. Seethaler, and G. Matz, "Vector perturbation precoding revisited," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 315–328, Jan, 2011.

[77] J. Zhan, B. Nazer, U. Erez, and M. Gastpar, "Integer-forcing linear receivers," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7661–7685, Oct. 2014.

[78] S. H. Chae, M. Jang, A. Seok-Ki, J. Park, and C. Jeong, "Multilevel coding scheme for integer-forcing MIMO receivers with binary codes," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5428–5441, Aug. 2017.

[79] O. Ordentlich and U. Erez, "Achieving the gains promised by integer-forcing equalization with binary codes," in *IEEE Conv. Elect. and Electron. Eng. in Israel (IEEEI)*, 2010, pp. 703–707.

[80] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[81] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug, 2002.

[82] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[83] S. Hu and F. Rusek, "A generalized zero-forcing precoder with successive dirty-paper coding in MISO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3632–3645, Jun. 2017.

[84] S. Hu, X. Gao, and F. Rusek, "Linear precoder design for MIMO-ISI broadcasting channels under channel shortening detection," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1207–1211, Sep. 2016.

[85] S. Gelfand and M. Pinsker, "Coding for channel with random parameters," *Problems Cont. and Inf. Theory*, vol. 9, no. 1, pp. 19–31, Jan. 1980.

[86] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice strategies for canceling known interference," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 3820–3833, Oct. 2005.

[87] U. Erez and S. Ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3417–3432, Oct. 2005.

[88] W. Yu, D. P. Varodayan, and J. M. Cioffi, "Trellis and convolutional precoding for transmitter-based interference presubtraction," *IEEE Trans. Commun.*, vol. 53, no. 7, pp. 1220–1230, Jul. 2005.

[89] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[90] S. Hu and F. Rusek, "A generalized zero-forcing precoder for multiple antenna Gaussian broadcast channels," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 556–560.

[91] S. Shi, M. Schubert, and H. Boche, "Downlink MMSE transceiver optimization for multiuser MIMO systems: Duality and sum-MSE minimization," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5436–5446, Nov. 2007.

[92] R. Hunger, M. Joham, and W. Utschick, "On the MSE-duality of the broadcast channel and the multiple access channel," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 698–713, Feb. 2009.

[93] S. Hu, F. Rusek, and O. Edfors, "Beyond massive-MIMO: The potential of data-transmission with large intelligent surfaces," *submitted to IEEE Trans. Signal Process., arXiv preprint: 1707.02887*, Jul. 2017.

[94] ——, "Beyond massive-MIMO: The potential of positioning with large intelligent surfaces," *accepted in IEEE Trans. Signal Process., arXiv preprint: 1705.06860*, Dec. 2017.

[95] L. Atzori, A. Iera, and G. Morabito, "The Internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[96] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas in Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[97] G. P. Fettweis, "A 5G wireless communications vision," *Microwave J.*, vol. 55, no. 12, pp. 24–36, Dec, 2012.

[98] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. University of Illinois press, 1998.

[99] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Trans. Emerg. Telecommun. Technol.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[100] S. Hu, K. Chitti, F. Rusek, and O. Edfors, "User assignment with distributed Large Intelligent Surface (LIS) systems," *acccepted in IEEE Wireless Commun. and Netw. Conf. (WCNC), arXiv preprint: 1709.01696*, Apr. 2018.

Part II

Included Papers

Paper 1

On the Design of Channel Shortening Demodulators for Iterative Receivers in Linear Vector Channels

We consider the problem of designing demodulators for linear vector channels with memory that use reduced-size trellis descriptions for the received signal. We assume an overall iterative receiver, and use interference cancellation (IC) based on the soft information provided by the outer decoder, to mitigate the parts of the signal that are not covered by the reduced-size trellis description. In order to reach a trellis description, a linear filter is applied as front-end to compress the signal structure into a small trellis. This process requires three parameters to be designed: (i) the front-end filter, (ii) the feedback filter through which the IC is done, and (iii) a target response which specifies the trellis. Demodulators of this form have been studied before under then name *channel shortening* (CS), but the interplay between CS, IC and the trellis-search process has not been adequately addressed in the literature. In this paper, we analyze two types of CS demodulators that are based on the Forney and Ungerboeck detection models, respectively. The parameters are jointly optimized based on a generalized mutual information (GMI) function. We also introduce a third type of CS demodulator that is in general suboptimal, but has closed-form solutions. Moreover, signal to noise ratio (SNR) asymptotic properties are analyzed and we show that the third CS demodulator asymptotically converges to the optimal CS demodulator in the sense of GMI-maximization.

Based on: S. Hu and F. Rusek, "On the design of channel shortening demodulators for iterative receivers in linear vector channels," submitted to *IEEE Trans. Inf. Theory*, May 2015.

1 Introduction

Channel shortening (CS) demodulators have a long and rich history, see [1–12, 61]. For intersymbol interference (ISI) channels, Forney [13] showed that the Viterbi Algorithm (VA) [14] implements maximum likelihood (ML) detection. However, the complexity of the VA is exponential in the memory of the channel which prohibits its use in many cases of interest. As a remedy, Falconer and Magee proposed in 1973 the concept of CS [1]. The concept is to filter the received signal with a prefilter so that the effective channel has much shorter duration than the original channel, and then apply the VA to the shorter effective channel.

Traditionally, CS demodulators have been optimized from a minimum mean square error (MMSE) perspective [2–10]. Two exceptions from this are the papers [11] and [12]. In [11], the authors attempt to minimize the error probability of an uncoded system which leads to a new notion of posterior equivalence between the target response and the filtered channel. However, since [11] works with uncoded error probabilities, the analysis in [11] does not adequately address the case of coded systems and Shannon capacity properties. The first paper that works with capacity-related cost measures is [12]. In [12] the authors consider the achievable rate, in the form of generalized mutual information (GMI) [15–19], that the transceiver system can achieve if a CS demodulator is adopted. However, [12] is limited to ISI channels only, and the design method in [12] of the CS demodulator is in fact not always possible to execute. The limitations of [12] were first dealt with in [18], which extended the CS concept to any linear vector channel and resulted in a closed-form optimization procedure.

Iterative receivers such as turbo equalization [51–56] followed as a natural extension to turbo codes as an iterative technique for detection and decoding of forward error correction (FEC) protected data that is transmitted over dispersive channel. However, when it comes to turbo equalization, common settings of the equalizer are [51] the maximum *a posterior* (MAP) demodulator [23] and its suboptimal variants such as dimension-reduction and subspace based detections [63, 64], and MMSE based approaches [52, 55, 56, 65] that replace the MAP demodulator with a linear equalizer or a decision feedback equalizer (DFE) to reduce the prohibitive complexity of the MAP demodulator. One important open problem in the area of turbo equalization is the development of other non-trellis-based detection methods that provide performance between that of MAP and MMSE performance [43, 51]. Instead of fully removing the trellis-based detection, another possible approach is to reduce the memory size of the original linear vector channel through an interference cancellation (IC) based prefiltering. To the best of our knowledge, there is only limited literature [54, 62] on such a design of demodulator that combines both IC based prefiltering and a memory-size shortened BCJR in iterative receiver design. A closely related concept is delayed-decision-feedback-sequence-estimation (DDFSE) [21, 57], which also reduces the number of states

in the BCJR. However, in DDFSE the IC is done within a single iteration, and not between the iterations of an iterative receiver.

In this paper, we generalize the idea in [18] of GMI-maximization based CS demodulators to iterative receivers. With iterative receivers it is reasonable to expect that better performance can be reached by allowing the parameters of the CS demodulator to change in each iteration. The CS demodulator in [18] does not take the prior information into account, rendering its design static in all iterations. We aim at constructing a CS demodulator that takes soft information provided by the outer decoder into account so that the parameters of the CS demodulator are designed for a particular level of prior knowledge. This procedure includes an IC mechanism to deal with the signal part that can not be handled by the trellis-search. Preliminary results for CS demodulators in iterative receivers are available in [20], but this paper non-trivially advances the state-of-the-art.

Although the trellis-search based detection is still utilized in the CS demodulator, the memory size ν of the linear vector channel has been reduced which results in significant complexity reduction compared to the MAP demodulator. Meanwhile, with different values of ν, the CS demodulator provides trade-off between the performance of MMSE and MAP. As will become clear later, the CS demodulator is closely related to the concept of linear MMSE receiver with parallel interference cancellation (LMMSE-PIC)[26–28], which cooperates the soft information into the filter coefficients and interference cancellation process. With setting $\nu = 0$, the CS demodulator is identical to the LMMSE-PIC demodulator whose trellis-search process is trivial since different symbols are assumed to be independent after the front-end filtering. The CS demodulator can also be viewed as an extension of the LMMSE-PIC to include a trellis-search, where the parameters of the front-end filter, IC, and trellis-search are jointly optimized. On the other hand, by setting ν to be equal to the original memory size of the linear vector channel, the CS demodulator is identical to MAP. Therefore, the CS demodulator is a generalized framework that includes both the MAP and LMMSE-PIC in iterative receiver design.

The rest of the paper is organized as follows: The linear vector channel model and the iterative receiver structure are introduced in Sec. II, while the general form of the CS demodulators and the GMI are described in Sec. III. In Sec. IV we analyze three types of CS demodulators for finite length linear vector channels. In Sec. V we deal with ISI channels as asymptotic versions of the results established in Sec. IV. The signal to noise ratio (SNR) asymptotic of the CS demodulators are discussed in Sec. VI. Empirical results are provided in Sec. VII, and Sec. VIII summarizes the paper. For improved readability, we have deferred some long proofs and derivations to Appendices A-K.

Notation

Throughout the paper, a capital bold letter such as $\boldsymbol{A}$ represents a matrix, a lower case bold letter $\boldsymbol{a}$ represents a vector, and a capital letter A represents a number. The expression $\boldsymbol{A} \prec 0$ means matrix $\boldsymbol{A}$ is negative definite, while $\boldsymbol{A} \succ 0$ means $\boldsymbol{A}$ is positive definite. Matrix $\boldsymbol{I}$ represents the identity matrix and in general the dimension will be omitted; when it cannot be understood from the context, we let $\boldsymbol{I}_K$ represent a $K \times K$ identity matrix. Our superscripts have the following meanings: $(\,)^*$ is complex conjugate, $(\,)^{\mathrm{T}}$ is matrix transpose, $(\,)^\dagger$ denotes the conjugate transpose of a matrix, $(\,)^{-1}$ is matrix inverse. In addition, $\propto$ means proportional to, $\mathbb{E}[\,]$ is the expectation operator, $\mathrm{Tr}(\,)$ takes the trace of a matrix, $\mathcal{R}\{\,\}$ returns the real part of a variable, $\otimes$ is the Kronecker multiplication operator, $\mathrm{vec}(\boldsymbol{A})$ is a column vector containing the columns of matrix $\boldsymbol{A}$ stacked on top of each other, and $[A, B]$ is the set of integers $\{k : A \leq k \leq B\}$. Furthermore, we say that a matrix $\boldsymbol{A}$ is banded within diagonals $[-\nu_1, \nu_2]$ ($\nu_1, \nu_2 \geq 0$), if the (k, ℓ)th element $A(k, \ell)$ satisfies[1]

$$A(k, \ell) = 0, \ \ell - k > \nu_1 \ \text{ or } \ k - \ell > \nu_2.$$

Moreover, we define two matrix operators $[\,]_\nu$ and $[\,]_{\backslash\nu}$ such that $\boldsymbol{A} = [\boldsymbol{A}]_\nu + [\boldsymbol{A}]_{\backslash\nu}$, with $[\boldsymbol{A}]_\nu$ banded within diagonals $[-\nu, \nu]$ where $[\boldsymbol{A}]_{\backslash\nu}$ is constrained to zero.

2 System Model

We consider linear vector channels according to

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{y}$ is an $N \times 1$ vector of received signal, $\boldsymbol{x}$ is a $K \times 1$ vector comprising unit energy coded symbols that belong to a constellation $\mathcal{X}$, $\boldsymbol{H}$ is an $N \times K$ matrix representing the communication channel which is perfectly known to the receiver and $\boldsymbol{n}$ is zero-mean complex Gaussian noise vector with covariance matrix $N_0\boldsymbol{I}$. Model (1) may represent many different communication systems, such as for example multi-input multi-output (MIMO) or ISI channels. In the MIMO case, the variables N and K are finite while they grow without bounds in the ISI case. For the former case, a block fading model is assumed, where the coherence time is infinite. The block fading model allows us to perform an analysis for a single symbol period.

Denote x_k as the kth element of $\boldsymbol{x}$ and $\boldsymbol{h}_k$ as the kth column vector of $\boldsymbol{H}$, (1) can be

[1] Note that ν_1 refers to the number of upper diagonals of $\boldsymbol{A}$ that are nonzero. We have this convention in order to subsequently follow standard notation for Toeplitz matrices [41].

rewritten as

$$\boldsymbol{y} = \sum_{k=0}^{K-1} \boldsymbol{h}_k x_k + \boldsymbol{n}. \tag{2}$$

In an iterative receiver, the feedback from the outer decoder can be utilized in the demodulator to improve the performance. As the outer decoder provides the demodulator with *a posteriori* probability (APP) and extrinsic information (in terms of bit log-likelihood ratios (LLRs)) [22, 60], side information is present about the symbols $\boldsymbol{x}$ and we represent this by the probability mass function $p_k(s) = \mathrm{P}(x_k = s), (0 \leq k \leq K-1)$. Note that the side-information does not consider the dependency among the symbols, but are symbol-wise marginal probabilities. This reflects the situation encountered in iterative receivers with perfect interleaving. In those cases, the prior probabilities provided from previous iterations are assumed independent, i.e., $\mathrm{P}(\boldsymbol{x} = \boldsymbol{s}) = \prod p_k(s)$. Due to the perfect interleaving assumption, the demodulator can compute $\hat{\boldsymbol{x}} = \mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}] = [\ \hat{x}_0, \hat{x}_1, \cdots, \hat{x}_{K-1}\ ]^{\mathrm{T}}$ in a per-entry fashion as

$$\hat{x}_k = \sum_{s \in \mathcal{X}} s p_k(s),$$

where the expectations are computed with respect to the prior distribution $p_k(s)$.

With soft information $\hat{\boldsymbol{x}}$, we define a $K \times K$ diagonal matrix $\boldsymbol{P}$ as follows. For finite length linear vector channels, $\boldsymbol{P}$ equals

$$\boldsymbol{P} = \mathbb{E}_{\mathrm{T}}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{x})}[\boldsymbol{x}\hat{\boldsymbol{x}}^{\dagger}]\big] = \mathbb{E}_{\mathrm{T}}[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\dagger}], \tag{3}$$

where the exception "$\mathbb{E}_{\mathrm{T}}$" is taken over the transmitted blocks of $\boldsymbol{x}$ under the block fading assumption. For ISI case, as the whole data block experiences the same channel, we let

$$\boldsymbol{P} = \alpha \boldsymbol{I}, \tag{4}$$

where the scalar

$$\alpha = \frac{1}{K} \sum_{k=0}^{K-1} |\hat{x}_k|^2. \tag{5}$$

The variable $\boldsymbol{P}$ in (3) can alternatively be written as

$$\boldsymbol{P} = \mathbb{E}_{\mathrm{T}}[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\dagger}] = \mathbb{E}_{\mathrm{T}}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{x})}[\boldsymbol{x}\boldsymbol{x}^{\dagger}]\big] - \mathbb{E}_{\mathrm{T}}\big[\mathrm{cov}(\boldsymbol{x})\big]. \tag{6}$$

Under the natural assumption of soft information that satisfies

$$\mathbb{E}_{\mathrm{T}}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{x})}[\boldsymbol{x}\boldsymbol{x}^{\dagger}]\big] = \boldsymbol{I}, \tag{7}$$

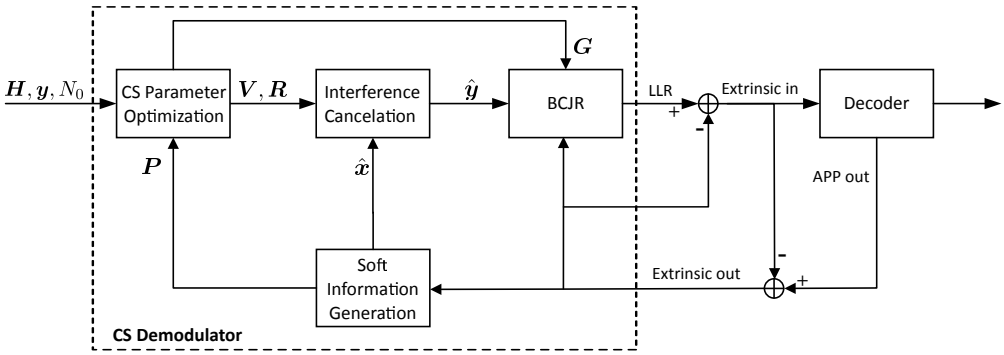


Figure 1: Iterative receiver structure with CS demodulator and outer decoder. The target of the CS demodulator is to maximize the GMI through jointly optimizing the parameters $\boldsymbol{V}$, $\boldsymbol{R}$ and $\boldsymbol{G}$, which are referred to as the front-end filter, IC matrix and trellis representation matrix, respectively.

it follows that $\mathbf{0} \preceq \boldsymbol{P} \preceq \boldsymbol{I}$, and the same also holds for ISI case. The variable $\boldsymbol{P}$ reflects the accuracy of the side information. That is, when there is no soft information available, we have $\boldsymbol{P} = \mathbf{0}$, while with perfect feedback we get $\boldsymbol{P} = \boldsymbol{I}$.

The task of the demodulator is to generate soft information about the symbols in $\boldsymbol{x}$ given the observable $\boldsymbol{y}$ and the side information $\{p_k(s)\}$. The optimal demodulator is the MAP demodulator [23, 24] which evaluates the posterior probabilities $\mathrm{P}(x_k = s|\boldsymbol{y})$. However, the number of leaves of the search tree corresponding to the MAP demodulator is in general $|\mathcal{X}|^K$ which is prohibitive for most practical applications. The purpose of the CS demodulator is to force the signal model to be an lower triangular matrix with only $\nu+1$ $(0 \leq \nu < K-1)$ nonzero diagonals by means of a linear filter[2], where ν is referred to as the memory size of the CS demodulator. Then, a BCJR [25] demodulator can be applied over a trellis with $|\mathcal{X}|^\nu$ states. Moreover, since there is side information present about $\boldsymbol{x}$, the parts of $\boldsymbol{H}$ that are outside the memory of the BCJR can be partly eliminated by means of IC through the prior mean $\hat{\boldsymbol{x}}$.

The structure of an iterative receiver utilizing a CS demodulator is depicted in Fig. 1. The extrinsic information from the outer decoder is used to compute an estimate $\hat{\boldsymbol{x}}$ and a matrix $\boldsymbol{P}$ that indicates the feedback quality. Based on the updated $\boldsymbol{P}$ in each iteration, the optimal CS parameters are found by maximizing the GMI. A prefiltering and IC process are then implemented on $\boldsymbol{y}$ with optimal $\boldsymbol{V}$ and $\boldsymbol{R}$ to obtain the signal $\hat{\boldsymbol{y}}$, which is sent to a memory ν BCJR module specified by an optimal $\boldsymbol{G}$. Moreover, the extrinsic information iteratively exchanged between the BCJR and the outer decoder is also used as *a priori* information for the transmitted symbols. Note that if we set $\nu = K-1$, the search space of the CS demodulator is no longer a trellis but corresponds to the original tree and is therefore equivalent to MAP, and LMMSE-PIC is a special case of the CS demodulation with $\nu = 0$.

[2]For finite length linear vector channels such as MIMO channel, "filtering" means matrix multiplication.

3 The General Form of the CS Demodulator

We state two lemmas that will be useful later, and Lemma 2 can be verified straightforwardly.

Lemma 1. *Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be two $K\times K$ matrices, where $\boldsymbol{A}_1$ is invertible and banded within diagonals $[-\nu,\nu]$. If $[\boldsymbol{A}_1^{-1}]_\nu=[\boldsymbol{A}_2]_\nu$, then*

$$\mathrm{Tr}\big(\boldsymbol{A}_1\boldsymbol{A}_2\big)=\mathrm{Tr}\big(\boldsymbol{I}\big).$$

Proof. Let $\boldsymbol{A}_3=\boldsymbol{A}_2-\boldsymbol{A}_1^{-1}$, then $[\boldsymbol{A}_3]_\nu=\boldsymbol{0}$ and $\boldsymbol{A}_3=[\boldsymbol{A}_3]_{\setminus\nu}$. As $\boldsymbol{A}_1=[\boldsymbol{A}_1]_\nu$, the elements along the main diagonal of $\boldsymbol{A}_1\boldsymbol{A}_3$ are zero. Therefore $\mathrm{Tr}\big(\boldsymbol{A}_1\boldsymbol{A}_2\big)=\mathrm{Tr}\big(\boldsymbol{A}_1(\boldsymbol{A}_1^{-1}+\boldsymbol{A}_3)\big)=\mathrm{Tr}(\boldsymbol{I})$. □

Lemma 2. *Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be two $K\times K$ matrices that are banded within diagonals $[-\nu_1,\nu_2]$ and $[-\nu_3,\nu_4]$, respectively. Then the product $\boldsymbol{A}_1\boldsymbol{A}_2$ is banded within diagonals $[\max(-(\nu_1+\nu_3),1-K),\min(\nu_2+\nu_4,K-1)]$.*

3.1 System Model of the CS Demodulator

The CS demodulators that we investigate operate on the basis of the mismatched[3] function

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x})=\exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{V}\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{x}})\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big) \tag{8}$$

instead of the true conditional probability

$$p(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{(\pi N_0)^N}\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2}{N_0}\right). \tag{9}$$

The matrices $\boldsymbol{V}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ are the front-end filter, IC matrix, and trellis representation matrix, respectively. Without loss of generality, we have absorbed N_0 into $\boldsymbol{V}$, $\boldsymbol{R}$, and $\boldsymbol{G}$. Models (8) and (9) are equivalent for demodulation if we set $\boldsymbol{V}=\boldsymbol{H}^\dagger/N_0$, $\boldsymbol{R}=\boldsymbol{0}$, and $\boldsymbol{G}=\boldsymbol{H}^\dagger\boldsymbol{H}/N_0$, in which case the CS demodulator represents the MAP demodulator.

The detection model (8) has its roots in Falconer and Magee's paper [1] with adding an IC step, where the system model of the demodulator is described as

$$\tilde{T}(\boldsymbol{y}|\boldsymbol{x})=\exp\big(-\|\boldsymbol{W}\boldsymbol{y}-\boldsymbol{T}\hat{\boldsymbol{x}}-\boldsymbol{F}\boldsymbol{x}\|^2\big) \tag{10}$$

[3]By "mismatched" we mean that $\tilde{p}(\boldsymbol{y}|\boldsymbol{x})$ may not be a valid probability distribution function and in general differs from the true conditional probability distribution function $p(\boldsymbol{y}|\boldsymbol{x})$ even with $\hat{\boldsymbol{x}}=\boldsymbol{0}$, but such a "mismatched" property is for the purpose of reducing the size of trellis description in the BCJR.

By setting $\boldsymbol{T}=\boldsymbol{0}$, we obtain the same system model as in [1]. If identifying $\boldsymbol{V}=\boldsymbol{F}^\dagger\boldsymbol{W}$, $\boldsymbol{R}=\boldsymbol{F}^\dagger\boldsymbol{T}$, and $\boldsymbol{G}=\boldsymbol{F}^\dagger\boldsymbol{F}$, model (10) is equivalent to (8) since

$$\begin{aligned}\tilde{T}(\boldsymbol{y}|\boldsymbol{x}) &\propto \exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{F}^\dagger\boldsymbol{W}\boldsymbol{y}-\boldsymbol{F}^\dagger\boldsymbol{T}\hat{\boldsymbol{x}})\}-\boldsymbol{x}^\dagger\boldsymbol{F}^\dagger\boldsymbol{F}\boldsymbol{x}\big)\\ &= \exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{V}\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{x}})\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big).\end{aligned}$$

The detection model (10) is usually denoted as "Forney" model [1] due to its Euclidean-distance form, while the more general model (8) is called "Ungerboeck" model [32, 36, 37]. An advantage of the Ungerboeck model over the Forney model is that the parameter optimization through GMI-maximization is simpler [18]. However, as both models can be viewed as "natural" CS demodulators, we shall investigate both in CS demodulator design for iterative receivers.

In order to optimize $(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$, we choose to work with the GMI which is an achievable rate for a receiver that operates on the basis of a mismatched version of the channel law. The GMI in nats/channel is defined as

$$I_{\mathrm{GMI}}=-\mathbb{E}_{p(\boldsymbol{y})}\left[\log\tilde{p}(\boldsymbol{y})\right]+\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{x})}\left[\log\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right] \tag{11}$$

where $\tilde{p}(\boldsymbol{y})=(1/\pi^K)\int\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\exp(-\|\boldsymbol{x}\|^2)\mathrm{d}\boldsymbol{x}$ and the expectation is taken over the true statistics $p(\boldsymbol{y})$ and $p(\boldsymbol{y},\boldsymbol{x})$. Although finite constellations $\mathcal{X}$ are almost always used in practice, they are hard to analyze. In order to obtain a mathematically tractable problem, here we use a zero-mean, unit variance, complex Gaussian constellation for each entry of $\boldsymbol{x}$. With Gaussian inputs, the trellis discussed earlier has no proper meaning as the number of states is infinite even for finite ν. However, the Gaussian assumption is only made in order to design the receiver parameters. We first state Theorem 1 which shows the calculation of the GMI for model (8).

Theorem 1. *The GMI for the detection model (8) equals*

$$\begin{aligned}I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G}) &= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)-\mathrm{Tr}(\boldsymbol{G})+2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}\\ &\quad-\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big(\boldsymbol{V}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^\dagger)\boldsymbol{V}^\dagger-2\mathcal{R}\big\{\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^\dagger\big\}+\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^\dagger\big)\big).\end{aligned} \tag{12}$$

The proof of Theorem 1 is given in Appendix A. Here we make the same assumption as in [18] that $\boldsymbol{I}+\boldsymbol{G}$ is positive definite, otherwise the GMI is not well defined. With any parameters $(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$, the GMI can be calculated in (12), although they may not be optimal in the sense GMI-maximization. We illustrate Theorem 1 with two examples.

Example 1. *Extended Zero-Forcing filter (EZF). We extend the zero-Forcing filter [30] to only partly invert the channel so that a trellis-search is necessary after the EZF front-end filter. In view of the CS demodulator, we can select the parameters in (8) as:*

$$\boldsymbol{V}=(\boldsymbol{I}+\boldsymbol{G})(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\boldsymbol{H}^\dagger,\ \boldsymbol{R}=\boldsymbol{0},$$

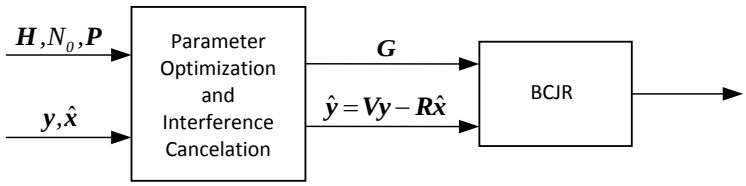


Figure 2: CS demodulator that maximizes the GMI based on the tuple $(\hat{\boldsymbol{y}}, \boldsymbol{x})$.

and then optimize (12) over G. To satisfy the constraint of having a trellis with $|\mathcal{X}|^\nu$ states, we should have $\boldsymbol{G} = [\boldsymbol{G}]_\nu$. The optimal $\boldsymbol{G}$, in the sense of maximizing (12), will be shown (Theorem 2) to satisfy

$$[(\boldsymbol{I}+\boldsymbol{G})^{-1}]_\nu = N_0[(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_\nu.$$

Utilizing Lemma 1, the GMI in (12) for the optimal $\boldsymbol{G}$ equals

$$I_{\mathrm{GMI}} = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\boldsymbol{I} - N_0(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}(\boldsymbol{I}+\boldsymbol{G})\big) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big).$$

Example 2. *Truncated Matched filter (TMF). As previously mentioned, the MAP demodulator (9) can be written in the form (8) by setting $\boldsymbol{V} = \boldsymbol{H}^\dagger/N_0$, $\boldsymbol{R} = \boldsymbol{0}$ and $\boldsymbol{G} = \boldsymbol{H}^\dagger\boldsymbol{H}/N_0$. The front-end is in this case a matched filter [31] and the BCJR needs to be implemented over the Ungerboeck model [32]. To reach a trellis with $|\mathcal{X}|^\nu$ states, we can truncate $\boldsymbol{G}$ to its center $2\nu+1$ diagonals, i.e., we can use the following parameters in (8):*

$$\boldsymbol{V} = \boldsymbol{H}^\dagger/N_0, \ \ \boldsymbol{R} = \boldsymbol{0}, \ \ \textit{and } \boldsymbol{G} = [\boldsymbol{H}^\dagger\boldsymbol{H}/N_0]_\nu.$$

With these choices, the GMI in (12) equals

$$I_{\mathrm{GMI}} = \log\big(\det(\boldsymbol{I}+[\boldsymbol{H}^\dagger\boldsymbol{H}/N_0]_\nu)\big) - \mathrm{Tr}\big(\boldsymbol{H}^\dagger\boldsymbol{H}\big(N_0\boldsymbol{I}+[\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu\big)^{-1}[\boldsymbol{H}^\dagger\boldsymbol{H}]_{\setminus\nu}\big).$$

3.2 Constraints on the Parameter $\boldsymbol{R}$ for the CS Demodulator

As mentioned earlier, optimization of the demodulator will be made on the basis of GMI which is evaluated for the statistical model of the tuple $(\boldsymbol{x}, \hat{\boldsymbol{y}})$. As illustrated in Fig. 2, our approach to design a CS demodulator consists of two steps:

- Construction of a signal $\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}}$ based on the received signal $\boldsymbol{y}$ and prior mean $\hat{\boldsymbol{x}}$;
- BCJR demodulation of $\hat{\boldsymbol{y}}$ operating on a reduced number of states $|\mathcal{X}|^\nu$.

This procedure is fully analogous to LMMSE-PIC demodulator which first subtracts the interference, applies a Wiener filter, and concludes by a BCJR that operates with a diagonal

matrix $\boldsymbol{G}$. The statistical behavior of $(\hat{\boldsymbol{y}}, \boldsymbol{x})$ may be superior to that of the original $(\boldsymbol{y}, \boldsymbol{x})$ as the former tuple corresponds to a statistically different channel than the true one. As what will be shortly shown in Example 3, the GMI obtained with tuple $(\hat{\boldsymbol{y}}, \boldsymbol{x})$ based on perfect feedback $\hat{\boldsymbol{x}}$ can be infinitely large, which exceeds the channel capacity with the original tuple $(\boldsymbol{y}, \boldsymbol{x})$. Therefore, the computed value of GMI may have little relevance for the performance of the transceiver system. In order for GMI to have bearing on performance, it is critical to put constraints on $\boldsymbol{R}$ as the next example will show.

Example 3. *Let the system model be*

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$$

with noise density N_0, and $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{n}$ are $K\times 1$ vectors. Assume perfect feedback information, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{x}$. The demodulator parameters are taken as $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{R} = -(1+\beta)\boldsymbol{I}$, and $\boldsymbol{G} = \beta\boldsymbol{I}$, β an arbitrary positive real value, then the statistical model for $\hat{\boldsymbol{y}}$ is

$$\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}} = (1+\beta)\boldsymbol{x}.$$

The GMI in (12) for the tuple $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ is

$$I_{\mathrm{GMI}}(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G}) = K\big(1+\log(1+\beta)\big).$$

In order to maximize the GMI, the demodulator will choose $\beta \to \infty$ to make I_{GMI} infinite. This is because, except for using the feedback information for IC, the demodulator uses the prior mean $\hat{\boldsymbol{x}}$ as a signal energy via $\boldsymbol{R}$. A demodulator equipped with these parameters will have significant error propagation and does not have much operational meaning for an iterative receiver. Thus, we conclude that unless constraints are put on $\boldsymbol{R}$, the GMI value is not relevant.

Three typical shapes of $\boldsymbol{R}$ are specified in Fig. 3. All three have in common that rather than adding signal energy, the rationale of $\boldsymbol{R}$ should be to remove interference. Therefore at the very minimum the diagonal elements of $\boldsymbol{R}$ should be constrained to zero, so that the demodulation of each symbol in $\boldsymbol{x}$ does not rely on its own prior mean $\hat{\boldsymbol{x}}$. Such a constraint is perfectly aligned with the operations of LMMSE-PIC, where $\hat{x}_\ell$ is not used for demodulation of x_ℓ. Furthermore, the rationale of the constraints we impose on $\boldsymbol{R}$ is to follow the principle of extrinsic information: The BCJR module should not rely on the prior information $\hat{x}_\ell$ when demodulating x_ℓ (this requires more than just the diagonal of $\boldsymbol{R}$ to be zero).

We point out that the fact that the GMI can exceed the channel capacity is a consequence of our choice not to include the side information as a prior distribution on $\boldsymbol{x}$ when evaluating the GMI. If we did, then the GMI is decaying with increasing quality of the side information (due to the mutual information $I(\boldsymbol{x}, \boldsymbol{y}|\hat{\boldsymbol{x}})$ goes to 0 as $\hat{\boldsymbol{x}}$ becomes perfect).

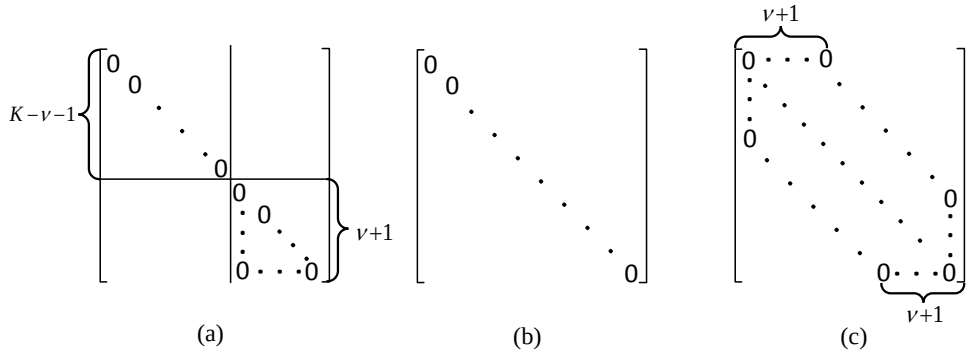


Figure 3: Three different types of shape of matrix $\boldsymbol{R}$, where ν is the memory size of $\boldsymbol{F}$ or $\boldsymbol{G}$, i.e., the memory size of the BCJR.

Finally, we acknowledge the fact that a permutation of the columns of $\boldsymbol{H}$ can boost the performance of the CS demodulator whenever $0<\nu<K-1$ for finite length linear vector channels. However, minimum-phase conversions of ISI channels are not beneficial as we will solve for the optimal front-end filter.

4 Parameter Optimization for Finite Length Linear Vector Channel

In this section, we elaborate the parameter optimization for finite length linear vector channels. We introduce three different methods, Method I, Method II and Method III. We start with the classical Forney model (10) based demodulator, i.e., Method I, and then extend the demodulation model into the Ungerboeck model (8), i.e., Method II. As both Method I and Method II need gradient-based approach for the optimization of target response, by carefully examining the properties of the CS demodulator with Ungerboeck model, we propose a suboptimal Method III which has an explicit construction based on an LMMSE-PIC and all parameters are in closed-forms.

4.1 Method I

In Method I, the CS demodulator is based on detection model (10) and the following structures of the CS parameters $(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$ are imposed:

- $\boldsymbol{W}$ is a $K\times N$ matrix with no constraints.
- $\boldsymbol{F}$ is a $K\times K$ lower triangular matrix where only the main diagonal and the first ν lower diagonals are nonzero, i.e., $\boldsymbol{F}$ is banded within diagonals $[0,\nu]$ $(0\leq\nu<K-1)$, where ν is denoted as the memory size of $\boldsymbol{F}$. Moreover, the main diagonal of $\boldsymbol{F}$ is constrained to only contain positive real values.

- $\boldsymbol{T}$ is a $K \times K$ matrix that is constrained to be zero wherever $\boldsymbol{F}$ can take nonzero values.

The constraint of $\boldsymbol{F}$ is to shorten the memory for the trellis-search in BCJR, while the purpose of the constraint on $\boldsymbol{T}$ is to cancel the signal part that $\boldsymbol{F}$ can not handle. From Theorem 1, and by identifying $\boldsymbol{V} = \boldsymbol{F}^\dagger \boldsymbol{W}$, $\boldsymbol{R} = \boldsymbol{F}^\dagger \boldsymbol{T}$, and $\boldsymbol{G} = \boldsymbol{F}^\dagger \boldsymbol{F}$, the GMI in (12) of Method I equals

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F}) \;=\; & \log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger \boldsymbol{F})\big) - \mathrm{Tr}(\boldsymbol{F}^\dagger \boldsymbol{F}) + 2\mathcal{R}\big\{\mathrm{Tr}\big(\boldsymbol{F}^\dagger(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{T}\boldsymbol{P})\big)\big\} \\ & - \mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{F}^\dagger \boldsymbol{F})^{-1} \boldsymbol{L}_1\big) \end{aligned} \tag{13}$$

where

$$\boldsymbol{L}_1 = \boldsymbol{F}^\dagger \boldsymbol{W}(N_0 \boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\dagger)\boldsymbol{W}^\dagger \boldsymbol{F} - 2\mathcal{R}\big\{\boldsymbol{F}^\dagger \boldsymbol{W}\boldsymbol{H}\boldsymbol{P}\boldsymbol{T}^\dagger \boldsymbol{F}\big\} + \boldsymbol{F}^\dagger \boldsymbol{T}\boldsymbol{P}\boldsymbol{T}^\dagger \boldsymbol{F}.$$

With the aforementioned constraints on $\boldsymbol{F}$ and $\boldsymbol{T}$, the matrix $\boldsymbol{R} = \boldsymbol{F}^\dagger \boldsymbol{T}$ has a form of shape (a) in Fig. 3. That is, all diagonal elements are zero as well as the lower triangular part of the $(\nu+1)\times(\nu+1)$ small matrix at the right bottom corner.

In order to optimize (13) over $(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F})$, we first introduce an $S \times K^2$ indication matrix $\boldsymbol{\Omega}$ only consisting of ones and zeros[4], having a single 1 in each row, and S equals the number of elements in $\boldsymbol{T}$ that are allowed to be nonzero. Let $\mathbb{I}(\mathrm{vec}(\boldsymbol{T}))$ be a vector that contains the positions where the vector $\mathrm{vec}(\boldsymbol{T})$ is allowed to be nonzero. Then the value of the kth entry in $\mathbb{I}(\mathrm{vec}(\boldsymbol{T}))$ gives the column where row k of $\boldsymbol{\Omega}$ is 1. That is, the $S \times 1$ vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T})$ stacks the columns of $\boldsymbol{T}$ on top of each other but with all elements that are constrained to zero removed.

With such a definition of $\boldsymbol{\Omega}$, and define two $K \times K$ matrices as,

$$\boldsymbol{M} \;=\; \boldsymbol{H}^\dagger (N_0 \boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\dagger)^{-1} \boldsymbol{H} - \boldsymbol{I}, \tag{14}$$

$$\tilde{\boldsymbol{M}} \;=\; \boldsymbol{P}(\boldsymbol{I}+\boldsymbol{M})\boldsymbol{P} - \boldsymbol{P}, \tag{15}$$

the GMI for the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$ is given in Proposition 1 and the proof is in Appendix B.

Proposition 1. *Define an $S \times K^2$ matrix $\boldsymbol{D} = \boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^*) \otimes \boldsymbol{I}_K\big)$, the optimal $\boldsymbol{W}$ maximizing the GMI in (13) is*

$$\boldsymbol{W}_{\mathrm{opt}} = \boldsymbol{F}^{-\dagger}(\boldsymbol{I} + \boldsymbol{F}^\dagger \boldsymbol{F} + \boldsymbol{F}^\dagger \boldsymbol{T}\boldsymbol{P})\boldsymbol{H}^\dagger (N_0 \boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\dagger)^{-1}, \tag{16}$$

[4]For instance, assuming $\boldsymbol{T} = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$, then the indication matrix $\boldsymbol{\Omega} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, and the vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

and when $\boldsymbol{P}\neq\boldsymbol{0}$, the optimal $\boldsymbol{T}$ maximizing the GMI is given by

$$\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) = -\boldsymbol{\Omega}^{\mathrm{T}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})^{-1}\boldsymbol{F}^{\dagger}\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}). \tag{17}$$

With the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, the GMI reads,

$$I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F}) = \begin{cases} I_1(\boldsymbol{F}), & \boldsymbol{P}=\boldsymbol{0} \\ I_1(\boldsymbol{F})+\delta_1(\boldsymbol{F}), & \boldsymbol{P}\neq\boldsymbol{0} \end{cases} \tag{18}$$

where the functions $I_1(\boldsymbol{F})$ and $\delta_1(\boldsymbol{F})$ are defined as

$$I_1(\boldsymbol{F}) \;=\; K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})\big), \tag{19}$$

$$\delta_1(\boldsymbol{F}) \;=\; -\mathrm{vec}(\boldsymbol{F})^{\dagger}\boldsymbol{D}^{\dagger}\Big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})^{-1}\boldsymbol{F}^{\dagger}\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}\Big)^{-1}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}). \tag{20}$$

Remark 1. *With the definitions in (14) and (15), $\boldsymbol{M}$ is the negative of the MSE matrix and $\tilde{\boldsymbol{M}}\preceq 0$ holds. Hence $\delta_1(\boldsymbol{F})\geq 0$ represents the GMI increments from the soft feedback.*

Before discussing the GMI-maximization of (18), we first state Theorem 2 that deals with a general maximization problem.

Theorem 2. *Define a scalar function I with respect to a $K\times K$ matrix $\boldsymbol{G}$ as*

$$I(\boldsymbol{G}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big) \tag{21}$$

where $\boldsymbol{G}$ satisfies $\boldsymbol{G}=[\boldsymbol{G}]_{\nu}$. Then the optimal $\boldsymbol{G}$ maximizing I is the unique solution that satisfies

$$[(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}]_{\nu} = -[\boldsymbol{M}]_{\nu}. \tag{22}$$

With $\boldsymbol{G}_{\mathrm{opt}}$, the maximal I equals

$$I(\boldsymbol{G}_{\mathrm{opt}}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})\big). \tag{23}$$

Proof. Taking the first order differential of I with respect to $\boldsymbol{G}$ and noticing that $\boldsymbol{G}$ is banded within diagonals $[-\nu,\nu]$, yields (22) after some manipulations. The existence and uniqueness of such an optimal solution for (22) is proved in [34, Theorem 2] and also illustrated in [18, Proposition 2]. By Lemma 1, $\mathrm{Tr}\big([\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}}]^{-1}\boldsymbol{M}\big)=-K$ from (22), and then (23) follows. □

Optimizing over $\boldsymbol{F}$ in (18) when $\boldsymbol{P}\neq\boldsymbol{0}$ is difficult and cannot be carried out in closed-form. In Appendix C we show by an example that (18) is in general non-concave. Therefore, a

gradient based numerical optimization procedure is utilized to search for the optimal $\boldsymbol{F}$. In the ith iteration, we construct

$$\boldsymbol{F}^{(i)} = \boldsymbol{F}^{(i-1)} + \nabla_{\boldsymbol{F}^*} I_{\mathrm{GMI}}\big(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F}^{(i-1)}\big)$$

where $\nabla_{\boldsymbol{F}^*} I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F})$ is the conjugate of the gradient of the GMI with respect to (the nonzero part of) $\boldsymbol{F}$, and is given in Appendix D.

With $\boldsymbol{P} = \boldsymbol{0}$, if replacing $\boldsymbol{F}^\dagger \boldsymbol{F}$ by $\boldsymbol{G}$, (19) has the same form as (21), and $\boldsymbol{G}_{\mathrm{opt}}$ is in closed-form as stated in Theorem 2. If $\boldsymbol{G}_{\mathrm{opt}} \succeq \boldsymbol{0}$, the optimal $\boldsymbol{F}$ then equals the Cholesky decomposition of $\boldsymbol{G}_{\mathrm{opt}}$. Whenever it is not, a gradient based numerical optimization procedure is utilized to optimize (19), and $\boldsymbol{G}_{\mathrm{opt}}$ from Theorem 2 is used to initialize the starting point of $\boldsymbol{F}$ for any $\boldsymbol{P}$, which has been observed to be highly reliable.

Next we establish a connection between the front-end filter $\boldsymbol{W}$ and IC matrix $\boldsymbol{T}$ in Method I.

Proposition 2. *For $\boldsymbol{P} \neq \boldsymbol{0}$, and with the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, the matrix $\boldsymbol{F}^\dagger(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{T}_{\mathrm{opt}})$ is banded within diagonals $[-\nu, K-1]$.*

Proof. Noting that $\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) = \mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}})$ and $\boldsymbol{\Omega}\boldsymbol{\Omega}^{\mathrm{T}} = \boldsymbol{I}$, from (17) and (81), it holds that

$$\begin{aligned}\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^* \otimes \big(\boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) &= \boldsymbol{\Omega}\mathrm{vec}\big(\boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\boldsymbol{T}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\big) \\ &= -\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}), \end{aligned} \tag{24}$$

which shows that, the elements of the matrix $\Delta = \boldsymbol{F}(\boldsymbol{I} + \boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\boldsymbol{T}_{\mathrm{opt}}\tilde{\boldsymbol{M}} + \boldsymbol{F}\boldsymbol{M}\boldsymbol{P}$ are zero wherever $\boldsymbol{T}$ can be nonzero. Hence Δ is banded within diagonals $[0, \nu]$. On the other hand, with the optimal $\boldsymbol{W}$ given in (16) and $\boldsymbol{M}$, $\tilde{\boldsymbol{M}}$ defined in (14) and (15), we have

$$\boldsymbol{F}^\dagger(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{T}_{\mathrm{opt}}) - (\boldsymbol{I} + \boldsymbol{F}^\dagger\boldsymbol{F}) = (\boldsymbol{I} + \boldsymbol{F}^\dagger\boldsymbol{F})\boldsymbol{F}^{-1}\Delta\boldsymbol{P}^{-1}. \tag{25}$$

Note that, $\boldsymbol{F}^{-1}$ is lower triangular, $\boldsymbol{I} + \boldsymbol{F}^\dagger\boldsymbol{F}$ is banded within diagonals $[-\nu, \nu]$, and $\boldsymbol{P}$ is diagonal. Utilizing Lemma 2, the r.h.s in (25) is banded within diagonals $[-\nu, K-1]$. Therefore $\boldsymbol{F}^\dagger(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{T}_{\mathrm{opt}})$ is also banded within diagonals $[-\nu, K-1]$. □

Proposition 2 reveals an interesting and somewhat surprising fact that, although the BCJR only has a memory size ν, the interference outside the memory size ν shall not be perfectly canceled with the optimal CS demodulator in Method I. As will be shown later, such a property also holds for the other two designs of CS demodulator, i.e., Method II and III.

4.2 Method II

Method II origins from Ungerboeck's 1974 paper [32]. Different from Method I, an Ungerboeck detection model (8) instead of the Forney model (10) is applied. The Ungerboeck model has been extensively discussed in [35–37]. The system model (8) has the following constraints:

- $\boldsymbol{V}$ is a $K \times N$ matrix with no constraints.
- $\boldsymbol{G}$ is a $K \times K$ Hermitian matrix satisfying $\boldsymbol{G} = [\boldsymbol{G}]_\nu$ and $\boldsymbol{I} + \boldsymbol{G} \succ 0$, where ν is the memory size of $\boldsymbol{G}$.
- $\boldsymbol{R}$ is a $K \times K$ matrix where the shape can be specified.

Instead of optimizing $(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F})$, in Method II we optimize $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$ for (12). The same definition of the indication matrix $\boldsymbol{\Omega}$ is used as in Method I, but now $\boldsymbol{\Omega}$ corresponds to $\boldsymbol{R}$ instead of $\boldsymbol{T}$. We continue to let S denote the number of elements that are allowed to be nonzero in $\boldsymbol{R}$. That is, the $S \times 1$ vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{R})$ stacks the columns of $\boldsymbol{R}$ on top of each other but with all elements that are constrained to zero removed. In Method II, we have Proposition 3 that shows the GMI calculation with optimal $\boldsymbol{V}$ and $\boldsymbol{R}$.

Proposition 3. *Define an $S \times 1$ vector $\boldsymbol{d} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P})$, the optimal $\boldsymbol{V}$ for the GMI in (12) is*

$$\boldsymbol{V}_{\mathrm{opt}} = (\boldsymbol{I} + \boldsymbol{G} + \boldsymbol{R}_{\mathbf{opt}}\boldsymbol{P})\boldsymbol{H}^{\dagger}(\boldsymbol{H}\boldsymbol{H}^{\dagger} + N_0\boldsymbol{I})^{-1}, \tag{26}$$

and when $\boldsymbol{P} \neq \boldsymbol{0}$, the optimal $\boldsymbol{R}$ maximizing the GMI is given by,

$$\mathrm{vec}(\boldsymbol{R}_{\mathrm{opt}}) = -\boldsymbol{\Omega}^{\mathrm{T}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I} + \boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}. \tag{27}$$

With the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, the GMI in (12) equals

$$I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G}) = \begin{cases} I_2(\boldsymbol{G}), & \boldsymbol{P} = \boldsymbol{0} \\ I_2(\boldsymbol{G}) + \delta_2(\boldsymbol{G}), & \boldsymbol{P} \neq \boldsymbol{0} \end{cases} \tag{28}$$

where the functions $I_2(\boldsymbol{G})$ and $\delta_2(\boldsymbol{G})$ are defined as,

$$I_2(\boldsymbol{G}) = K + \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big), \tag{29}$$

$$\delta_2(\boldsymbol{G}) = -\boldsymbol{d}^{\dagger}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}. \tag{30}$$

The proof is given in Appendix E. Similar to $\delta_1(\boldsymbol{F})$ in Method I, $\delta_2(\boldsymbol{G}) \geq 0$ represents the GMI increment from the soft information.

When $\boldsymbol{P} \neq \boldsymbol{0}$, the optimization over $\boldsymbol{G}$ in (28) also uses a gradient based numerical optimization, and the gradient of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G})$ with respect to (the nonzero part of) $\boldsymbol{G}$ is provided in Appendix F. The closed-from $\boldsymbol{G}$ from Theorem 2 with $\boldsymbol{P} = \boldsymbol{0}$ is still used as the starting point for $\boldsymbol{P} \neq \boldsymbol{0}$. However, different from Method I, the optimization procedure is concave and the proof is given in Appendix G.

Although the optimal $\boldsymbol{R}$ is solved for in closed-form as in (27), we shall specify the constraint (reflected by $\boldsymbol{\Omega}$) on it. We consider two types of $\boldsymbol{R}$ in Method II. Firstly, as we are interested in the comparison between Method I and Method II, we also consider the shape (a) in Fig. 3, which has the same shape as for $\boldsymbol{R} = \boldsymbol{F}^{\dagger}\boldsymbol{T}$ in Method I. Secondly, we consider a band-shaped $\boldsymbol{R}$ with memory size ν_{R}, where shape (b) and (c) in Fig. 3 are typical cases with $\nu_{\mathrm{R}} = 0$ and $\nu_{\mathrm{R}} = \nu$, respectively. With shape (b), we only limit the diagonal elements of $\boldsymbol{R}$ to be zero and intend to eliminate the interference as much as possible. With shape (c), we limit $\boldsymbol{R}$ to have the opposite form of $\boldsymbol{G}$, that is, the elements of $\boldsymbol{R}$ are constrained to be zero wherever $\boldsymbol{G}$ is nonzero. The intention is to only cancel the interference that the BCJR represented by $\boldsymbol{G}$ cannot handle. Shape (c) is based on the same idea as Method I, but now operates on the Ungerboeck model.

The connection between the optimal front-end filter $\boldsymbol{V}$ and IC matrix $\boldsymbol{R}$ in Method II is now established in Proposition 4.

Proposition 4. *For $\boldsymbol{P} \neq \boldsymbol{0}$ and the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$,*

$$[\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}]_{\setminus(\nu+\nu_{\mathrm{R}})} = [\boldsymbol{R}_{\mathrm{opt}}]_{\setminus(\nu+\nu_{\mathrm{R}})}. \tag{31}$$

That is, the elements of $\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}$ and $\boldsymbol{R}_{\mathrm{opt}}$ are equal outside the center $2(\nu+\nu_{\mathrm{R}})+1$ diagonals for any $\boldsymbol{G}$ that is banded within diagonals $[-\nu, \nu]$, where $\nu_{\mathrm{R}} = 0$ for $\boldsymbol{R}$ with both shape (a) and (b), while $\nu_{\mathrm{R}} = \nu$ for $\boldsymbol{R}$ with shape (c).

Proof. Following similar steps as in the proof of Proposition 2, (27) can be rewritten as,

$$\boldsymbol{\Omega}\mathrm{vec}\big((\boldsymbol{I} + \boldsymbol{G})^{-1}\boldsymbol{R}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\big) = -\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}). \tag{32}$$

It shows that, the elements of the matrix $\Delta = (\boldsymbol{I} + \boldsymbol{G})^{-1}\boldsymbol{R}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\boldsymbol{P}^{-1} + \boldsymbol{M}$ are zero wherever $\boldsymbol{R}$ can be nonzero. On the other hand, with the optimal $\boldsymbol{V}$ in (26) we have

$$\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{R}_{\mathrm{opt}} - (\boldsymbol{I}+\boldsymbol{G}) = (\boldsymbol{I}+\boldsymbol{G})\Delta. \tag{33}$$

As $\boldsymbol{I}+\boldsymbol{G}$ is banded within diagonals $[-\nu, \nu]$, utilizing Lemma 2 ($\boldsymbol{R}$ with shape (a) is slightly different, but it can be verified straightforwardly), and with the three shapes of $\boldsymbol{R}$ in Fig. 3, it can be shown that the r.h.s in (33) is banded within diagonals $[-(\nu+\nu_{\mathrm{R}}), \nu+\nu_{\mathrm{R}}]$, where $\nu_{\mathrm{R}} = 0$ for the shape (a) and (b), and $\nu_{\mathrm{R}} = \nu$ fot the shape (c). Therefore, $\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{R}_{\mathrm{opt}}$ on the l.h.s in (33) is banded within diagonals $[-(\nu + \nu_{\mathrm{R}}), \nu + \nu_{\mathrm{R}}]$. □

The same as Proposition 2 for Method I, Proposition 4 shows that the signal part that is not considered in $\boldsymbol{G}$ (the BCJR) shall not be perfectly canceled inside the center $2(\nu+\nu_{\mathrm{R}})+1$ diagonals for Method II, instead of the center $2\nu+1$ diagonals where $\boldsymbol{G}$ is constrained to be nonzero. With LMMSE-PIC, we have $\nu=\nu_{\mathrm{R}}=0$ and Proposition 4 is natural and frequently used. However, when $\nu_{\mathrm{R}}>0$, a more general property is revealed that, $\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}$ and $\boldsymbol{R}$ are only equal outside the center $2(\nu+\nu_{\mathrm{R}})+1$ diagonals.

4.3 Method III

So far we have discussed two types of CS demodulators based on Forney and Ungerboeck detection models, respectively. One disadvantage of them is that, in general both methods need an numerical optimization to obtain the optimal target response. Next, we construct a third method that has closed-form solutions for all CS parameters, although its GMI is suboptimal in general.

Method III relies on the same operations as Method II for $\boldsymbol{P}=\boldsymbol{0}$. By inserting $\boldsymbol{V}_{\mathrm{opt}}$ in (26) into (8) and setting $\boldsymbol{R}=\boldsymbol{0}$, the demodulator actually operates on the mismatched function

$$\begin{aligned}\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) &= \exp\bigl(2\mathcal{R}\{\boldsymbol{x}^{\dagger}\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{y}\}-\boldsymbol{x}^{\dagger}\boldsymbol{G}\boldsymbol{x}\bigr)\\ &= \exp\bigl(2\mathcal{R}\bigl\{\boldsymbol{x}^{\dagger}(\boldsymbol{I}+\boldsymbol{G})\breve{\boldsymbol{x}}\bigr\}-\boldsymbol{x}^{\dagger}\boldsymbol{G}\boldsymbol{x}\bigr)\end{aligned} \tag{34}$$

where $\breve{\boldsymbol{x}}=\boldsymbol{H}^{\dagger}(\boldsymbol{H}\boldsymbol{H}^{\dagger}+N_0\boldsymbol{I})^{-1}\boldsymbol{y}$ is the LMMSE estimate. As can be seen from (34), the BCJR is based on $\breve{\boldsymbol{x}}$. With soft feedback, we can therefore replace $\breve{\boldsymbol{x}}$ by LMMSE-PIC estimates $\tilde{\boldsymbol{x}}$. That is, instead of (34) we operate on

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}}) = \exp\bigl(2\mathcal{R}\bigl\{\boldsymbol{x}^{\dagger}(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{x}}\bigr\}-\boldsymbol{x}^{\dagger}\boldsymbol{G}\boldsymbol{x}\bigr) \tag{35}$$

where $\boldsymbol{G}$ has the same banded-shape as the first two methods, but optimized according to $\tilde{\boldsymbol{x}}$. The estimate $\tilde{\boldsymbol{x}}$ is constructed as follows. As we prefer to handle the interference through the trellis-search process, the IC should not be present within the memory size ν. In other words, the signal vector after the IC that is used to form the kth symbol of $\tilde{\boldsymbol{x}}$ is denoted as $\tilde{\boldsymbol{y}}_k$ and

$$\tilde{\boldsymbol{y}}_k=\boldsymbol{y}-\sum_{n\in\mathcal{A}_k}\boldsymbol{h}_n\hat{x}_n \tag{36}$$

where $\mathcal{A}_k=\bigl\{0\le n\le K-1: n\notin[\max(0,k-\nu),\min(k+\nu,K-1)]\bigr\}$. Denote p_n as the nth diagonal element of $\boldsymbol{P}$, the Wiener filtering coefficients [38] for the kth symbol are calculated through

$$\hat{\boldsymbol{w}}_k=\boldsymbol{h}_k^{\dagger}\bigl(\boldsymbol{H}^{\dagger}\boldsymbol{C}_k\boldsymbol{H}+N_0\boldsymbol{I}\bigr)^{-1} \tag{37}$$

where $\boldsymbol{C}_k$ is a diagonal matrix with the nth diagonal element defined as

$$C_k(n) = \begin{cases} 1-p_n, & k \in \mathcal{A}_k \\ 1, & \text{otherwise.} \end{cases} \tag{38}$$

The estimate $\tilde{\boldsymbol{x}}$ is then obtained through

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} \hat{\boldsymbol{w}}_1\tilde{\boldsymbol{y}}_1 & \hat{\boldsymbol{w}}_2\tilde{\boldsymbol{y}}_2 & \cdots & \hat{\boldsymbol{w}}_K\tilde{\boldsymbol{y}}_K \end{bmatrix}^{\mathrm{T}} = \hat{\boldsymbol{W}}\boldsymbol{y} - \hat{\boldsymbol{C}}\hat{\boldsymbol{x}} \tag{39}$$

where the coefficient matrix $\hat{\boldsymbol{W}}$ and IC matrix $\hat{\boldsymbol{C}}$ defined as

$$\hat{\boldsymbol{W}} = [\, \hat{\boldsymbol{w}}_1^{\mathrm{T}} \; \hat{\boldsymbol{w}}_2^{\mathrm{T}} \; \cdots \; \hat{\boldsymbol{w}}_K^{\mathrm{T}} \,]^{\mathrm{T}}, \tag{40}$$

$$\hat{\boldsymbol{C}} = [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\backslash\nu}. \tag{41}$$

Inserting $\tilde{\boldsymbol{x}}$ in (39) back into (35), the detection model we operate on reads

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) = \exp\big(2\mathcal{R}\big\{\boldsymbol{x}^\dagger\big((\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{W}}\boldsymbol{y} - (\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}\hat{\boldsymbol{x}}\big)\big\} - \boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big). \tag{42}$$

Note that, (42) is a also special case of (8) by identifying

$$\begin{aligned} \boldsymbol{V} &= (\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{W}}, \\ \boldsymbol{R} &= (\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{C}}. \end{aligned}$$

The GMI in (12) in this case reads, after some manipulations,

$$I_{\mathrm{GMI}}(\boldsymbol{G}) = K + \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\hat{\boldsymbol{M}}(\boldsymbol{I}+\boldsymbol{G})\big) \tag{43}$$

with $\hat{\boldsymbol{M}}$ (the updated $\boldsymbol{M}$ in Method II) defined as

$$\begin{aligned} \hat{\boldsymbol{M}} = \; & \hat{\boldsymbol{W}}\boldsymbol{H}\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger + \hat{\boldsymbol{W}}\boldsymbol{H} - \boldsymbol{P}\hat{\boldsymbol{C}}^\dagger + \big(\hat{\boldsymbol{W}}\boldsymbol{H}\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger + \hat{\boldsymbol{W}}\boldsymbol{H} - \boldsymbol{P}\hat{\boldsymbol{C}}^\dagger\big)^\dagger \\ & - \hat{\boldsymbol{W}}(\boldsymbol{H}\boldsymbol{H}^\dagger + N_0\boldsymbol{I})\hat{\boldsymbol{W}}^\dagger - \hat{\boldsymbol{C}}\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger - \boldsymbol{I}, \end{aligned} \tag{44}$$

which can be shown to be the negative of the MSE matrix since

$$\hat{\boldsymbol{M}} = -\mathbb{E}\big[(\boldsymbol{x}-\tilde{\boldsymbol{x}})(\boldsymbol{x}-\tilde{\boldsymbol{x}})^\dagger\big] = -\mathbb{E}\big[(\boldsymbol{x}-\hat{\boldsymbol{W}}\boldsymbol{y}+\hat{\boldsymbol{C}}\hat{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{W}}\boldsymbol{y}+\hat{\boldsymbol{C}}\hat{\boldsymbol{x}})^\dagger\big].$$

The optimal $\boldsymbol{G}$ for (43) is then obtained from Theorem 2, and the optimal GMI reads

$$I_{\mathrm{GMI}}(\boldsymbol{G}_{\mathrm{opt}}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})\big).$$

An graphical overview of Method III for $K=4$ and $\nu=1$ is illustrated in Fig. 4. For any $\boldsymbol{G}$ with memory size ν, the IC matrix $(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{C}}$ is zero along the main diagonal, which guarantees that the extrinsic information will not be used for current symbols in the IC process. In GMI sense, Method III will not outperform Method II with a shape (b) $\boldsymbol{R}$, but it may outperform the GMI of Method II with a shape (c) $\boldsymbol{R}$, as it can be verified that a shape (c) $\boldsymbol{R}$ has zeros at the positions where $(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}$ are also zeros.

Remark 2. *As $\hat{\boldsymbol{W}}\boldsymbol{H} - \hat{\boldsymbol{C}} = [\hat{\boldsymbol{W}}\boldsymbol{H}]_\nu$, by Lemma 2 $(\boldsymbol{I}+\boldsymbol{G})(\hat{\boldsymbol{W}}\boldsymbol{H} - \hat{\boldsymbol{C}})$ is banded within diagonals $[-2\nu, 2\nu]$, which shows that, $[(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{W}}\boldsymbol{H}]_{\backslash 2\nu} = [(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}]_{\backslash 2\nu}$. Therefore, Proposition 4 also holds for Method III with $\nu_{\mathrm{R}} = \nu$.*

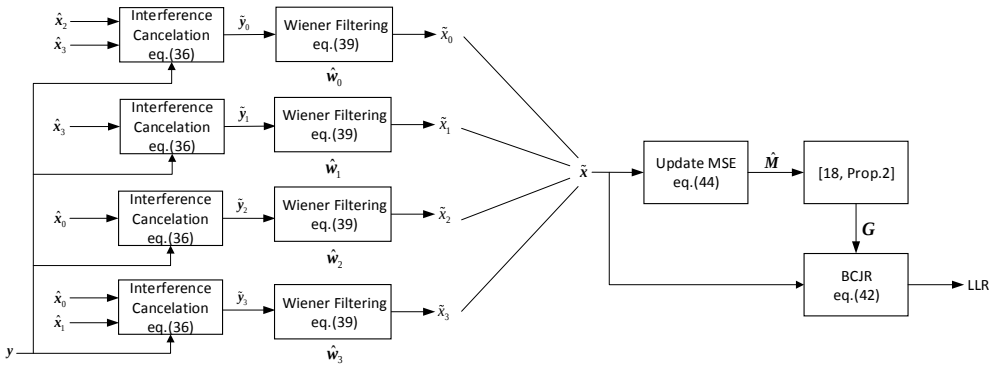


Figure 4: An graphical overview of Method III with $K=4$ and $\nu=1$.

5 Parameter Optimization for ISI Channel

In this section, we extend the CS demodulators to ISI channels where the matrix $\boldsymbol{P}=\alpha\boldsymbol{I}$ and the block length K is infinitely large. The formulas for the achievable rates in (12), (13) and (43) can be directly applied to (1), but as the achievable rate I_{GMI} (as a function of the specified CS parameters) is then dependent on the block length K, we are interested in asymptotic rate

$$\bar{I}=\lim_{K\to\infty}\frac{1}{K}I_{\mathrm{GMI}}.$$

Ideally, in the ISI case the front-end matrix $\boldsymbol{V}$ and IC matrix $\boldsymbol{R}$ correspond to linear filtering operations and the filters are infinitely long, but in practice filters with finite tap lengths are used. Therefore, we analyze the properties of $\boldsymbol{V}$, $\boldsymbol{R}$ (and $\boldsymbol{W}$, $\boldsymbol{T}$) with a finite number of taps and approximate them by band-shaped Toeplitz matrices. Furthermore, the trellis representation matrix $\boldsymbol{G}$ (and $\boldsymbol{F}$), and channel matrix $\boldsymbol{H}$ are also band-shaped Toeplitz matrices. Therefore, in the ISI case all matrices we consider are assumed to be band-shaped Toeplitz matrices, and the band size can be arbitrary and sufficiently large so that we can analyze the asymptotic properties. In [39] a complete theoretic machinery for ISI channels is derived and a result is that, as $K\to\infty$ the linear convolution in (1) can be replaced with a circular convolution.

In the following, we denote the Fourier series associated to a band-shaped Toeplitz matrix $\boldsymbol{E}$ with infinitely large dimensions by $E(\omega)$, where $\boldsymbol{E}$ is constrained to be zero outside the middle $2N_{\mathrm{E}}+1$ diagonals, and N_{E} is referred to as the tap length of $E(\omega)$. The series $E(\omega)$ defined as

$$E(\omega)=\sum_{k=-N_{\mathrm{E}}}^{N_{\mathrm{E}}}e_k\exp(jk\omega)$$

is specified by a vector $\boldsymbol{e} = [\, e_{-N_\mathrm{E}} \ \ldots \ e_{-1}\ e_0\ e_1\ \ldots\ e_{N_\mathrm{E}}\,]$, where e_0 is the element on the main diagonal and e_k is the element on kth lower ($k>0$) or upper ($k<0$) diagonal. As all quantities are evaluated as the block length K grows large, $E(\omega)$ approaches the eigenvalue distribution of $\boldsymbol{E}$ (see [40, 41] for a precise statement of this result). We first state Theorem 3, which is an asymptotic version of Theorem 2 for ISI channels.

Theorem 3. *Assume that two band-shaped Toeplitz matrices $\boldsymbol{G}$ and $\boldsymbol{M}$ with infinitely large dimensions satisfying $[\boldsymbol{G}]_{\setminus\nu} = \boldsymbol{0}$, $\boldsymbol{I}+\boldsymbol{G} \succ 0$ and $\boldsymbol{M} \prec 0$. Define a scalar function*

$$\bar{I} = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \big(\log(1+G(\omega)) + M(\omega)(1+G(\omega))\big)\mathrm{d}\omega. \tag{45}$$

Then, the optimal $G(\omega)$ that maximizes $\bar{I}$ is

$$G_{\mathrm{opt}}(\omega) = |u_0 + \hat{\boldsymbol{u}}\varphi(\omega)|^2 - 1,$$

where the $1\times\nu$ vector $\varphi(\omega) = [\, \exp(j\omega)\ \ \exp(j2\omega)\ \ldots \exp(j\nu\omega)\,]^{\mathrm{T}}$, and

$$\begin{aligned} u_0 &= \frac{1}{\sqrt{\boldsymbol{\tau}_1^\dagger \boldsymbol{\tau}_2^{-1}\boldsymbol{\tau}_1 - \tau_0}}, \\ \hat{\boldsymbol{u}} &= -u_0 \boldsymbol{\tau}_1^\dagger \boldsymbol{\tau}_2^{-1}. \end{aligned} \tag{46}$$

The real scalar τ_0, $\nu\times 1$ vector $\boldsymbol{\tau}_1$, and $\nu\times\nu$ matrix $\boldsymbol{\tau}_2$ are defined as

$$\begin{aligned} \tau_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\mathrm{d}\omega, \\ \boldsymbol{\tau}_1 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\varphi(\omega)\mathrm{d}\omega, \\ \boldsymbol{\tau}_2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\varphi(\omega)\varphi(\omega)^\dagger\mathrm{d}\omega. \end{aligned} \tag{47}$$

Furthermore, with $G_{\mathrm{opt}}(\omega)$ the optimal $\bar{I}$ reads

$$\bar{I} = 2\log(u_0). \tag{48}$$

Proof. As $\boldsymbol{I}+\boldsymbol{G} \succ 0$, we assume that $1+G(\omega) = |U(\omega)|^2$, with $U(\omega) = u_0 + \hat{\boldsymbol{u}}\varphi(\omega)$ and $\hat{\boldsymbol{u}} = [\, u_1\ u_2\ \ldots\ u_\nu\,]$. Then $\bar{I}$ in (45) can be rewritten as

$$\bar{I} = 1 + 2\log(u_0) + \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\big(u_0^2 + 2\mathcal{R}\{u_0\hat{\boldsymbol{u}}\varphi(\omega)\} + \hat{\boldsymbol{u}}\varphi(\omega)\varphi^\dagger(\omega)\hat{\boldsymbol{u}}^\dagger\big)\mathrm{d}\omega. \tag{49}$$

Taking the first order differentials with respect to u_0 and $\hat{\boldsymbol{u}}$ and optimizing them directly results in the optimal solution (46). Inserting (46) back into (49) and after some manipulations, the optimal asymptotic rate is then in (48). □

5.1 Method I

The structures of $(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F})$ are the same as in Section 4.1, except that now the matrices have infinite dimensions. Applying Szegő's eigenvalue distribution theorem [40] to (13), the asymptotic rate reads

$$\begin{aligned}\bar{I}\big(W(\omega),T(\omega),F(\omega)\big) &= \lim_{K\to\infty}\frac{1}{K}I_{\mathrm{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\bigg(\log\big(1+|F(\omega)|^2\big)-|F(\omega)|^2-\frac{L_1(\omega)}{1+|F(\omega)|^2}\bigg)\mathrm{d}\omega\\ &\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)\big(W(\omega)H(\omega)-\alpha T(\omega)\big)\big\}\mathrm{d}\omega \qquad (50)\end{aligned}$$

where

$$\begin{aligned}L_1(\omega) = |F(\omega)W(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\alpha|F(\omega)T(\omega)|^2\\ -2\alpha|F(\omega)|^2\mathcal{R}\{H(\omega)W(\omega)T^*(\omega)\}\,.\end{aligned}$$

Note that, the Fourier series associated to $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ in (14) and (15) are

$$M(\omega) = \frac{|H(\omega)|^2}{N_0+|H(\omega)|^2}-1, \qquad (51)$$

$$\tilde{M}(\omega) = \alpha^2(M(\omega)+1)-\alpha. \qquad (52)$$

Further, define a $(2N_{\mathrm{T}}-\nu)\times 1$ vector

$$\phi(\omega)=\big[\exp\big(-jN_{\mathrm{T}}\omega\big)\ \ldots\ \exp\big(-j(\nu+1)\omega\big)\ \exp\big(j(\nu+1)\omega\big)\ \ldots\ \exp\big(jN_{\mathrm{T}}\omega\big)\big]^{\mathrm{T}}, \qquad (53)$$

a $(2N_{\mathrm{T}}-\nu)\times 1$ vector $\boldsymbol{\varepsilon}_1$, and a $(2N_{\mathrm{T}}-\nu)\times(2N_{\mathrm{T}}-\nu)$ Hermitian matrix $\boldsymbol{\varepsilon}_2$ as

$$\begin{aligned}\boldsymbol{\varepsilon}_1 &= \frac{\alpha}{2\pi}\int_{-\pi}^{\pi}M(\omega)F^*(\omega)\phi(\omega)\mathrm{d}\omega,\\ \boldsymbol{\varepsilon}_2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2\phi(\omega)\phi(\omega)^{\dagger}}{1+|F(\omega)|^2}\mathrm{d}\omega, \qquad (54)\end{aligned}$$

where N_{T} is the tap length of $T(\omega)$, and $\nu+1$ is the band size where matrix $\boldsymbol{T}$ is constrained to zero. Then, we have Proposition 5 with the proof[5] given in Appendix H.

Proposition 5. *The optimal $W(\omega)$ for the asymptotic rate in (50) is*

$$W_{\mathrm{opt}}(\omega) = \frac{H^*(\omega)}{F^*(\omega)(N_0+|H(\omega)|^2)}\big(1+|F(\omega)|^2+\alpha F^*(\omega)T_{\mathrm{opt}}(\omega)\big), \qquad (55)$$

[5]Proposition 5 is the same as [62, Theorem 1] which has been derived for hard feedback symbols. For completeness, we restate the proof in Appendix H.

and when $0<\alpha\leq 1$*, the optimal* $T(\omega)$ *reads*

$$T_{\text{opt}}(\omega) = -\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\phi(\omega). \tag{56}$$

With the optimal $W(\omega)$ *and* $T(\omega)$*, the asymptotic rate equals*

$$\bar{I}\big(W_{\text{opt}}(\omega), T_{\text{opt}}(\omega), F(\omega)\big) = \begin{cases} \bar{I}_1(F(\omega)), & \alpha = 0 \\ \bar{I}_1(F(\omega)) + \bar{\delta}_1(F(\omega)), & 0 < \alpha \leq 1. \end{cases} \tag{57}$$

The functions $\bar{I}_1(F(\omega))$ *and* $\bar{\delta}_1(F(\omega))$ *are defined as*[6]*,*

$$\bar{I}_1(F(\omega)) = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big) + M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega, \tag{58}$$

$$\bar{\delta}_1(F(\omega)) = -\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1. \tag{59}$$

In the ISI case, Method I is still not concave an example is also provided in Appendix C, and a gradient based optimization is used to optimize $F(\omega)$ with the optimal solution of $G_{\text{opt}}(\omega)$ from Theorem 3 is used to initialize the starting point.

The connection between the optimal front-end filter $W(\omega)$ and the IC filter $T(\omega)$ in Proposition 2 also holds for ISI channels. An asymptotic version of Proposition 2 is stated in Proposition 6.

Proposition 6. *When* $0<\alpha\leq 1$*,* $a_k = b_k$ *holds for* $k<-(\nu+1)$*, where*

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F^*(\omega)W_{\text{opt}}(\omega)H(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega$$

$$b_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F^*(\omega)T_{\text{opt}}(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega.$$

Proof. In Appendix H, the optimal $\tilde{\boldsymbol{t}}$ in (95) satisfies $\tilde{\boldsymbol{t}}_{\text{opt}}\boldsymbol{\varepsilon}_2 = -\boldsymbol{\varepsilon}_1^{\dagger}$. With the definitions of $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$ in (54), this is equivalent to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2 T_{\text{opt}}(\omega)\phi(\omega)^{\dagger}}{1+|F(\omega)|^2}\mathrm{d}\omega = -\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}F(\omega)M(\omega)\phi(\omega)^{\dagger}\mathrm{d}\omega. \tag{60}$$

On the other hand, with W_{opt} in (55) and $M(\omega)$, $\tilde{M}(\omega)$ defined in (51) and (52), we have

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(F^*(\omega)W_{\text{opt}}(\omega)H(\omega) - F^*(\omega)T_{\text{opt}} - \big(1+|F(\omega)|^2\big)\big)\exp\big(-jk\omega\big)\mathrm{d}\omega \\ &\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{\tilde{M}(\omega)F^*(\omega)T_{\text{opt}}(\omega)}{\alpha} + \big(1+|F(\omega)|^2\big)M(\omega)\Big)\exp\big(-jk\omega\big)\mathrm{d}\omega. \end{aligned}\tag{61}$$

[6]Similar to finite length linear vector channels, $\bar{\delta}_1(F(\omega))$ in (59) is only defined for $\alpha\neq 0$ which represents the rate increment with soft information. The same holds for $\bar{\delta}_2(G(\omega))$ in (69) for Method II.

Transforming (60) and (61) back into matrix forms, we have that (24) and (25) hold. Following the same arguments as in the proof of Proposition 2, $\boldsymbol{F}^{\dagger}(\boldsymbol{W}_{\rm opt}\boldsymbol{H}-\boldsymbol{R}_{\rm opt})$ is banded within diagonals $[-\nu, K-1]$. Therefore we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(F^*(\omega)W_{\rm opt}(\omega)H(\omega)-F^*(\omega)T_{\rm opt}(\omega)\big)\exp\big(-jk\omega\big)\mathrm{d}\omega=0$$

whenever $k<-(\nu+1)$, which proves Proposition 6. □

5.2 Method II

The matrices $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$ have the same constraints as in Section 4.2 while the dimensions of these matrices are infinitely large. However, as the shape (a) of $\boldsymbol{R}$ in Fig. 3 is not meaningful as $N, K\to\infty$, it is not considered for ISI case. Applying Szegő's eigenvalue distribution theorem to (12), the asymptotic rate of Method II reads

$$\begin{aligned}\bar{I}(V(\omega),R(\omega),G(\omega)) &= \lim_{K\to\infty}\frac{1}{K}I_{\rm GMI}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log\big(1+G(\omega)\big)-G(\omega)-\frac{L_2(\omega)}{1+G(\omega)}\right)\mathrm{d}\omega\\ &\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{\big(V(\omega)H(\omega)-\alpha R(\omega)\big)\big\}\mathrm{d}\omega \qquad (62)\end{aligned}$$

where

$$L_2(\omega)=|V(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\alpha|R(\omega)|^2-2\alpha\mathcal{R}\big\{H(\omega)V(\omega)R^*(\omega)\big\}.$$

Define a $2(N_{\rm R}-\nu_{\rm R})\times 1$ vector

$$\psi(\omega)=\big[\exp\big(-jN_{\rm R}\omega\big)\ \ldots\ \exp\big(-j(\nu_{\rm R}+1)\omega\big)\ \exp\big(j(\nu_{\rm R}+1)\omega\big)\ \ldots\ \exp\big(jN_{\rm R}\omega\big)\big]^{\rm T}, \qquad (63)$$

a $2(N_{\rm R}-\nu_{\rm R})\times 1$ vector $\boldsymbol{\zeta}_1$, and a $2(N_{\rm R}-\nu_{\rm R})\times 2(N_{\rm R}-\nu_{\rm R})$ Hermitian matrix $\boldsymbol{\zeta}_2$ as

$$\begin{aligned}\boldsymbol{\zeta}_1 &= \frac{\alpha}{2\pi}\int_{-\pi}^{\pi}M(\omega)\psi(\omega)\mathrm{d}\omega,\\ \boldsymbol{\zeta}_2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)\psi(\omega)\psi(\omega)^{\dagger}}{1+G(\omega)}\mathrm{d}\omega, \qquad (64)\end{aligned}$$

where $N_{\rm R}$ denotes the tap length of $R_{\rm opt}(\omega)$, $2\nu_{\rm R}+1$ is the band size where $\boldsymbol{R}$ is constrained to zero, and $M(\omega)$ and $\tilde{M}(\omega)$ are in (51) and (52). Then, we have Proposition 6 with the proof in Appendix I, where we also show that $R(\omega)$ is real and $\boldsymbol{R}$ has Hermitian symmetry.

Proposition 7. *The optimal $V(\omega)$ for (62) is,*

$$V_{\mathrm{opt}}(\omega) = \frac{H^*(\omega)}{N_0+|H(\omega)|^2}\big(1+G(\omega)+\alpha R_{\mathrm{opt}}(\omega)\big), \tag{65}$$

and when $0<\alpha\leq 1$, the optimal $R(\omega)$ reads

$$R_{\mathrm{opt}}(\omega) = -\boldsymbol{\zeta}_1^{\dagger}\boldsymbol{\zeta}_2^{-1}\psi(\omega). \tag{66}$$

With the optimal $V(\omega)$ and $R(\omega)$, the asymptotic rate equals

$$\bar{I}\big(V_{\mathrm{opt}}(\omega), R_{\mathrm{opt}}(\omega), G(\omega)\big) = \begin{cases} \bar{I}_2(G(\omega)), & \alpha = 0 \\ \bar{I}_2(G(\omega)) + \bar{\delta}_2(G(\omega)), & 0 < \alpha \leq 1. \end{cases} \tag{67}$$

The functions $\bar{I}_1(G(\omega))$ and $\bar{\delta}_2(G(\omega))$ are defined as,

$$\bar{I}_2(G(\omega)) = 1+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big)+M(\omega)\big(1+G(\omega)\big)\Big)\mathrm{d}\omega, \tag{68}$$

$$\bar{\delta}_2(G(\omega)) = -\boldsymbol{\zeta}_1^{\dagger}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1. \tag{69}$$

For $0<\alpha\leq 1$, it still needs a gradient based optimization to find the optimal $G(\omega)$ for (68), and the closed-form solution in Theorem 3 is utilized as the starting point. The asymptotic rate $\bar{I}(V_{\mathrm{opt}}(\omega), R_{\mathrm{opt}}(\omega), G(\omega))$ is also concave with respect to $G(\omega)$, which is shown in Appendix J.

Proposition 8. *When $0<\alpha\leq 1$, $a_k=b_k$ holds for $|k|>\nu+\nu_{\mathrm{R}}$, where*

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} V_{\mathrm{opt}}(\omega)H(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega, \tag{70}$$

$$b_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} R_{\mathrm{opt}}(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega. \tag{71}$$

The connection of the optimal $V(\omega)$ and $R(\omega)$ stated in Proposition 8 is an asymptotic version of Proposition 4, and the proof follows the similar approach as Proposition 6. We show an example in Fig. 5 to illustrate Proposition 8 with Method II and $\nu=\nu_{\mathrm{R}}=1$. The Proakis-C [42] channel is tested at an SNR of 10 dB and α equals 0.1, 0.4 and 0.8, respectively. As $\nu_{\mathrm{R}}=1$, b_k as defined in (71) is constrained to zero for $0\leq k\leq 1$. As can be seen, a_k as defined in (70) equals b_k only for $|k|>2$, and when $|k|=2$, a_k and b_k are not identical. This shows that with the optimal $V(\omega)$ and $R(\omega)$, the signal part along the second upper and lower diagonals that is not considered in $G(\omega)$ shall not be perfectly canceled out. This behavior cannot be seen in [43] which treats LMMSE-PIC for ISI channels, due to $\nu=\nu_{\mathrm{R}}=0$.

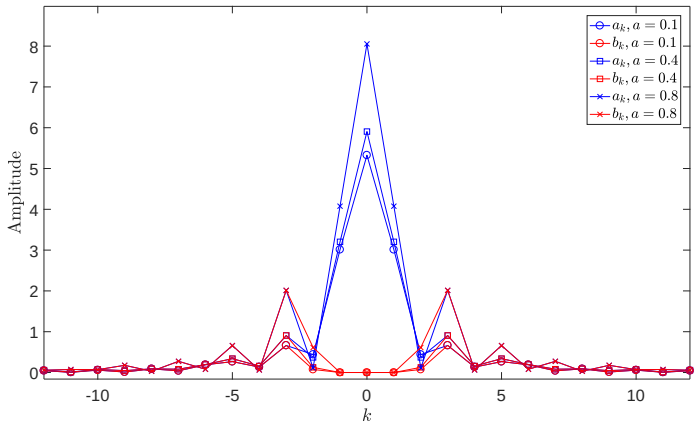


Figure 5: Comparison between a_k and b_k for Method II under Proakis-C channel $\boldsymbol{h} = [\,0.227\ 0.46\ 0.688\ 0.46\ 0.227\,]$.

5.3 Method III

In Method III, from (43) the asymptotic rate reads

$$\bar{I}(G(\omega)) = \lim_{K\to\infty} \frac{1}{K} I_{\mathrm{GMI}}(\boldsymbol{G}) = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \big(\log\big(1+G(\omega)\big) + \hat{M}(\omega)(1+G(\omega))\big)\mathrm{d}\omega \tag{72}$$

where according to (44),

$$\begin{aligned}\hat{M}(\omega) \;=\; & 2\mathcal{R}\big\{\alpha\hat{W}(\omega)H(\omega)\hat{C}^*(\omega) + \hat{W}(\omega)H(\omega) \\ & -\alpha\hat{C}^*(\omega)\big\} - \frac{|\hat{W}(\omega)|^2}{N_0+|H(\omega)|^2} - \alpha|\hat{C}(\omega)|^2 - 1.\end{aligned}$$

Replacing $M(\omega)$ by $\hat{M}(\omega)$, the optimal $G(\omega)$ and asymptotic rate $\bar{I}$ follow from Theorem 3.

Remark 3. *Proposition 8 also holds for Method III with* $\nu_{\mathrm{R}} = \nu$*, due to the fact that* $[(\boldsymbol{I} + \boldsymbol{G})(\hat{\boldsymbol{W}}\boldsymbol{H} - \hat{\boldsymbol{C}})]_{\setminus 2\nu} = \boldsymbol{0}$.

6 SNR Asymptotics

In this section, we analyze asymptotic properties of the CS demodulators, and show that, as N_0 goes to 0 and ∞, Method III and Method II are asymptotically equivalent. As Method I is inferior to Method II in GMI sense, we limit our investigations to Method II and Method III, and start the analysis for finite length linear vector channels first. The

following limits can be verified straightforwardly:

$$\lim_{N_0\to 0} \boldsymbol{M}/N_0 = -(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1},$$
$$\lim_{N_0\to \infty} N_0(\boldsymbol{I}+\boldsymbol{M}) = \boldsymbol{H}^\dagger\boldsymbol{H}. \tag{73}$$

Moreover, it also holds that

$$\lim_{N_0\to 0} \tilde{\boldsymbol{M}} = \boldsymbol{P}^2 - \boldsymbol{P},$$
$$\lim_{N_0\to \infty} \tilde{\boldsymbol{M}} = -\boldsymbol{P}. \tag{74}$$

As $\tilde{\boldsymbol{M}}$ should be invertible from the definition of $\delta_2(\boldsymbol{G})$ in (30), we restrict that $\boldsymbol{P} \prec \boldsymbol{I}$.

Lemma 3. *When $N_0 \to 0$ and ∞, the optimal $\boldsymbol{G}$ for (28) in Method II satisfies (22), and the following limits hold,*

$$\lim_{N_0\to 0} \left[(N_0(\boldsymbol{I}+\boldsymbol{G}_{\rm opt}))^{-1}\right]_\nu = [(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_\nu, \tag{75}$$
$$\lim_{N_0\to \infty} [N_0\boldsymbol{G}_{\rm opt}]_\nu = [\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu. \tag{76}$$

Proof. When $\boldsymbol{P} = \boldsymbol{0}$, from Theorem 2 the optimal $\boldsymbol{G}$ for (28) satisfies (22). From (73), when $N_0 \to 0$, $\boldsymbol{M} \to \boldsymbol{0}$ and $N_0 \to \infty$, $\boldsymbol{M} \to -\boldsymbol{I}$. Therefore, by the definition of $\boldsymbol{\Omega}$,

$$\lim_{N_0\to 0,\infty} \boldsymbol{d} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}) = \boldsymbol{0}.$$

This implies that the gradient $d_{\boldsymbol{G}}(\delta_2)$ in (90) (Appendix F) converges to zero. Hence the differentials of $I_{\rm GMI}(\boldsymbol{V}_{\rm opt}, \boldsymbol{R}_{\rm opt}, \boldsymbol{G})$ in (28) with $\boldsymbol{P} \neq \boldsymbol{0}$ converges to the differentials with $\boldsymbol{P} = \boldsymbol{0}$. From (22) and (73), the limit (75) follows.

Next, since

$$\lim_{N_0\to \infty} \left[N_0\left(\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}\right)\right]_\nu = \lim_{N_0\to \infty} [N_0(\boldsymbol{I}+\boldsymbol{M})]_\nu = [\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu, \tag{77}$$

it shows that $\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{G}_{\rm opt})^{-1} \to \boldsymbol{0}$[7] as $N_0 \to \infty$. By the matrix inversion lemma, $\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{G}_{\rm opt})^{-1} \to \boldsymbol{G}_{\rm opt}$ as $N_0 \to \infty$, and combining this with (77) proves the limit (76). □

Lemma 4. *In Method II, with the optimal $\boldsymbol{G}$, when $N_0 \to 0$ the GMI increment $\delta_2(\boldsymbol{G})$ in (30) converges to zero with speed $\mathcal{O}(1/N_0)$[8] and when $N_0 \to \infty$ the GMI increment $\delta_2(\boldsymbol{G})$ converges to zero with speed $\mathcal{O}(N_0^2)$.*

[7] A matrix $\boldsymbol{A} \to \boldsymbol{B}$ or a vector $\boldsymbol{a} \to \boldsymbol{b}$ means the nonzero elements of $\boldsymbol{A}-\boldsymbol{B}$ or $\boldsymbol{a}-\boldsymbol{b}$ converges to zero.

[8] Two scalars A and B as functions of a variable n converging to each other with speed $\mathcal{O}(n)$ means that, there exists a constant C such that $\lim_{n\to\infty} n|A - B| < C$.

Proof. As $N_0 \to 0$, from (73) we have

$$\lim_{N_0\to 0} \boldsymbol{d}/N_0 = \lim_{N_0\to 0} \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{MP}/N_0) = -\boldsymbol{\Omega}\mathrm{vec}\big((\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\boldsymbol{P}\big).$$

Based on (74) and Lemma 3, the below equalities hold,

$$\delta_2(\boldsymbol{G}_{\mathrm{opt}}) = N_0 \frac{\boldsymbol{d}^\dagger}{N_0}\Big(\boldsymbol{\Omega}\Big(\tilde{\boldsymbol{M}}^* \otimes \frac{(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}}{N_0}\Big)\boldsymbol{\Omega}^{\mathrm{T}}\Big)^{-1}\frac{\boldsymbol{d}}{N_0} = \mathcal{O}(N_0).$$

On the other hand, as $N_0 \to \infty$, and by the definition of $\boldsymbol{\Omega}$, from (73) we also have

$$\lim_{N_0\to\infty} N_0\boldsymbol{d} = \lim_{N_0\to\infty} \boldsymbol{\Omega}\mathrm{vec}(N_0\boldsymbol{MP}) = \lim_{N_0\to\infty} \boldsymbol{\Omega}\mathrm{vec}\big(N_0(\boldsymbol{I}+\boldsymbol{M})\boldsymbol{P}\big) = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{H}^\dagger\boldsymbol{HP}).$$

Again utilizing (74) and Lemma 3, the below equalities hold,

$$\delta_2(\boldsymbol{G}_{\mathrm{opt}}) = \frac{1}{N_0^2}(N_0\boldsymbol{d}^\dagger)\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^* \otimes (\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}(N_0\boldsymbol{d}) = \mathcal{O}(1/N_0^2).$$

Therefore, Lemma 4 holds. □

Lemma 5. *When $N_0 \to 0$ and ∞, the optimal GMI in Method III is independent of $\boldsymbol{P}$ and converges to the optimal GMI with $\boldsymbol{P}=\boldsymbol{0}$. Moreover, (75) and (76) hold.*

The proof is given in Appendix K. Combining Lemmas 3-5, and using the fact that Method III and Method II are equivalent with $\boldsymbol{P}=\boldsymbol{0}$, we have the following Theorem 4.

Theorem 4. *Assume that $\boldsymbol{P} \prec \boldsymbol{I}$, when $N_0 \to 0$ and ∞, the optimal GMI in Method III converges to the optimal GMI in Method III with $\boldsymbol{P}=\boldsymbol{0}$. Moreover, the optimal GMI in Method II also converges to the optimal GMI in Method III with $\boldsymbol{P}=\boldsymbol{0}$, with speed $\mathcal{O}(1/N_0)$ when SNR increase and $\mathcal{O}(N_0^2)$ when SNR decreases. The optimal $\boldsymbol{G}$ for both methods has the limits (75) and (76).*

From Theorem 4 we know that, except for the case where one of the elements in the diagonal matrix $\boldsymbol{P}$ is 1, the soft feedback information becomes asymptotically insignificant for the design of the CS parameters. The reason is that, when $N_0 \to 0$, $\hat{\boldsymbol{x}}$ is overwhelmed by the noise, while when $N_0 \to \infty$, the optimal front-end filter will null out $\hat{\boldsymbol{x}}$ since the filter can perfectly reconstruct the transmitted symbols without using the side information.

Remark 4. *When $N_0 \to 0$ and ∞, the optimal CS demodulator is the EZF demodulator defined in Example 1, and the TMF defined in Example 2, respectively.*

With ISI channels, as the same constraint $\boldsymbol{P} = \alpha\boldsymbol{I} \prec \boldsymbol{I}$ shall hold, we make the restriction that $\alpha < 1$. The asymptotic properties for ISI channels are presented in Corollary 1, which is an asymptotic version of Theorem 4 when the channel matrix $\boldsymbol{H}$ and CS parameters are band-shaped Toeplitz matrices with infinite dimensions. The detailed proof is following the same analysis as for the finite linear vector channels and omitted.

Corollary 1. *Assume that* $0 \leq \alpha < 1$*, when* $N_0 \to 0$ *and* ∞*, the optimal GMI in Method III converges to the optimal GMI in Method III with* $\alpha = 0$*. Moreover, the optimal GMI in Method II also converges to the optimal GMI in Method III with* $\alpha = 0$*, with speed* $\mathcal{O}(1/N_0)$ *when SNR increase and* $\mathcal{O}(N_0^2)$ *when SNR decreases. The optimal* $\boldsymbol{G}$ *for both methods has the following asymptotic properties hold for* $|k| \leq \nu$*:*

$$\lim_{N_0 \to 0} \int_{-\pi}^{\pi} \frac{1}{N_0(1 + G_{\text{opt}}(\omega))} \exp(-jk\omega) \mathrm{d}\omega = \int_{-\pi}^{\pi} \frac{1}{|H(\omega)|^2} \exp(-jk\omega) \mathrm{d}\omega,$$
$$\lim_{N_0 \to \infty} \int_{-\pi}^{\pi} N_0 G_{\text{opt}}(\omega) \exp(-jk\omega) \mathrm{d}\omega = \int_{-\pi}^{\pi} |H(\omega)|^2 \exp(-jk\omega) \mathrm{d}\omega.$$

7 Empirical Results

In this section, we provide empirical results to show the behaviors of CS demodulators in an iterative detection and decoding receiver designs. With the considered MIMO channels, all channel elements are assumed to be independent identically distributed (IID) complex Gaussian with zero-means, and the received signal power at each receive antenna is normalized to unity. For ISI case, we test with the typical Proakis-C channel as in Fig. 5.

7.1 GMI Evaluation

We first evaluate the GMI under 5×5 MIMO channels with memory size $\nu = 1$ for all CS demodulators. We simulate 10000 channel realizations for each SNR point. The GMIs are compared with that of the static CS demodulator [18], which is equivalent to Method II with $\boldsymbol{P} = \boldsymbol{0}$. The channel capacity is also presented for comparison. The results of GMI are shown in Fig. 6. As the quality of soft information improves beyond $\boldsymbol{P} = \boldsymbol{0}$, Method II with $\nu_{\text{R}} = 0$ performs the best among all CS demodulators, as it has the most degrees of freedom (DoF) in $\boldsymbol{R}$. Method II with $\nu_{\text{R}} = \nu$ is the worst among Method I and Method II, while Method I is slightly worse than Method II with $\boldsymbol{R}$ of shape (a) in Fig. 3, which is because although the IC matrix $\boldsymbol{R}$ is shape (a) in both cases, $\boldsymbol{R}$ in Method II is more general than in Method I which is constrained to $\boldsymbol{R} = \boldsymbol{F}^{\dagger}\boldsymbol{T}$. The GMI of Method III is inferior to Method II as expected.

The results show consistent GMI increments for all CS demodulators when the feedback quality improves. When $\boldsymbol{P}$ increases from $\boldsymbol{P} = \boldsymbol{0}$ to the ideal case $\boldsymbol{P} = \boldsymbol{I}$, the channel capacity becomes inferior to the GMI as the pair $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ is superior to $(\boldsymbol{x}, \boldsymbol{y})$ for information transfer.

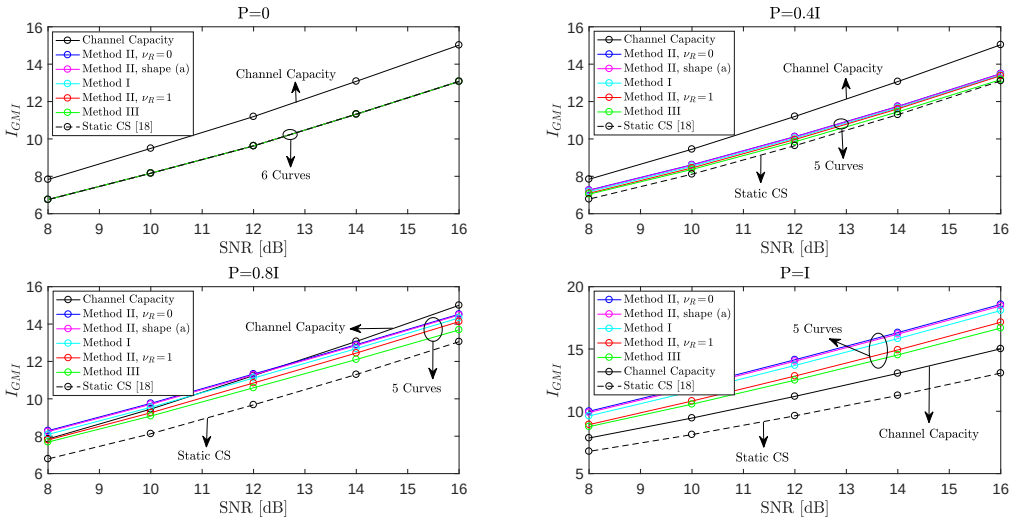


Figure 6: GMI of CS demodulators under 5×5 MIMO IID complex Gaussian channels with $\nu=1$.

7.2 SNR Asymptotic of the GMI

Next, we evaluate the asymptotic properties of the GMI described in Theorem 4 under 5×5 MIMO channels. As shown in Fig. 7, the GMIs of Method II and Method III both converge to Method III with $\boldsymbol{P}=\mathbf{0}$. Moreover, the GMI of the CS demodulators converge to the EZF in Example 1 at high SNR, and the TMF in Example 2 at low SNR, respectively, which are well aligned with Theorem 4.

7.3 EXIT Charts of CS Demodulators

In order to predict the dynamics of iterative receivers, we use the tool of extrinsic information transfer (EXIT) charts invented by ten Brink [59, 60] for analysis of iterative receiver behavior. For EXIT analysis, the CS demodulator and the decoder measure the output extrinsic information I_E based on a sequence of observations $\boldsymbol{y}$ and *a priori* information I_A into a new sequences.

In Fig. 8, we evaluate the EXIT charts for CS demodulators under 4×6 MIMO channels with $\nu=2$ for $\boldsymbol{F}$ and $\boldsymbol{G}$ at an SNR of 10dB. With Method II, we test different values of ν_{R}. As can be seen, when $\nu_{\mathrm{R}}>\nu$, the demodulation performance is inferior to $\nu_{\mathrm{R}}\leq\nu$. This is because, the interference outside memory size ν and inside memory size ν_{R} is neither considered in the IC process nor in the BCJR module. Moreover, with $\nu_{\mathrm{R}}\leq\nu$, the CS demodulators with Method II performs quite close to each other as well as Method I and III. For Method II with $\nu_{\mathrm{R}}<\nu$, the interference inside memory size ν and outside ν_{R} are considered both in the IC and BCJR processes. However, an interesting observation is that, with large *a priori* input I_A, Method II with $\nu_{\mathrm{R}}=0$ is inferior to $\nu_{\mathrm{R}}=1$ and

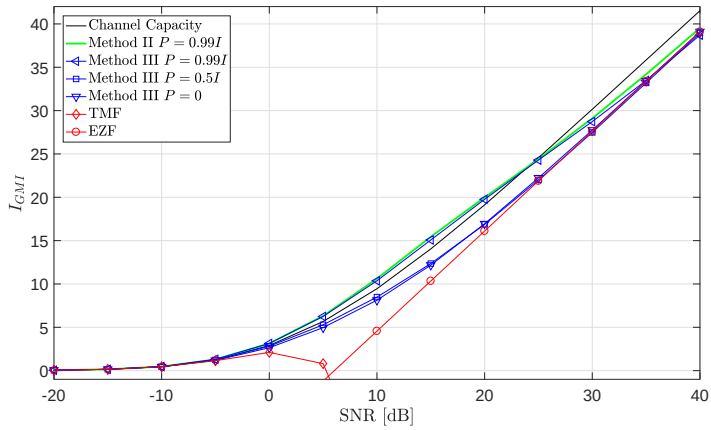


Figure 7: SNR asymptotic under 5×5 MIMO IID complex Gaussian channels with $\nu = 1$.

$\nu_{\mathrm{R}} = 2$. Therefore, a conservative approach with Method II is to set $\nu_{\mathrm{R}} = \nu$ such that the interference is either removed in IC process or dealt with in the BCJR module, to get rid of potential error propagation caused by redundant processings of the same part of interference.

In Fig. 9, we show the iterative detection and decoding trajectories for CS demodulators under Proakis-C channel with $\nu = 2$ and at an SNR of 10dB. We use an [7, 5] convolutional code [51] with a coded block-length $K = 2004$, and a random permutation is applied to the coded bits. As can be seen, the CS demodulators with Method II and Method III are superior to the LMMSE-PIC demodulator, and the iterative detection and decoding trajectories are well aligned with the measured EXIT charts.

7.4 Link Performance

We next turn to link-level simulations with turbo codes [44] where the outer decoder uses 8 internal iterations. A single code-block over all transmit symbols is used. At each SNR point 20000 data blocks are simulated and the block-error-ratio (BLER) is measured. In all simulations, at most three global iterations are used between the demodulators, and the decoder the tap length of the front-end and IC filters are all set to $8L$, and $\nu_{\mathrm{R}} = \nu$ for Method II.

In Fig. 10, we evaluate the BLER under Proakis-C channel with QPSK symbols and $\nu = 2$ for all CS demodulators. A (1064, 1600) turbo code is used. Note that, at the first iteration when there is no soft information, Method II and III overlap with each other. With CS demodulators, the gap to the MAP demodulator is less than 1 dB, while the LMMSE-PIC has a gap to the MAP that is up to 10 dB. Moreover, Method II performs slightly better than Method I, and Method III is slightly inferior to both methods. However, Method III

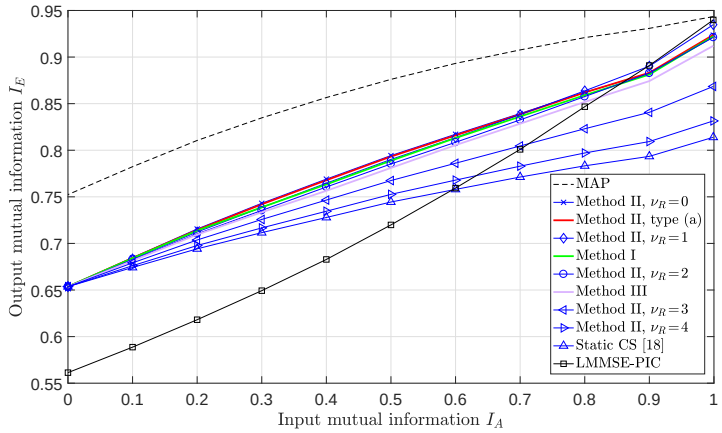


Figure 8: EXIT charts under 4×6 IID complex Gaussian MIMO channels with $\nu=2$ and different values of ν_{R}.

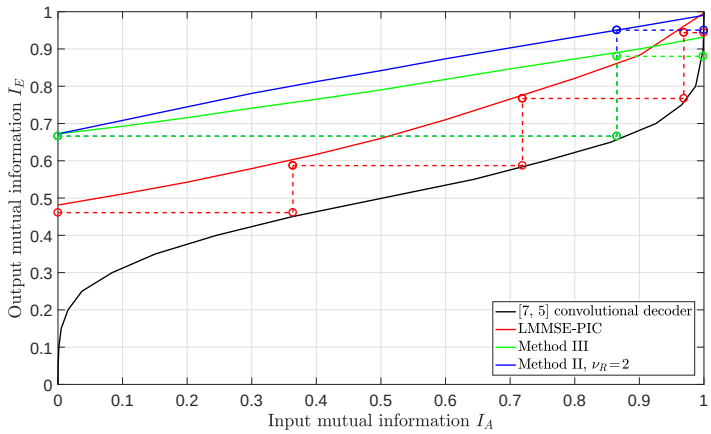


Figure 9: Iterative detection and decoding trajectories under Proakis-C channel at an SNR of 10 dB. The outer code is an [7, 5] convolutional code with generator polynomials $g_0(D)=1+D^2$ and $g_1(D)=1+D+D^2$. A random permutation of the code block is used and the black curve is the decoding EXIT chart. The dashed lines are the iterative detection and decoding trajectories for LMMSE-PIC, Method III and II, respectively.

has the advantage of less computational complexity than the others since all parameters are in closed-forms.

In Fig. 11, we evaluate the BLER under 4×6 MIMO channels with QPSK symbols and $\nu=3$ for all CS demodulators. A (1064, 1800) turbo code is used. As $N<K$, the LMMSE-PIC fails [46] at the first iteration due to the lack of receive diversity. However, the CS demodulators with $\nu=3$ significantly improve the performance and with less than 1 dB gap at 10% BLER to the MAP. CS demodulators with $\nu=1$ after three iterations is less than 2 dB away from the MAP. With less computational cost, Method III still performs close to Method II.

Finally we remark that, for the sake of complexity savings, both for finite linear vector channels and ISI channels, the parameters of CS demodulators do not need to be updated

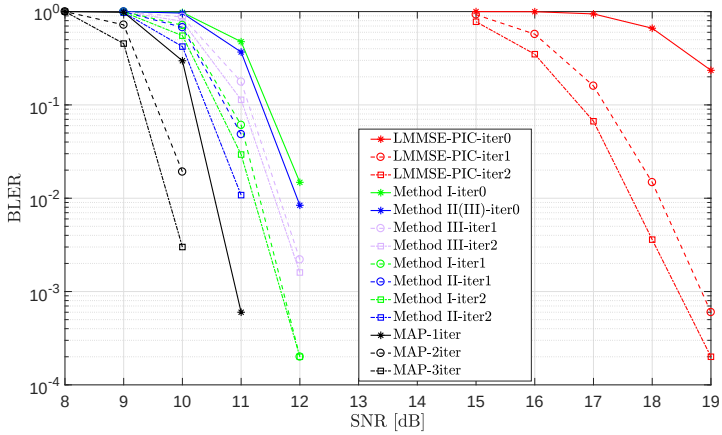


Figure 10: BLER performance of the LMMSE-PIC, Method I-III, and MAP under Proakis-C channel with QPSK modulation.

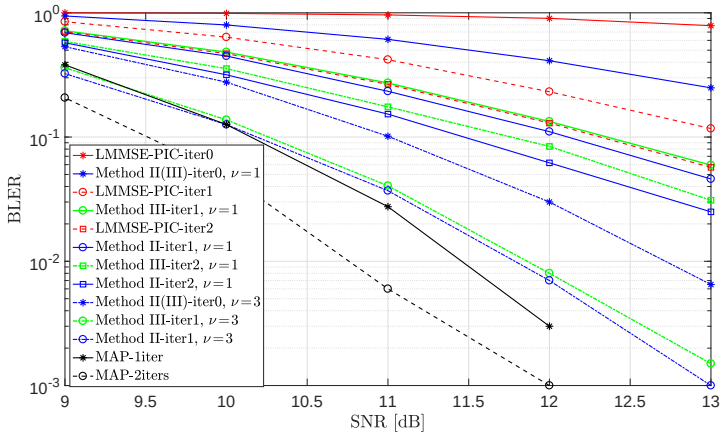


Figure 11: BLER performance of the LMMSE-PIC, Method II, Method III, and MAP under 4×6 MIMO channels with QPSK modulation.

through all iterations. Once the feedback information quality is good enough and the parameter $\boldsymbol{P}$ or α are close to ideal, the CS parameters can be kept unchanged in successive iterations.

8 Summary

In this paper we considered the design of channel shortening (CS) demodulators for linear vector channels that use a trellis representation of the received signal in combination with interference cancellation (IC) of the signal part that is not appropriately modeled by the trellis. In order to reach a trellis representation, a linear filter is applied as front-end. It is an extension of the well studied CS demodulators to iterative receivers and a generalization

of the LMMSE-PIC demodulator to cooperate with trellis-search in turbo equalization.

We analyzed the properties of three different approaches for designing such optimal CS demodulators as all of them may come across as natural "CS" demodulators. In the used framework, there are three parameters that need to be optimized. Based on a generalized mutual information (GMI) cost function, two of these are solved for in closed-form, while the third needs to be numerically optimized except for Method III where we constructed it explicitly at the cost of a small performance loss. A simple gradient based optimization is used and turns out to perform well.

Numerical results are provided to illustrate the behavior of the proposed CS demodulators. In general, Method II based on the Ungerboeck model is superior to Method I that is based on the Forney model. Method II has the advantage over Method I that the optimization procedure is concave. The suboptimal Method III performs close to Method I and Method II, and it has all parameters in closed-forms. An interesting result is that the interference cancellation of the CS demodulators should not cancel the effective channel perfectly outside the memory size ν, a property that cannot be seem in LMMSE-PIC demodulator as $\nu=0$. Moreover, we have also analyzed asymptotic properties of the CS demodulators and showed that, Method III converges to Method II asymptotically when the noise density goes to zero or infinity.

Appendix A: Derivation of the GMI

By making the eigenvalue decomposition $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\dagger = \boldsymbol{G}$ and letting $\boldsymbol{s} = \boldsymbol{Q}^\dagger\boldsymbol{x}$. As $\boldsymbol{x}$ is assumed to be zero mean complex Gaussian random vector with covariance matrix $\boldsymbol{I}$, we can write $\tilde{p}(\boldsymbol{y}|\boldsymbol{x})$ in (8) as

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(2\mathcal{R}\big\{\boldsymbol{s}^\dagger\boldsymbol{d}\big\} - \boldsymbol{s}^\dagger\boldsymbol{\Lambda}\boldsymbol{s}\big), \tag{78}$$

where $\boldsymbol{d}=\boldsymbol{Q}^\dagger(\boldsymbol{V}\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{x}})$. We can now evaluate

$$\begin{aligned}
\tilde{p}(\boldsymbol{y}) &= \int \tilde{p}(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{\pi^K}\int \exp\big(2\mathcal{R}\big\{\boldsymbol{s}^\dagger\boldsymbol{d}\big\} - \boldsymbol{s}^\dagger\boldsymbol{\Lambda}\boldsymbol{s}\big)\exp\big(-\boldsymbol{s}^\dagger\boldsymbol{s}\big)\mathrm{d}\boldsymbol{s} \\
&= \prod_{k=0}^{K-1}\frac{1}{1+\lambda_k}\exp\left(\frac{|d_k|^2}{1+\lambda_k}\right).
\end{aligned}$$

where λ_k is the kth diagonal element of $\boldsymbol{\Lambda}$ and d_k is the kth entry of $\boldsymbol{d}$. Taking the expectation over $\boldsymbol{y}$ gives

$$-\mathbb{E}_{p(\boldsymbol{y})}[\log(\tilde{p}(\boldsymbol{y}))] = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) - \mathrm{Tr}\big(\boldsymbol{L}(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)$$

where the matrix $\boldsymbol{L}=\mathbb{E}_{p(\boldsymbol{y})}[\boldsymbol{Q}\boldsymbol{d}\boldsymbol{d}^{\dagger}\boldsymbol{Q}^{\dagger}]$ is given by

$$\boldsymbol{L}=\boldsymbol{V}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\dagger})\boldsymbol{V}^{\dagger}-\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^{\dagger}-\boldsymbol{R}\boldsymbol{P}\boldsymbol{H}^{\dagger}\boldsymbol{V}^{\dagger}+\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^{\dagger}. \tag{79}$$

On the other hand, we have

$$-\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{x})}\left[\log(\tilde{p}(\boldsymbol{y}|\boldsymbol{x}))\right]=\mathrm{Tr}(\boldsymbol{G})-2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}.$$

Combining the two expectations, the GMI reads,

$$\begin{aligned}I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})=&\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)-\mathrm{Tr}\big(\boldsymbol{L}(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)-\mathrm{Tr}(\boldsymbol{G})+2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}\\ =&\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)-\mathrm{Tr}(\boldsymbol{G})+2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}\\ &-\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big(\boldsymbol{V}[\boldsymbol{H}\boldsymbol{H}^{\dagger}+N_0\boldsymbol{I}]\boldsymbol{V}^{\dagger}-2\mathcal{R}\big\{\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^{\dagger}\big\}+\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^{\dagger}\big)\big).\end{aligned}$$

Appendix B: The Proof of Proposition 1

As the formula of GMI in (13) is quadratic in $\boldsymbol{W}$ and no constraints apply to $\boldsymbol{W}$, taking the gradient of $I_{\mathrm{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$ with respect to $\boldsymbol{W}$ and setting it to zero, the optimal $\boldsymbol{W}$ is given in (16). Inserting $\boldsymbol{W}_{\mathrm{opt}}$ into (13) gives, after some manipulations,

$$\begin{aligned}I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T},\boldsymbol{F})\;=\;&K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})\big)+\mathrm{Tr}\big(\boldsymbol{T}^{\dagger}\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})^{-1}\boldsymbol{F}^{\dagger}\boldsymbol{T}\tilde{\boldsymbol{M}}\big)\\ &+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})\big)+2\mathcal{R}\big\{\mathrm{Tr}\big(\boldsymbol{P}\boldsymbol{M}\boldsymbol{F}^{\dagger}\boldsymbol{T}\big)\big\}.\end{aligned} \tag{80}$$

where $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are defined in (14) and (15). If $\boldsymbol{P}=\boldsymbol{0}$, (80) equals

$$I_1(\boldsymbol{F})=K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})\big).$$

In this case, there is no soft information available and the matrix $\boldsymbol{T}$ is not included in the formula. When $\boldsymbol{P}\neq\boldsymbol{0}$, the terms of I_{GMI} in (80) related to $\boldsymbol{T}$ are

$$f(\boldsymbol{T})=\mathrm{Tr}\big(\boldsymbol{T}^{\dagger}\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\dagger}\boldsymbol{F})^{-1}\boldsymbol{F}^{\dagger}\boldsymbol{T}\tilde{\boldsymbol{M}}\big)+2\mathcal{R}\big\{\mathrm{Tr}\big(\boldsymbol{P}\boldsymbol{M}\boldsymbol{F}^{\dagger}\boldsymbol{T}\big)\big\}.$$

Let $\boldsymbol{t}_k$ denote the kth column of $\boldsymbol{T}$, but all elements in rows $[k,\min(k+\nu,K-1)]$ removed, and define the column vector $\boldsymbol{t}=[\,\boldsymbol{t}_0^{\mathrm{T}}\;\boldsymbol{t}_1^{\mathrm{T}}\;\ldots\;\boldsymbol{t}_{K-1}^{\mathrm{T}}\,]^{\mathrm{T}}$, then by the definition of the indication matrix $\boldsymbol{\Omega}$, we have

$$\boldsymbol{t}=\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}).$$

Similarly, let $\boldsymbol{z}_k$ denote the kth column of the matrix $\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}$ but with all elements in rows $[k,\min(k+\nu,K-1)]$ removed, and define a row vector $\boldsymbol{z}=[\,\boldsymbol{z}_0^{\mathrm{T}}\;\boldsymbol{z}_1^{\mathrm{T}}\;\ldots\;\boldsymbol{z}_{K-1}^{\mathrm{T}}\,]^{\mathrm{T}}$, then we have

$$\boldsymbol{z}=\Omega\mathrm{vec}(\boldsymbol{F}\boldsymbol{M}\boldsymbol{P})=\boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^{*})\otimes\boldsymbol{I}_K\big)\mathrm{vec}(\boldsymbol{F}).$$

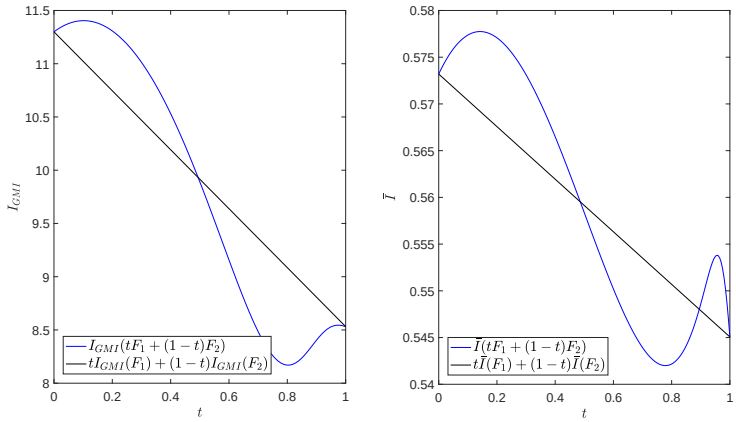


Figure 12: Non-concaveness of Method I under 5×5 MIMO channel (left figure) and Proakis-C ISI channel (right figure).

Finally, defining a Hermitian matrix $\hat{\boldsymbol{B}}_1=\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}$, and with that we can rewrite $f(\boldsymbol{T})$ as $f(\boldsymbol{T})=\boldsymbol{t}^\dagger\hat{\boldsymbol{B}}_1\boldsymbol{t}+2\mathcal{R}\{\boldsymbol{z}^\dagger\boldsymbol{t}\}$. Taking the gradient with respect to $\boldsymbol{t}$ and setting it to zero yields,

$$\boldsymbol{t}_{\mathrm{opt}}=-\hat{\boldsymbol{B}}_1^{-1}\boldsymbol{z}. \tag{81}$$

Transferring $\boldsymbol{t}_{\mathrm{opt}}$ back into $\boldsymbol{T}_{\mathrm{opt}}$ given the optimal $\boldsymbol{T}$ in (81) and inserting this into $f(\boldsymbol{T})$ gives

$$f(\boldsymbol{T}_{\mathrm{opt}})=-\boldsymbol{z}^\dagger\hat{\boldsymbol{B}}_1^{-1}\boldsymbol{z}.$$

Thus, with the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, when $\boldsymbol{P}\neq\boldsymbol{0}$ the GMI equals

$$\begin{aligned}I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F})=K&+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big)\\&-\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\Big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}\Big)^{-1}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}).\end{aligned}$$

where $\boldsymbol{D}=\boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^*)\otimes\boldsymbol{I}_K\big)$.

Appendix C: Non-Concavity Examples of Method I

We give examples to demonstrate the non-concavity of Method I for MIMO and ISI channels with assuming that $\boldsymbol{P}=\boldsymbol{I}$ and $\alpha=1$, respectively. The memory size $\nu=1$ and the noise density N_0 equals 1 in both cases. A 5×5 MIMO channel and the Proakis-C channel are used.

Example 4. *MIMO case:*

$$\boldsymbol{H}=\begin{bmatrix} 2 & 0 & -3 & 5 & 4 \\ -5 & 2 & -1 & 0 & 2 \\ 2 & -4 & 3 & 3 & 3 \\ -1 & -5 & -4 & 1 & 2 \\ 0 & -2 & 0 & 5 & 5 \end{bmatrix}, \boldsymbol{F}_1=\begin{bmatrix} 4.94 & 4.45 & 0 & 0 & 0 \\ 0 & 0.21 & 3.85 & 0 & 0 \\ 0 & 0 & 5.56 & 1.76 & 0 \\ 0 & 0 & 0 & 0.61 & 7.10 \\ 0 & 0 & 0 & 0 & 2.79 \end{bmatrix}, \boldsymbol{F}_2=\begin{bmatrix} 2.03 & 6.17 & 0 & 0 & 0 \\ 0 & 5.22 & 3.56 & 0 & 0 \\ 0 & 0 & 7.43 & 0.73 & 0 \\ 0 & 0 & 0 & 4.98 & 4.32 \\ 0 & 0 & 0 & 0 & 10.11 \end{bmatrix}.$$

Example 5. *ISI case:*

$$\boldsymbol{h}=\begin{bmatrix} 0.227 & 0.460 & 0.688 & 0.460 & 0.227 \end{bmatrix}, \boldsymbol{f}_1=\begin{bmatrix} 0.1606 & 0.9009 \end{bmatrix}, \boldsymbol{f}_2=\begin{bmatrix} 0.2230 & 0.2035 \end{bmatrix}.$$

The $I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F})$ given in (18) as a function $\boldsymbol{F}$ is plotted on the left in Fig. 12, while the $\bar{I}(W_{\mathrm{opt}}(\omega), T_{\mathrm{opt}}(\omega), F(\omega))$ given in (57) as a function of $F(\omega)$ is plotted on the right. If $I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F})$ and $\bar{I}(W_{\mathrm{opt}}(\omega), T_{\mathrm{opt}}(\omega), F(\omega))$ are concave or convex, the blue curves lie above or below the black curves, which clearly does not hold in our examples.

Appendix D: The Gradient in Method I for Finite Linear Vector Channel

In this section we derive the first order differential of the GMI given in (18) with respect to $\boldsymbol{F}$. In order to utilize the differential with respect to a matrix, we use the α-differential as defined in [47]. Assume a matrix $\boldsymbol{Y}_{N,K}$ with dimension $N \times K$ and a matrix $\boldsymbol{X}_{M,S}$ with dimension $M \times S$, define $d_{\boldsymbol{X}}\boldsymbol{Y}$ as the α-differential of $\boldsymbol{Y}$ with respect to $\boldsymbol{X}$. Furthermore, defining y_ℓ and x_ℓ as $\begin{bmatrix} y_1 & y_2 & \cdots & y_{NK} \end{bmatrix} = \mathrm{vec}(\boldsymbol{Y})^{\mathrm{T}}$ and $\begin{bmatrix} x_1 & x_2 & \cdots & x_{MS} \end{bmatrix} = \mathrm{vec}(\boldsymbol{X})^{\mathrm{T}}$, the α-differential $d_{\boldsymbol{X}}\boldsymbol{Y}$ is

$$d_{\boldsymbol{X}}\boldsymbol{Y} = \frac{\partial \mathrm{vec}(\boldsymbol{Y})}{\partial \mathrm{vec}(\boldsymbol{X})^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_{MS}} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_{MS}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{NK}}{\partial x_1} & \frac{\partial y_{NK}}{\partial x_2} & \cdots & \frac{\partial y_{NK}}{\partial x_{MS}} \end{bmatrix}.$$

The reason for adopting the α-differential is because it keeps the chain rule and the product rule. We introduce an $NK \times NK$ permutation matrix $\boldsymbol{Z}_{N,K}$, which satisfies the condition $\mathrm{vec}(\boldsymbol{Y}^{\mathrm{T}}) = \boldsymbol{Z}_{N,K}\mathrm{vec}(\boldsymbol{Y})$. It is easy to verify that $\boldsymbol{Z}_{N,K}^{-1} = \boldsymbol{Z}_{K,N}$, and when $N = 1$ or $K = 1$, $\boldsymbol{Y}$ is a vector and $\mathrm{vec}(\boldsymbol{Y}^{\mathrm{T}}) = \mathrm{vec}(\boldsymbol{Y})$, hence $\boldsymbol{Z}_{N,1} = \boldsymbol{I}_N$ and $\boldsymbol{Z}_{1,K} = \boldsymbol{I}_K$. Furthermore, by definition we have $d_{\boldsymbol{F}}(\boldsymbol{F}) = d_{\boldsymbol{F}}(\mathrm{vec}(\boldsymbol{F})) = \boldsymbol{I}$, and $d_{\boldsymbol{F}}(\boldsymbol{F}^{\dagger}) = d_{\boldsymbol{F}}\left(\mathrm{vec}(\boldsymbol{F}^{\dagger})\right) = \boldsymbol{0}$. We start by reviewing a few properties [47, 48] of α-differential below

that will be used later, where both matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are functions of $\boldsymbol{F}$ and the dimensions are specified by the subscripts associated to them:

$$\begin{aligned}
&d_{\boldsymbol{F}}(\boldsymbol{X}_{K,K}^{-1}) = -(\boldsymbol{X}_{K,K}^{-\mathrm{T}} \otimes \boldsymbol{X}_{K,K}^{-1}) d_{\boldsymbol{F}} \boldsymbol{X}_{K,K} \\
&d_{\boldsymbol{F}}(\boldsymbol{Y}_{N,K}\boldsymbol{X}_{K,S}) = (\boldsymbol{X}_{K,S}^{\mathrm{T}} \otimes \boldsymbol{I}_N) d_{\boldsymbol{F}} \boldsymbol{Y}_{N,K} + (\boldsymbol{I}_S \otimes \boldsymbol{Y}_{N,K}) d_{\boldsymbol{F}} \boldsymbol{X}_{K,S} \\
&d_{\boldsymbol{F}}(\log(\det(\boldsymbol{X}_{K,K}))) = \mathrm{vec}(\boldsymbol{X}_{K,K}^{-\mathrm{T}})^{\mathrm{T}} d_{\boldsymbol{F}} \boldsymbol{X}_{K,K} \\
&d_{\boldsymbol{F}}(\boldsymbol{Y}_{N,K} \otimes \boldsymbol{X}_{M,S}) = (\boldsymbol{I}_K \otimes \boldsymbol{Z}_{S,N} \otimes \boldsymbol{I}_M)(\boldsymbol{I}_{NK} \otimes \mathrm{vec}(\boldsymbol{X}) d_{\boldsymbol{F}} \boldsymbol{Y}_{N,K} \\
&\qquad\qquad + (\boldsymbol{I}_K \otimes \boldsymbol{Z}_{S,N} \otimes \boldsymbol{I}_M)(\mathrm{vec}(\boldsymbol{Y}) \otimes \boldsymbol{I}_{MS}) d_{\boldsymbol{F}} \boldsymbol{X}_{M,S}.
\end{aligned}$$

The α-differential of $I_1(\boldsymbol{F})$ with respect to $\boldsymbol{F}$ is

$$\begin{aligned}
d_{\boldsymbol{F}}(I_1) &= \mathrm{vec}((\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-\mathrm{T}})^{\mathrm{T}}(\boldsymbol{I}_K \otimes \boldsymbol{F}^\dagger) + \mathrm{vec}(\boldsymbol{F}^* \boldsymbol{M}^{\mathrm{T}})^{\mathrm{T}} \\
&= \mathrm{vec}(\boldsymbol{F}\boldsymbol{M} + \boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-\dagger})^\dagger. \qquad (82)
\end{aligned}$$

Defining a $K \times K$ matrix $\boldsymbol{B} = \boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger$ and an $S \times S$ matrix $\boldsymbol{\Pi} = \big(\boldsymbol{\Omega}(\tilde{\boldsymbol{M}}^* \otimes \boldsymbol{B})\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}$, the α-differential of $\delta_1(\boldsymbol{F})$ with respect to $\boldsymbol{F}$ is

$$d_{\boldsymbol{F}}(\delta_1) = -\mathrm{vec}(\boldsymbol{F})^\dagger \boldsymbol{D}^\dagger \left((\mathrm{vec}(\boldsymbol{F})^{\mathrm{T}} \boldsymbol{D}^{\mathrm{T}}) \otimes \boldsymbol{I}_S\right) d_{\boldsymbol{F}}(\boldsymbol{\Pi}) - \mathrm{vec}(\boldsymbol{F})^\dagger \boldsymbol{D}^\dagger \boldsymbol{\Pi} \boldsymbol{D}, \qquad (83)$$

where

$$\begin{aligned}
d_{\boldsymbol{F}}(\boldsymbol{\Pi}) &= -(\boldsymbol{\Pi}^{\mathrm{T}} \otimes \boldsymbol{\Pi}) d_{\boldsymbol{F}}(\boldsymbol{\Omega}(\tilde{\boldsymbol{M}}^* \otimes \boldsymbol{B})\boldsymbol{\Omega}^{\mathrm{T}}) \\
&= -\big((\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{\Pi}\boldsymbol{\Omega})\big)(\boldsymbol{I}_K \otimes \boldsymbol{Z}_{K,K} \otimes \boldsymbol{I}_K)(\mathrm{vec}(\tilde{\boldsymbol{M}}^*) \otimes \boldsymbol{I}_{K^2}) d_{\boldsymbol{F}} \boldsymbol{B} \qquad (84)
\end{aligned}$$

and

$$\begin{aligned}
d_{\boldsymbol{F}}(\boldsymbol{B}) &= d_{\boldsymbol{F}}\big(\boldsymbol{I} - (\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-1}\big) \\
&= \big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-\mathrm{T}}\big) \otimes \big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-1}\big)(\boldsymbol{F}^* \otimes \boldsymbol{I}_K) \\
&= \big(\boldsymbol{F}^*(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-\mathrm{T}}\big) \otimes (\boldsymbol{I}-\boldsymbol{B}). \qquad (85)
\end{aligned}$$

Then, defining a $K \times K$ matrix $\tilde{\boldsymbol{F}} = (\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger$ and a $K^4 \times K^2$ matrix

$$\boldsymbol{\Psi} = (\boldsymbol{I}_K \otimes \boldsymbol{Z}_{K,K} \otimes \boldsymbol{I}_K)\big(\mathrm{vec}(\tilde{\boldsymbol{M}}^*) \otimes \boldsymbol{I}_{K^2}\big), \qquad (86)$$

and by combing (82)-(86), we finally have when $\boldsymbol{P} \neq \boldsymbol{0}$,

$$\begin{aligned}
&d_{\boldsymbol{F}}\big(I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F})\big) \\
&= d_{\boldsymbol{F}}(I_1) + d_{\boldsymbol{F}}(\delta_1) \\
&= \mathrm{vec}\big(\boldsymbol{F}\boldsymbol{M} + \tilde{\boldsymbol{F}}^\dagger\big)^\dagger - \mathrm{vec}(\boldsymbol{F})^\dagger \boldsymbol{D}^\dagger \boldsymbol{\Pi} \boldsymbol{D} \\
&\quad + \mathrm{vec}(\boldsymbol{F})^\dagger \boldsymbol{D}^\dagger \big((\boldsymbol{\Omega}^{\mathrm{T}} \boldsymbol{\Pi} \boldsymbol{D} \mathrm{vec}(\boldsymbol{F}))^{\mathrm{T}} \otimes (\boldsymbol{\Pi}\boldsymbol{\Omega})\big) \boldsymbol{\Psi} \big(\tilde{\boldsymbol{F}}^{\mathrm{T}} \otimes (\boldsymbol{I}-\boldsymbol{B})\big).
\end{aligned}$$

Appendix E: The Proof of Proposition 3

As the formula of GMI in (12) is quadratic in $\boldsymbol{V}$ and no constraints apply to $\boldsymbol{V}$, taking the gradient of $I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$ with respect to $\boldsymbol{V}$ and setting it to zero, yields the optimal $\boldsymbol{V}$ given in (26). Inserting $\boldsymbol{V}_{\mathrm{opt}}$ into (12) gives, after some manipulations

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R},\boldsymbol{G}) \;=\; & K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{M}\boldsymbol{R})\big\} \\ & +\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}\tilde{\boldsymbol{M}}\boldsymbol{R}^{\dagger}\big) \end{aligned} \tag{87}$$

where $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are defined in (14) and (15). If $\boldsymbol{P}=\boldsymbol{0}$, (87) equals

$$I_2(\boldsymbol{G}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big).$$

When $\boldsymbol{P}\neq\boldsymbol{0}$, the terms of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R},\boldsymbol{G})$ in (87) related to $\boldsymbol{R}$ are

$$g(\boldsymbol{R})=2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{M}\boldsymbol{R})\big\}+\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}\tilde{\boldsymbol{M}}\boldsymbol{R}^{\dagger}\big).$$

Let $\boldsymbol{r}_k$ denote the kth column of $\boldsymbol{R}$, but where all elements in rows $[\max(0,k-\nu_{\mathrm{R}}),\min(k+\nu_{\mathrm{R}},K-1)]$ are removed, and define the column vector $\boldsymbol{r}=[\,\boldsymbol{r}_0^{\mathrm{T}}\;\boldsymbol{r}_1^{\mathrm{T}}\;\ldots\;\boldsymbol{r}_{K-1}^{\mathrm{T}}\,]^{\mathrm{T}}$, then we have $\boldsymbol{r}=\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{R})$. Moreover, let $\boldsymbol{d}_k$ denote the kth column of the matrix $\boldsymbol{M}\boldsymbol{P}$ but with all elements in rows $[\max(0,k-\nu_{\mathrm{R}}),\min(k+\nu_{\mathrm{R}},K-1)]$ are removed and define the vector $\boldsymbol{d}=[\,\boldsymbol{d}_0^{\mathrm{T}}\;\boldsymbol{d}_1^{\mathrm{T}}\;\ldots\;\boldsymbol{d}_{K-1}^{\mathrm{T}}\,]^{\mathrm{T}}$. From the definition of $\boldsymbol{d}$, we have $\boldsymbol{d}=\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P})$. Defining a Hermitian matrix $\hat{\boldsymbol{B}}_2$ as

$$\hat{\boldsymbol{B}}_2=\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}},$$

we can write $f(\boldsymbol{R})$ as $g(\boldsymbol{R})=\boldsymbol{r}^{\dagger}\hat{\boldsymbol{B}}_2\boldsymbol{r}+2\mathcal{R}\{\boldsymbol{d}^{\dagger}\boldsymbol{r}\}$. Therefore, the optimal $\boldsymbol{r}$ is

$$\boldsymbol{r}_{\mathrm{opt}}=-\hat{\boldsymbol{B}}_2^{-1}\boldsymbol{d}. \tag{88}$$

Transferring $\boldsymbol{r}_{\mathrm{opt}}$ back into $\boldsymbol{R}_{\mathrm{opt}}$ gives the optimal $\boldsymbol{R}$ in (27) and inserting this into $g(\boldsymbol{R})$ gives

$$g(\boldsymbol{R}_{\mathrm{opt}})=-\boldsymbol{d}^{\dagger}\hat{\boldsymbol{B}}_2^{-1}\boldsymbol{d}.$$

Thus, with the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, when $\boldsymbol{P}\neq\boldsymbol{0}$ the GMI equals

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G}) \;=\; & K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big) \\ & -\boldsymbol{d}^{\dagger}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}. \end{aligned}$$

Appendix F: The Gradient in Method II for Finite Linear Vector Channel

Now we calculate the α-differential of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G})$ given in (28) with respect to $\boldsymbol{G}$ when $\boldsymbol{P} \neq \boldsymbol{0}$. Taking the α-differential of $I_2(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ yields,

$$d_{\boldsymbol{G}}(I_2) = \mathrm{vec}((\boldsymbol{I}+\boldsymbol{G})^{-1}+\boldsymbol{M})^{\dagger}. \tag{89}$$

Define an $S \times S$ Hermitian matrix $\boldsymbol{\Phi} = \big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}$ and taking the α-differential of $\delta_2(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ yields,

$$\begin{aligned} d_{\boldsymbol{G}}(\delta_2) &= -(\boldsymbol{d}^{\mathrm{T}} \otimes \boldsymbol{d}^{\dagger}) d_{\boldsymbol{G}}(\boldsymbol{\Phi}) \\ &= (\boldsymbol{d}^{\mathrm{T}} \otimes \boldsymbol{d}^{\dagger})(\boldsymbol{\Phi}^{\mathrm{T}} \otimes \boldsymbol{\Phi}) d_{\boldsymbol{G}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big) \\ &= \big((\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}) \otimes (\boldsymbol{d}^{\dagger}\boldsymbol{\Phi})\big)(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega}) d_{\boldsymbol{G}}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \\ &= \big((\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{d}^{\dagger}\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi} d_{\boldsymbol{G}}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \\ &= -\big((\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{d}^{\dagger}\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}\big((\boldsymbol{I}+\boldsymbol{G})^{-\mathrm{T}} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \end{aligned} \tag{90}$$

where $\boldsymbol{\Psi}$ is defined in (86). Combining (89) and (90), we can obtain

$$\begin{aligned} d_{\boldsymbol{G}}\big(I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G})\big) &= d_{\boldsymbol{G}}(I_2) + d_{\boldsymbol{G}}(\delta_2) \\ &= \mathrm{vec}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}+\boldsymbol{M}\big)^{\dagger} \\ &\quad -\big((\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{d}^{\dagger}\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}\big((\boldsymbol{I}+\boldsymbol{G})^{-\mathrm{T}} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big). \end{aligned}$$

Appendix G: The Concavity Proof of Method II with Finite Linear Vector Channels

When $\boldsymbol{P} = \boldsymbol{0}$, as $\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)$ is concave [49] and $\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big)$ is linear in $\boldsymbol{G}$, the function $I_2(\boldsymbol{G})$ in (29) is concave with respect to $\boldsymbol{G}$ whenever $\boldsymbol{I}+\boldsymbol{G}$ is positive definite.

The concavity when $\boldsymbol{P} \neq \boldsymbol{0}$ can be deduced from the composition theorem in [49, Chpater 3.6]. For a positive definite matrix $\boldsymbol{X}$, $\boldsymbol{d}^{\dagger}\boldsymbol{X}^{-1}\boldsymbol{d}$ is convex and non-increasing (with respect to the generalized inequality for positive definite Hermitian matrices, see [49, 50]) for any column vector $\boldsymbol{d}$. Furthermore, since $\boldsymbol{I}+\boldsymbol{G}$ is positive definite, $(\boldsymbol{I}+\boldsymbol{G})^{-1}$ is convex. As $\tilde{\boldsymbol{M}} \prec \boldsymbol{0}$ $\boldsymbol{X} = \boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}$ is concave in $\boldsymbol{G}$. By the composition theorem, $\boldsymbol{d}^{\dagger}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes [\boldsymbol{I}+\boldsymbol{G}]^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}$ is convex, and $\delta_2(\boldsymbol{G})$ is then concave. Therefore the function $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G})$ in (28) is concave with respect to $\boldsymbol{G}$ whenever $\boldsymbol{I}+\boldsymbol{G}$ is positive definite.

Appendix H: The Proof of Proposition 5

The Fourier series associated to the Toeplitz matrix $\boldsymbol{W}$ is

$$W(\omega) = \sum_{k=-\infty}^{\infty} w_k \exp(jk\omega)\,,$$

and the differential of $\bar{I}(W(\omega), T(\omega), F(\omega))$ in (50) with respect to w_k (where ω is fixed) is

$$\begin{aligned}\frac{\partial \bar{I}}{\partial w_k} = &-\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|F(\omega)|^2\big(N_0+|H(\omega)|^2\big)W^*(\omega)}{1+|F(\omega)|^2}\exp(jk\omega)\,\mathrm{d}\omega \\ &+\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(F^*(\omega)H(\omega)+\frac{\alpha|F(\omega)|^2H(\omega)T^*(\omega)}{1+|F(\omega)|^2}\Big)\exp\big(jk\omega\big)\mathrm{d}\omega. \qquad (91)\end{aligned}$$

Since (91) should equal zero for all k, the optimal $W(\omega)$ is given in (55). Inserting $W_{\mathrm{opt}}(\omega)$ back into (50) yields,

$$\begin{aligned}\bar{I}(W_{\mathrm{opt}}(\omega), T(\omega), F(\omega)) = &1+\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)T(\omega)M(\omega)\big\}\mathrm{d}\omega \\ &+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big)+\frac{\tilde{M}(\omega)|T(\omega)F(\omega)|^2}{1+|F(\omega)|^2}+M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega. \qquad (92)\end{aligned}$$

where $M(\omega)$ and $\tilde{M}(\omega)$ are defined in (51) and (52). When $\alpha=0$, the GMI in (92) equals (58), and when $0<\alpha\leq 1$, the terms related to $T(\omega)$ in (92) are

$$f(T(\omega)) = \frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)T(\omega)M(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|T(\omega)F(\omega)|^2}{1+|F(\omega)|^2}\mathrm{d}\omega. \qquad (93)$$

As the elements of the main diagonal and the first ν lower diagonals of matrix $\boldsymbol{T}$ are constrained to zero, we define the vector $\tilde{\boldsymbol{t}}$ that specifies the Toeplitz matrix $\boldsymbol{T}$ as

$$\tilde{\boldsymbol{t}} = [\; t_{-N_{\mathrm{T}}} \;\ldots\; t_{-1}\; t_{\nu+1} \;\ldots\; t_{N_{\mathrm{T}}}\;],$$

and with $\phi(\omega)$ defined in (53), the Fourier series $T(\omega)$ with a finite tap length N_{T} is

$$T(\omega) = \sum_{-N_{\mathrm{T}}\leq k\leq N_{\mathrm{T}}, k\notin[0,\nu]} t_k \exp\big(jk\omega\big) = \tilde{\boldsymbol{t}}\phi(\omega). \qquad (94)$$

Furthermore, with $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ defined in (54), (93) can be rewritten as

$$f(T(\omega)) = \tilde{\boldsymbol{t}}\boldsymbol{\varepsilon}_2\tilde{\boldsymbol{t}}^{\dagger}+2\mathcal{R}\big\{\tilde{\boldsymbol{t}}\boldsymbol{\varepsilon}_1\big\}.$$

Therefore, the optimal $\tilde{\boldsymbol{t}}$ is

$$\tilde{\boldsymbol{t}}_{\mathrm{opt}} = -\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}. \qquad (95)$$

Putting $\tilde{\boldsymbol{t}}_{\mathrm{opt}}$ back into (92)-(94), the optimal $T(\omega)$ is given in (56) and $\bar{I}(W(\omega), T(\omega), F(\omega))$ for the optimal $W(\omega)$ and $T(\omega)$ is given in (57).

Appendix I: The Proof of Proposition 7

The Fourier series associated to the Toeplitz matrix $\boldsymbol{V}$ is $V(\omega)=\sum_{k=-\infty}^{\infty} v_k \exp(jk\omega)$ and the differential of $\bar{I}(V(\omega),R(\omega),G(\omega))$ in (62) with respect to v_k (where ω is fixed) is

$$\begin{aligned}\frac{\partial \bar{I}}{\partial v_k} &= -\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\big(N_0+|H(\omega)|^2\big)V^*(\omega)}{1+G(\omega)}\exp(jk\omega)\,\mathrm{d}\omega \\ &\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(H(\omega)+\frac{\alpha H(\omega)R^*(\omega)}{1+G(\omega)}\Big)\exp\big(jk\omega\big)\mathrm{d}\omega. \end{aligned} \tag{96}$$

Since (96) shall equal zero for all k, the optimal $V(\omega)$ is given in (65). Putting $V_{\mathrm{opt}}(\omega)$ in (65) back into (62) yields,

$$\begin{aligned}\bar{I}(V_{\mathrm{opt}}(\omega),R(\omega),G(\omega)) &= 1+\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{M(\omega)R(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big) \\ &\quad+\frac{\tilde{M}(\omega)|R(\omega)|^2}{1+G(\omega)}+M(\omega)\big(1+G(\omega)\big)\Big)\mathrm{d}\omega, \end{aligned} \tag{97}$$

where $M(\omega)$ and $\tilde{M}(\omega)$ are defined in (51) and (52). When $\alpha=0$, the GMI in (97) equals (68), and when $0<\alpha\leq 1$, the terms of $\bar{I}(V_{\mathrm{opt}}(\omega),R(\omega),G(\omega))$ related to $R(\omega)$ in (97) are

$$g(R(\omega))=\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{M(\omega)R(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|R(\omega)|^2}{1+G(\omega)}\mathrm{d}\omega. \tag{98}$$

Define the vector $\tilde{\boldsymbol{r}}$ that specifies the Toeplitz matrix $\boldsymbol{R}$ as

$$\tilde{\boldsymbol{r}}=[\ r_{-N_{\mathrm{R}}}\ \ldots\ r_{-\nu_{\mathrm{R}}-1}\ r_{\nu_{\mathrm{R}}+1}\ \ldots\ r_{N_{\mathrm{R}}}\ ],$$

and with $\psi(\omega)$ defined in (63), the Fourier series $R(\omega)$ with a finite tap length N_{R} is

$$R(\omega)=\sum_{-N_{\mathrm{R}}\leq k\leq N_{\mathrm{R}},k\notin[-\nu_{\mathrm{R}},\nu_{\mathrm{R}}]} r_k\exp\big(jk\omega\big)=\tilde{\boldsymbol{r}}\psi(\omega) \tag{99}$$

where $2\nu_{\mathrm{R}}+1$ is the band size that $\boldsymbol{R}$ is constrained to zero. With $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ defined in (64), (98) can be written as $g(R(\omega))=\tilde{\boldsymbol{r}}\boldsymbol{\zeta}_2\tilde{\boldsymbol{r}}^{\dagger}+2\mathcal{R}\{\tilde{\boldsymbol{r}}\boldsymbol{\zeta}_1\}$. Therefore, the optimal $\tilde{\boldsymbol{r}}$ is

$$\tilde{\boldsymbol{r}}_{\mathrm{opt}}=-\boldsymbol{\zeta}_1^{\dagger}\boldsymbol{\zeta}_2^{-1}. \tag{100}$$

This shows that $\tilde{\boldsymbol{r}}_{\mathrm{opt}}$ has Hermitian symmetry as $G(\omega)$, $M(\omega)$ and $\tilde{M}(\omega)$ are all real valued, thus $R_{\mathrm{opt}}(\omega)$ is real. Putting $\tilde{\boldsymbol{r}}_{\mathrm{opt}}$ back into (97)-(99), the optimal $R(\omega)$ is given in (66) and $\bar{I}(V(\omega),R(\omega),G(\omega))$ for the optimal $V(\omega)$ and $R(\omega)$ is given in (67).

Appendix J: The Concavity Proof of Method II with ISI Channels

To prove $\bar{I}(V_{\rm opt}(\omega), R_{\rm opt}(\omega), G(\omega))$ in (67) is concave with respect to $G(\omega)$, it is sufficient to prove that $\boldsymbol{\zeta}_1^\dagger \boldsymbol{\zeta}_2^{-1} \boldsymbol{\zeta}_1$ is convex with respect to $G(\omega)$. For a positive definite matrix $\boldsymbol{\zeta}_2$, $\boldsymbol{\zeta}_1^\dagger \boldsymbol{\zeta}_2^{-1} \boldsymbol{\zeta}_1$ is convex and non-increasing (with respect to a generalized inequality for positive definite Hermitian matrices) in $G(\omega)$ for any vector $\boldsymbol{\zeta}_1$ and with arbitrary finite tap length $N_{\rm R}$. As matrix $\tilde{\boldsymbol{M}}$ is negative definite, $\boldsymbol{\zeta}_2$ in (64) is concave with respect to $G(\omega)$ under the constraint that $\boldsymbol{I}+\boldsymbol{G}$ is positive definite. Hence $\boldsymbol{\zeta}_1^\dagger \boldsymbol{\zeta}_2^{-1} \boldsymbol{\zeta}_1$ is convex in $G(\omega)$ by the composition theorem [49].

Appendix K: The Proof of Lemma 5

From Theorem 2, the optimal $\boldsymbol{G}$ in Method III satisfies $[(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}]_\nu = -[\hat{\boldsymbol{M}}]_\nu$. Note that when $\boldsymbol{P}=\boldsymbol{0}$, Method III and Method II are equivalent as $\hat{\boldsymbol{M}}=\boldsymbol{M}$. Hence, in order to prove Lemma 4, it is sufficient to show that $[\hat{\boldsymbol{M}}]_\nu$ converges to $[\boldsymbol{M}]_\nu$ as $N_0 \to 0$ and ∞. When $\boldsymbol{P} \prec \boldsymbol{I}$, $\boldsymbol{C}_k$ in (38) is positive definite, and as $N_0 \to 0$,

$$\begin{aligned}\boldsymbol{H}^\dagger(\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}^\dagger+N_0\boldsymbol{I})^{-1}\boldsymbol{H} &= \boldsymbol{C}_k^{-1}(\boldsymbol{H}^\dagger\boldsymbol{H}+N_0\boldsymbol{C}_k^{-1})^{-1}\boldsymbol{H}^\dagger\boldsymbol{H} \\ &= \boldsymbol{C}_k^{-1}\big(\boldsymbol{I}-N_0\boldsymbol{C}_k^{-1}(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\big)+\mathcal{O}(N_0^2).\end{aligned}$$

Therefore with $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{C}}$ defined in (37)-(41),

$$\begin{aligned}\hat{\boldsymbol{W}}\boldsymbol{H} &= \boldsymbol{I}-N_0(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}+\mathcal{O}(N_0^2), \\ \hat{\boldsymbol{C}} &= [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\setminus\nu} = -N_0[(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_{\setminus\nu}+\mathcal{O}(N_0^2). \qquad (101)\end{aligned}$$

With (101) and $\hat{\boldsymbol{M}}$ in (44), it can be verified that $\lim_{N_0\to 0}[\hat{\boldsymbol{M}}/N_0]_\nu = -[(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_\nu$. On the other hand, when $N_0 \to \infty$, from (37)-(44) we have

$$\begin{aligned}N_0\hat{\boldsymbol{W}} &= \boldsymbol{H}^\dagger(\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}^\dagger/N_0+\boldsymbol{I})^{-1} = \boldsymbol{H}^\dagger+\mathcal{O}(1/N_0), \\ N_0\hat{\boldsymbol{C}} &= [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\setminus\nu} = [\boldsymbol{H}^\dagger\boldsymbol{H}]_{\setminus\nu}+\mathcal{O}(1/N_0), \qquad (102)\end{aligned}$$

With (102) and $\hat{\boldsymbol{M}}$ defined in (44), it can be verified that $\lim_{N_0\to\infty}[N_0(\boldsymbol{I}+\hat{\boldsymbol{M}})]_\nu = [\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu$. Hence, from (73) $[\hat{\boldsymbol{M}}]_\nu$ converges to $[\boldsymbol{M}]_\nu$ as $N_0 \to 0$ and ∞, which completes the proof.

References

[1] D. D. Falconer, and F. R. Magee, "Adaptive channel memory truncation for maximum likelihood sequence estimation," *The Bell Syst. Tech. J.*, vol.51, no. 9, pp. 1541-1562, Nov. 1973.

[2] S. A. Fredricsson, "Joint optimization of transmitter and receiver filter in digital PAM systems with a Viterbi detector," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 2, pp. 200-210, Mar. 1976.

[3] C. T. Beare, "The choice of the desired impulse response in combined linear-Viterbi algorithm equalizers," *IEEE Trans. Commun.*, vol. 26, pp. 1301-1307, 1978.

[4] N. Sundström, O. Edfors, P. Ödling, H. Eriksson, T. Koski, and P. O. Börjesson, "Combined linear-Viterbi equalizers-a comparative study and a minimax design," *Proc.* IEEE Veh. Technol. Conf. (VTC), Stockholm, Sweden, Jun. 1994, vol. 2, pp. 1263-1267.

[5] N. Al-Dhahir, and J. M. Cioffi, "Efficiently computed reduced-parameter input-aided MMSE equalizers for ML detection: A unified approach," *IEEE Trans. Inf. Theory*, vol. 42, pp. 903-915, Apr. 1996.

[6] M. A. Lagunas, A. I. Perez-Neia, and J. Vidal, "Joint beamforming and Viterbi equalizer in wireless communications," *Proc.* Asilomar Conf. Signals, Syst. & Comput. (ACSSC), Pacific Grove, (CA) , Nov.1997, vol. 1, pp. 915-919.

[7] S. A. Aldosari, S. A. Alshebeili, and A. M. Al-Sanie, "A new MSE approach for combined linear-Viterbi equalizers," *Proc.* IEEE Veh. Technol. Conf. (VTC), Tokyo, Japan, May 2000, vol. 3, pp. 1707-1711.

[8] R. Venkataramani and S. Sankaranarayanan, "Optimal channel shortening equalization for MIMO ISI channels," *Proc.* IEEE Global Telecommun. Conf. (GLOBECOM), New Orleans, (LO), Dec. 2008.

[9] A. Shaheem, *Iterative detection for wireless communications*, Ph.D. thesis, School of Electrical, Electronic and Computer Engineering, University of Western Australia, 2008.

[10] U. L. Dang, W. H. Gerstacker, and D. T. M. Slock, "Maximum SINR prefiltering for reduced state trellis based equalization," *Proc.* IEEE Int. Conf. Commun. (ICC), Kyoto, Japan, Jun. 2011.

[11] R. Venkataramani and M. F. Erden, "A posteriori equivalence: A new perspective for design of optimal channel shortening equalizers," *arXiv preprint: 0710.3802v1*.

[12] I. Abou-Faycal and A. Lapidoth, "On the capacity of reduced complexity receivers for intersymbol interference channels", *Proc.* Conf. Inf. Sciences and Systems (CISS), Princeton University, Mar. 2000, pp. WA4 32-37.

[13] G. D. Forney Jr., "Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 363-378, May 1972.

[14] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260-269, Apr. 1967.

[15] N. Merhav, G. Kaplan, A. Lapidoth and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, Nov. 1994.

[16] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, pp. 2315-2328, Nov. 2000.

[17] M. R. McKay, I. B. Collings, and A. M. Tulino, "Achievable sum rate of MIMO MMSE receivers: A general analytic framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, Jan. 2010.

[18] F. Rusek and A. Prlja, "Optimal channel shortening of MIMO and ISI channels," *IEEE Trans. Wireless Commun.,* vol. 11, no. 2, pp. 810-818, Feb. 2012.

[19] H. Weingarten, Y. Steinberg, and S. Shamai, "Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1665-1686, Aug. 2004.

[20] F. Rusek, N. Al-Dhahir, and A. Gomaa, "A rate-maximizing channel-shortening detector with soft feedback side information," *Proc.* IEEE Global Telecommun. Conf. (GLOBECOM), Anaheim, (CA), Dec., 2012, pp. 1-6.

[21] A. Duel-Hallen and C. Heegard, "Delayed decision-feedback sequence estimation," *IEEE Trans. Commun.,* vol. 37, no. 5, pp. 428-436, May 1989.

[22] J. Hagenauer, "Source-controlled channel decoding," *IEEE Trans. Commun.,* vol. 43, no. 9, pp. 2449-2457, Sep. 1995.

[23] S. M. Kay, "Fundamentals of statistical signal processing, volume I: Estimation theory," Prentice Hall, Apr. 1993.

[24] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.,* vol. 51, no. 3, pp. 389-399, Mar. 2003.

[25] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 284-287, Mar. 1974.

[26] C. Studer, S. Fateh, and D. Seethaler, "ASIC Implementation of soft-input soft-output MIMO detection using parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754-1765, Jul. 2011.

[27] M. Witzke, S. Bäro, F. Schreckenbach, and J. Hagenauer, "Iterative detection of MIMO signals with linear detectors," *Proc.* Asilomar Conf. Signals, Syst. and Comput. (ACSSC), Monterey, (CA), Nov. 2002, pp. 289-293.

[28] J. Zhang, H. Nguyen, and G. Mandyam, "LMMSE-based iterative and turbo equalization methods for CDMA downlink channels," *Proc.* IEEE 6th Workshop Signal Process. Advances Wireless Commun., Jun. 2005, pp. 231-235.

[29] F. Rusek, D. Fertonani, "Bounds on the information rate of intersymbol interference channels based on mismatched receivers," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1470-1482, 2012.

[30] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, Jul. 2003.

[31] G. L.Turin, "An introduction to digital matched filters," *Proc. IEEE*, vol. 64, pp. 1092-1112, Jul. 1972.

[32] G. Ungerboeck, "Adaptive maximum likelihood receiver for carrier-modulated data-transmission systems," *IEEE Trans. Commun.,* vol. 22, pp. 624-636, May 1974.

[33] G. H. Golub and C. F. Van Loan, "Matrix computations," third edition, Baltimore, MD: Johns Hopkins, 1996.

[34] A. Kavčić and J. M. F. Moura, "Matrix with banded inverses: algorithms and factorization of Gauss-Markov processes," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1495-1509, Jul. 2000.

[35] G. Colavolpe and A. Barbieri, "On MAP symbol detection for ISI channels using the Ungerboeck observation model," *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 720-722, Aug. 2005.

[36] F. Rusek, G. Colavolpe, and, C. Sundberg, "40 years with the Ungerboeck model: a Look at its potentialities [Lecture Notes]," *Signal Process. Mag.*, vol. 32, pp. 156-161, May 2015.

[37] F. Rusek, M. Loncar and A. Prlja, "A comparison of Ungerboeck and Forney models for reduced complexity lSI equalization," *Proc.* IEEE Global Telecommun. Conf. (GLOBECOM), Washington D.C., Dec. 2007, pp. 1431-1436.

[38] O. Edfors, M. Sandell, J. J. Van de Beek, S. K. Wilson and P. O. Borjesson, "OFDM channel estimation by singular value decomposition," *IEEE Trans. Commun.,* vol. 46, no. 7, pp. 931-939, Jul. 1998.

[39] W. Hirt, *Capacity and information rates of discrete-time channels with memory*, Ph.D thesis, no. ETH 8671, Inst. Signal and Inf. Process., Swiss Federal Inst. Technol., Zürich, 1988.

[40] U. Grenander and G. Szegő, *Toeplitz forms and their applications*, University of Calif. Press Berkeley and Los Angeles, 1958.

[41] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Commun. and Inf. Theory*, vol. 2, no. 3, pp 155-239, 2006.

[42] J. G. Proakis and M. Salehi, *Digital communications*, the fifth edition, McGraw-Hill, 2008.

[43] M. Tüchler, A. Singer and R. Kötter, "Minimum mean squared error (MMSE) equalization using priors," *IEEE Trans. Signal Process.*, vol. 50, pp. 673-683, 2000.

[44] 3GPP TS 36.212: *Evolved universal terrestrial radio access (E-UTRA); Multiplexing and channel coding*, Release 12, v12.4.0, Mar. 2015.

[45] P. Kabal and S. Pasupathy, "Partial-response signaling," *IEEE Trans. Commun.,* vol. COM-23, pp. 921-934, Sep. 1975.

[46] F. Rusek and O. Edfors, "An information theoretic characterization of channel shortening receivers," Asilomar Conf. Signals, Syst. and Comput. (ACSSC), Pacific Grove, (CA), Nov. 2013, pp. 2108-2112.

[47] J. R. Magnus, "On the concept of matrix derivative," *J. Multivariate Anal.*, vol. 101, no. 9, pp. 2200-2206, Oct. 2001.

[48] P. L. Fackler, "Notes on Matrix Calculus," North Carolina State University, Sep. 2005.

[49] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[50] C. Davis, "Notions generalizing convexity for functions defined on spaces of matrices," *Proc. Symp. Pure Math.,* vol. 7, pp. 187-201, 1963.

[51] M. Tüchler and A. C. Singer, "Turbo Equalization: An Overview," *IEEE Trans. Inf. Theory.,* vol. 57, no. 2, pp. 920-952, Feb. 2011.

[52] S. J. Lee, A. C. Singer, and N. R. Shanbhag, "Linear turbo equalization analysis via BER transfer and EXIT charts," *IEEE Trans. Signal Process.,* vol. 53, no. 8, pp. 2883-2897, Aug. 2005.

[53] S. Hu and F. Rusek, "On the design of reduced state demodulators with interference cancellation for iterative receivers," IEEE Annual Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC), Sep. 2015, pp. 981-985.

[54] A. Shaheem, H. Zepernick, and M. Caldera, "Enhanced channel shortened turbo equalization," Int. Conf. Advanced Technol. Commun. (ATC), 6-9 Oct. 2008, pp. 8-11.

[55] A. Glavieux, C. Laot, and J. Labat, "Turbo equalization over a frequency selective channel," *Proc.* Int. Symp. Turbo Codes Related Topics, Brest, France, Sep. 1997, pp. 96-102.

[56] R. Lopes and J. Barry, "The soft-feedback equalizer for turbo equalization of highly dispersive channels," *IEEE Trans. Commun.,* vol. 54, no. 5, pp. 783-788, May 2006.

[57] A. Berthet, R. Visoz, and P. Tortelier, "Sub-optimal turbo-detection for coded 8-PSK signals over ISI channels with application to EDGE advanced mobile system," *Proc.* IEEE Veh. Technol. Conf. (VTC), Sep. 2000.

[58] N. Al-Dhahir, "FIR channel-shortening equalizers for MIMO ISI channels," *IEEE Trans. Commun.*, vol. 49, pp. 213-218, Feb. 2001.

[59] S. ten Brink, "Convergence of iterative decoding," *Electron. Lett.*, vol.35, pp. 806-808, May 1999.

[60] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 40, pp. 1727-1737, Oct. 2001.

[61] D. Darsena and F. Verde, "Minimum-mean-output-energy blind adaptive channel shortening for multicarrier SIMO transceivers," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5755-5771, Dec. 2007.

[62] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "An Optimal channel shortener design for reduced-state soft-output viterbi equalizer in single-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2568-2582, Jun. 2017.

[63] J. W. Choi, B. Shim, A. C. Singer, and N. I. Cho, "Low-complexity decoding via reduced dimension maximum-likelihood search," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1780-1794, Mar. 2010.

[64] J. W. Choi, B. Lee, and B. Shim, "Iterative group detection and decoding for large MIMO systems," *J. Commun. and Netw.*, vol 17, no. 6, pp. 609-621, Dec. 2015.

[65] J. W. Choi, A. C. Singer, J. W. Lee, and N. I. Cho, "Improved linear soft-input soft-output detection via soft feedback successive interference cancellation," *IEEE Trans. Commun.*, vol. 58, no. 3, pp. 986-996, Mar. 2010.

Paper II

Optimal Channel Shortener Design for Reduced-State Soft-Output Viterbi Equalizer in Single-Carrier Systems

We consider optimal channel shortener design for reduced-state soft-output Viterbi equalizer (RS-SOVE) in single-carrier (SC) systems. To use RS-SOVE, three receiver filters need to be designed: a prefilter, a target response and a feedback filter. The collection of these three filters are commonly referred to as the "channel shortener". Conventionally, the channel shortener is designed to transform an intersymbol interference (ISI) channel into an equivalent minimum-phase equivalent form. In this paper, we design the channel shortener to maximize a mutual information lower bound (MILB) based on a mismatched detection model. By taking the decision-feedback quality in the RS-SOVE into consideration, the prefilter and feedback filter are found in closed forms, while the target response is optimized via a gradient-ascending approach with the gradient explicitly derived. The information theoretical properties of the proposed channel shortener are analyzed. Moreover, we show through numerical results that, the proposed channel shortener design achieves superior detection performance compared to previous channel shortener designs at medium and high code-rates.

S. Hu, H. Kröll, Q. Huang, and F. Rusek,
"Optimal channel shortener design for reduced-state soft-output Viterbi equalizer in single-carrier systems,"
IEEE Trans. Commun., vol. 65, no. 6, pp. 2568-2582, Jun. 2017.

1 Introduction

Communication systems based on single carrier (SC) modulation are currently used in 2G networks [1] which have the largest number of subscribers worldwide. Besides personal mobile communication they play a key role in the latest LTE-Advanced Release 13 [2], where on the path to 5G Internet of Things (IoT) networks, the standard EC-GSM-IoT was released together with SC waveforms for high power efficiency requirements [3]. Moreover, SC modulation is also used in satellite communications and high-speed serial links[4]. The advantages of a low peak-to-average-power ratio (PAPR), low device complexity, straightforward synchronization, and the absence of cyclic-prefix (CP) overhead favor its use in many low data rate scenarios over multi-carrier (MC) systems[5, 6]. However, SC systems suffer from intersymbol interference (ISI) caused by delay dispersion along the multi-path propagation from the transmitter to the receiver.

In order to combat intersymbol interference (ISI) caused by delay dispersion in propagation channels in SC systems, reduced-state based soft-output equalizers are commonly deployed, which has a long and rich history [7–14]. In 1972, Forney proposed the Viterbi algorithm (VA) [14] that implements maximum log-likelihood sequence estimation (MLSE). With error correcting codes such as turbo codes[15] and low-density parity-check (LDPC) codes[16], it is well-known that soft-decisions generated by the soft output Viterbi algorithm (SOVA)[17], i.e., the reliability information, are superior to hard-decisions. To reduce the prohibitive detection complexity of the SOVA (or BCJR [18]) algorithm, reduced-state based ISI equalizers are extensively developed which are based on techniques such as delayed-decision-feedback[7], state partitioning[11], state-merging[8, 9], and list-type Viterbi equalizer (LVE)[12, 13].

In [19], Koch and Baier proposed the soft output Viterbi equalizer (SOVE). Rather than minimizing the sequence error probability in SOVA, the SOVE uses a trellis-based algorithm that minimizes the bit error probability. To further reduce the receiver complexity, the authors in [19] also proposed the suboptimal reduced-state SOVE (RS-SOVE). Different from the SOVE whose trellis spans over all L taps, where L is the tap-length of the considered channel impulse response (CIR) $\boldsymbol{h}$, the trellis in RS-SOVE only spans the first $(\nu+1)$ taps, and the signal part corresponding to the remaining $(L-\nu-1)$ channel tails is canceled by a state-dependent decision-feedback along the detection. The RS-SOVE is simple to implement and performs nearly as good as the full-complexity SOVE. Note that, the RS-SOVE can also be reviewed as a soft-output extension of the delayed decision-feedback sequence estimation (DDFSE)[7], which combines VA and the decision-feedback detection to approximate the MLSE.

To obtain high performance, a minimum-phase CIR is essential for reduced-state ISI equalizers, and a discrete-time prefilter, which ideally has an all-pass characteristic, should be

introduced in front of equalization in order to transform the $\boldsymbol{h}$ into its minimum-phase equivalent. Therefore, in order to transform $\boldsymbol{h}$ into a new target response, which renders better performance in conjunction with the RS-SOVE, the channel shorteners are commonly utilized prior to the RS-SOVE. Due to its low-complexity, simple-implementation, and good-performance, the RS-SOVE together with channel shortener is widely used in the receiver design of devices in SC systems. A typical overview of such systems is depicted in Fig. 1. Normally, the channel shortener requires three receiver filters to be designed: a prefilter (the tap-length is up to design), a $(\nu+1)$-tap target response, and a $(L-\nu-1)$-tap feedback filter.

Traditionally, there are two types of processing schemes for designing the channel shortener, namely, the Forney detection model[20] which assumes white noise, and the Ungerboeck detection model[21] which assumes that the noise is colored according to the target response autocorrelation. A conventional design of the Forney model based channel shortener is to use an all-phase filter to transform $\boldsymbol{h}$ into the minimum-phase equivalent $\tilde{\boldsymbol{h}}$. Then, the target response is set to the first $(\nu+1)$ taps of $\tilde{\boldsymbol{h}}$, while the feedback filter is set to the remaining taps. The all-pass prefilter can be designed based on various criteria [22–25] such as linear minimum-mean-square-error (LMMSE), linear prediction, and homomorphic filtering. The authors in [25] showed that, the homomorphic filter has lower-complexity, simpler hardware-implementation, and superior performance than the other prefilter designs. We refer to such a conventional channel shortener design as the "HOM" shortener.

In [21], the Ungerboeck model based channel shortener design was developed. A prefilter $\boldsymbol{v}$ and target response $\boldsymbol{g}$ are designed to maximize a mutual information lower bound (MILB) based on a mismatched detection model. However, the feedback filter is not utilized in the detection model, which means that the $(L-\nu-1)$ channel tails are truncated directly. We refer to such a state-of-the-art design as the "UBM" shortener. As there is no feedback filter, with the UBM shortener there is no decision-feedback process in the RS-SOVE. In [1, 26], the UBM shortener was successfully implemented for GSM/EDGE systems, and showed superior detection performance, yet with a much lower complexity than the HOM shortener. However, as shown in [26], in the high signal-to-noise (SNR) regime[1], the UBM shortener suffers from performance losses and renders a bit-error-rate (BER) error floor.

In this paper, we propose a novel channel shortener design for RS-SOVE aiming to overcome the performance losses of the UBM shortener. Due to the lack of a probabilistic meaning of the branch metric definition[27, 28], the UBM shortener cannot be extended by decision-feedback using the methods introduced in [1, 21, 26] and therefore is not applicable for the RS-SOVE. Instead we show that we can overcome the performance losses of the UBM by applying the information theoretical MILB approach to the Forney model instead of the Ungerboeck model. Since we derive a Forney model equalizer

[1]In relation to higher-order modulations and code-rates, which require high SNRs to decode.

Table 1: Channel Shortener Designs and Parameter Notations

Name	prefilter	target response	feedback filter	RS-SOVE cooperates with feedback?
FOM	$\boldsymbol{w}$	$\boldsymbol{f}$	$\boldsymbol{b}$	yes
UBM	$\boldsymbol{v}$	$\boldsymbol{g}$	$\mathbf{0}$	no
HOM	$\boldsymbol{w}_{\mathrm{hom}}$	$\boldsymbol{h}_{\mathrm{f}}$	$\boldsymbol{h}_{\mathrm{b}}$	yes

that is equipped with MILB-maximization channel shortening filters, we refer to this approach as the "FOM" shortener. Note that, both the HOM and FOM shorteners adopt the same the Forney model for channel shortener designs. The difference is that, the HOM shortener is a conventional design, while the FOM shortener optimizes the receiver filters to maximize the information rate[2]. Therefore, the FOM shortener always performs better than the HOM shortener from an information-theoretical perspective. Moreover, as the UBM shortener is constrained to the case that the feedback filter equals $\mathbf{0}$, while the FOM shortener can jointly optimize all three receiver filters, the FOM shortener is superior to the UBM shortener when the feedback has good quality.

We show that although at low code-rates the UBM shortener performs better than both the HOM and FOM shorteners, it suffers from significant performance losses at medium and high code-rates. This phenomenon, however, does not exist for the FOM shortener, which outperforms the UBM shortener at medium and high code-rates, and is better than the conventional HOM shortener in all cases. These three different channel shorteners considered in this paper are listed in Table 1, with FOM shortener being the proposed channel shortener design and the remaining two are the reference designs. The main contributions of this paper are as follows.

- Firstly, we propose the FOM shortener for RS-SOVE with the MILB derived in closed form. The prefilter and feedback filter are optimized through MILB perspective and found in closed forms, and the target response utilizes a gradient-ascending optimization.
- Secondly, we analyze the optimal parameter design of the FOM channel shortener by considering the feedback quality, and show that the FOM shortener can be designed for perfect feedback. We further show that, the FOM shortener outperforms the UBM shortener at medium and high code-rates, and is superior to the HOM shortener in all cases.
- Thirdly, we analyze information-theoretic properties and information rates of the FOM shortener in relation to Shannon capacity $\mathcal{C}$ and the previous channel shortener designs of the HOM and UBM shorteners.

[2]The information rate is a bound on the rate that can be transmitted, but are not a capacity since there are constraints on the transmit signals and the decoding operations.

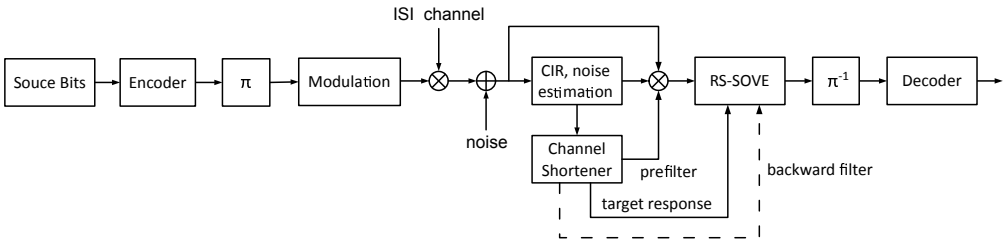


Figure 1: Discrete time transmission and receive model with the channel shortener and RS-SOVE. Note that, with the UBM shortener the feedback filter is not needed and no decision-feedback is performed in the RS-SOVE. The CIR and noise estimation can be based on, e.g., pilot symbols.

- Lastly, we extend the RS-SOVE to an arbitrary delay D, and show an interesting fact that, the trellis search process in RS-SOVE is equivalent to a full forward recursion and D-depth backward recursion.

The rest of the paper is organized as follows. In Sec. II, the received signal model, conventional HOM shortener, and RS-SOVE are introduced. In Sec. III, the proposed FOM shortener is derived, and the optimal design of the filters ($\boldsymbol{w}$, $\boldsymbol{f}$, $\boldsymbol{b}$) with feedback quality is elaborated. In Sec. IV, the links of theoretical information rates among all three channel shorteners are established. Empirical results are provided in Sec. V, and Sec. VI concludes the paper.

Notations

Throughout this paper, boldface lowercase letters indicate vectors and boldface uppercase letters designate matrices. Superscripts $(\cdot)^{-1}$, $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^\dagger$ stand for the inverse, complex conjugate, transpose, and Hermitian transpose, respectively. Furthermore, $\mathbb{E}[\cdot]$ is the expectation operator, and $\mathcal{R}\{\cdot\}$ takes the real part of the arguments. We reserve '$\star$' to denote linear convolution, $\boldsymbol{I}$ to represent an identity matrix, and $\mathrm{vec}\,(\boldsymbol{A})$ to stack the columns of $\boldsymbol{A}$ on top of each other.

2 Received Signal Model and the HOM Detector

The considered SC system that applies channel shortening and RS-SOVE is depicted in Fig. 1. With sufficiently good interleaving, we assume the transmit bits to be independent. The transmit symbols x_k have unit-energy and are drawn from a constellation $\mathcal{X}$, whose cardinality is $|\mathcal{X}|$. Considering the data transmission over a dispersive channel with additive

noise, the received sample y_k at time epoch k is modeled as

$$y_k = \sum_{\ell=0}^{L-1} h_\ell x_{k-\ell} + n_k, \tag{1}$$

where L is the ISI duration, and h_ℓ is the ℓth tap of the CIR $\boldsymbol{h} = (h_0\ h_1\ \ldots\ h_{L-1})$. The noise variables n_k are identical and independently distributed (IID) zero-mean complex Gaussian variables with variance N_0. For a transmit block comprising K symbols[3], we denote the signal vector $\boldsymbol{y}$, receive vector $\boldsymbol{x}$, and noise vector $\boldsymbol{n}$ as

$$\begin{aligned}\boldsymbol{y} &= (y_0\ y_1\ \ldots\ y_{K+L-1})^{\mathrm{T}},\\ \boldsymbol{x} &= (x_0\ x_1\ \ldots\ x_{K-1})^{\mathrm{T}},\\ \boldsymbol{n} &= (n_0\ n_1\ \ldots\ n_{K-1})^{\mathrm{T}},\end{aligned}$$

respectively. The signal model (1) that comprises $(K+L-1)$ received samples can be written as

$$\boldsymbol{y} = \boldsymbol{h} \star \boldsymbol{x} + \boldsymbol{n}, \tag{2}$$

or equivalently,

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{3}$$

where the $(K+L-1)\times K$ Toeplitz matrix $\boldsymbol{H}$ is generated from $\boldsymbol{h}$ as

$$\boldsymbol{H} = \begin{bmatrix} h_0 & & & \\ h_1 & h_0 & & \\ \vdots & h_1 & \ddots & \\ h_{L-1} & \vdots & \ddots & h_0 \\ & h_{L-1} & \ddots & h_1 \\ & & \ddots & \vdots \\ & & & h_{L-1} \end{bmatrix}. \tag{4}$$

2.1 Conventional HOM Channel Shortener

Prior to the RS-SOVE, the HOM shortener utilizes homomorphic filtering to obtain the minimum-phase equivalent form of the causal response $\boldsymbol{h}$. With the prefilter $\boldsymbol{w}_{\mathrm{hom}}$ designed based on the cepstrum of $\boldsymbol{h}$ [23], the target response $\tilde{\boldsymbol{h}} = \boldsymbol{w}_{\mathrm{hom}} \star (\boldsymbol{h}/\sqrt{N_0})$, and

[3]We assume that $L-1$ zero-symbols are inserted between continuous data blocks (i.e., the guard period) to prevent inter-block interference.

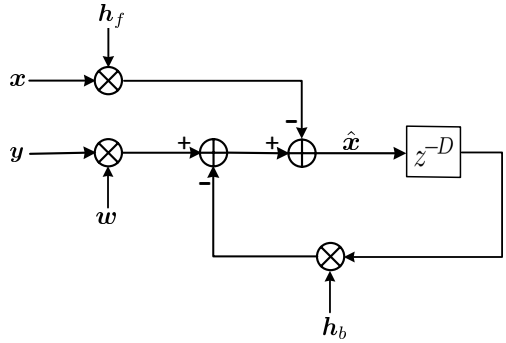


Figure 2: The decision-feedback process in RS-SOVE. The hard feedback $\hat{\boldsymbol{x}}$ is associated to each state and updated along the detection stages.

the filtered samples $\tilde{\boldsymbol{y}} = \boldsymbol{w}_{\mathrm{hom}} \star (\boldsymbol{y}/\sqrt{N_0})$, the detection model after prefiltering reads

$$\tilde{y}_k = \sum_{\ell=0}^{\nu} \tilde{h}_\ell x_{k-\ell} + \sum_{\ell=\nu+1}^{L-1} \tilde{h}_\ell x_{k-\ell} + \tilde{n}_k, \tag{5}$$

where ν denotes the memory length considered by the RS-SOVE so that its number of states becomes $|\mathcal{X}|^\nu$. Denoting

$$\boldsymbol{h}_{\mathrm{f}} = \big(\tilde{h}_0, \tilde{h}_1, \cdots, \tilde{h}_\nu\big), \tag{6}$$

$$\boldsymbol{h}_{\mathrm{b}} = \big(\underbrace{0, \cdots, 0}_{\nu+1}, \tilde{h}_{\nu+1}, \tilde{h}_{\nu+2}, \cdots, \tilde{h}_{L-1}\big), \tag{7}$$

the second term in (5) is canceled by the hard feedback $\hat{x}_\ell$ on the surviving path that leads to each state after filtered by the feedback filter $\boldsymbol{h}_{\mathrm{b}}$. By setting $\nu = 0$, the RS-SOVE becomes the decision-feedback detector, while with $\nu = L-1$, the RS-SOVE is the full-complexity SOVE.

In contrast to BCJR algorithm [18] or Max-Log-Map (MLM)[29], the backward recursions are omitted in RS-SOVE [19]. In order to improve the quality of soft-decisions, we extend the decision-delay in RS-SOVE to an arbitrary value D, which can set to be larger than L–1. As shown next, the RS-SOVE with a delay D can be viewed as the the MLM equalizer with a full forward recursion and D-step backward recursion at each detection stage. Hence, when D is sufficiently large, the RS-SOVE performs as well as MLM. Such a modification

$$\begin{aligned} L(x_{k,n}) = &\log\left(\sum_{x_{k,n}=1} \exp\left(-\alpha^i_{k+\nu-1} - \gamma^{i,j}_{k+\nu} - \beta^j_{k+\nu}\right)\right) \\ &- \log\left(\sum_{x_{k,n}=-1} \exp\left(-\alpha^i_{k+\nu-1} - \gamma^{i,j}_{k+\nu} - \beta^j_{k+\nu}\right)\right) \\ \approx &\min_{x_{k,n}=-1}\left(\alpha^i_{k+\nu-1} + \gamma^{i,j}_{k+\nu} + \beta^j_{k+\nu}\right) - \min_{x_{k,n}=1}\left(\alpha^i_{k+\nu-1} + \gamma^{i,j}_{k+\nu} + \beta^j_{k+\nu}\right). \end{aligned} \tag{8}$$

only increases the equalization latency from ν to D, and introduces a small overheard by the D-step backward recursion process in the RS-SOVE. In [30], an improvement of RS-SOVE is also proposed by introducing an expanded memory, however, the number of states is exponentially increased and results in higher memory cost.

2.2 RS-SOVE with Arbitrary Decision-Delay D

In Fig. 2, we illustrate the decision-feedback detection in the RS-SOVE with the prefilter $\boldsymbol{w}$, the feedback filter $\boldsymbol{h}_{\mathrm{b}}$, and the target response $\boldsymbol{h}_{\mathrm{f}}$. Utilizing Jacobian approximation [29],

$$\log\big(\exp(-a)+\exp(-b)\big)\approx -\min(a,b),$$

the soft-decisions of the nth bit $x_{k,n}$ in x_k, i.e., the log-likelihood ratio (LLR), is calculated according to (8) with delay ν. The forward path metric α_k^j corresponding to state j at stage k is recursively computed through

$$\alpha_k^j=\min_i\left\{\alpha_{k-1}^i+\gamma_k^{i,j}\right\}, \tag{9}$$

where the branch metric $\gamma_k^{i,j}$ in (5) associated to state transition $i\to j$ is calculated as

$$\gamma_k^{i,j}=\left|\tilde{y}_k-\sum_{\ell=0}^{\nu}\tilde{h}_\ell x_{k-\ell}-\sum_{\ell=\nu+1}^{L-1}\tilde{h}_\ell\hat{x}_{k-\ell}\right|^2. \tag{10}$$

In (10), the symbol vector $(x_k,\cdots,x_{k-\nu})$ are determined from state transition $i\to j$, while $(\hat{x}_{k-\nu-1},\cdots,\hat{x}_{k-L+1})$ are the hard decisions associated to each state i at stage k. As for each state there is a survival path that leads to it, with decision-feedback determined from such a path, the feedback varies on different states. In addition, an update of all survival paths is needed along the detection stages.

In [19], with RS-SOVE the backward recursions are omitted by setting $\beta_{k+\nu}^j=0$ for all states, and the LLR in (8) is simplified to

$$L(x_{k,n})\approx\min_{x_{k,n}=-1}\left(\alpha_{k+\nu-1}^i+\gamma_{k+\nu}^{i,j}\right)-\min_{x_{k,n}=1}\left(\alpha_{k+\nu-1}^i+\gamma_{k+\nu}^{i,j}\right).$$

However, a drawback of such an approximation is that, the short decision delay ν in RS-SOVE limits its performance, especially with higher-order modulations and code-rates[30]. Therefore, we increase the delay ν to an arbitrary value D by initializing $\beta_{k+D}^j=0$ for all states at detection stage $k+D$, and define the backward recursion for state transition $j\to i$ as

$$\beta_{k-1}^i=\min_j\left\{\beta_k^j+\gamma_k^{i,j}\right\}. \tag{11}$$

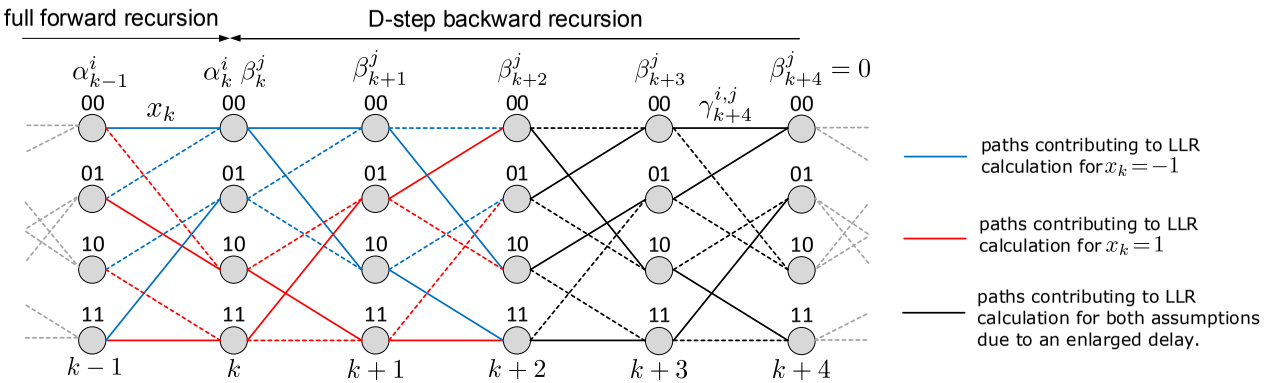


Figure 3: A trellis diagram for binary-phase-shift-keying (BPSK) modulation with memory length $\nu = 2$ and an enlarged delay $D = 4$. In RS-SOVE, the decision-feedback is determined by the survival path on each state. The last $(L - \nu - 1)$ symbols associated to the survival path that leading to current state are preserved and updated along the detection stages. The dashed lines are the discarded paths in the forward and backward recursions due to the Jacobian approximation.

Note that, from detection stage k up to $k+\nu-1$, the state transactions corresponding to different symbol assumptions x_k do not merge with each other at the same state (and on both directions). This is so, since the state transactions from stage k to $k+\nu-1$ follow the below pattern

$$\underbrace{\circ\underbrace{\times\times\cdots\times}_{\nu-1}\longrightarrow\ \times\circ\underbrace{\times\times\cdots\times}_{\nu-2}\longrightarrow\cdots\longrightarrow\ \times\times\cdots\times\underbrace{\circ}_{\nu-1}}_{\nu\text{ stages}},$$

where "$\circ$" denotes the symbol assumption x_k at stage k, and "$\times$" represents all the possible choices for the other $\nu-1$ symbols on each state. There all in total $|\mathcal{X}|$ possible assumptions for x_k, and with each assumption, the sub-trellises formed by the transition pattern above are non-intersecting within stage k and $k+\nu-1$. Hence, by utilizing (9) and (11), the minimal path metric of each symbol assumption x_k in (8) can be recursively computed as

$$\begin{aligned}\min_{x_k}\left(\alpha^i_{k+\nu-1}+\gamma^{i,j}_{k+\nu}+\beta^j_{k+\nu}\right) &= \min_{x_k}\left(\alpha^i_{k+\nu-1}+\beta^i_{k+\nu-1}\right)\\ &= \min_{x_k}\left(\alpha^i_k+\beta^i_k\right).\end{aligned}\tag{12}$$

Then, for each bit assumption $x_{k,n}$ the minimal path metric is the minimum of all $|\mathcal{X}|/2$ symbols $x_k\in\mathcal{X}$ that the nth bit equals to such an assumption. Therefore, the LLR in (8) can be equivalently expressed as

$$L(x_{k,n})=\min_{x_{k,n}=-1}\left(\alpha^j_k+\beta^j_k\right)-\min_{x_{k,n}=1}\left(\alpha^j_k+\beta^j_k\right).\tag{13}$$

In Fig. 3, we illustrate the forward and backward recursions in the RS-SOVE at detection stage k with a binary trellis with $\nu=2$ and $D=4$. As can be seen, the state transactions represented by the red lines and blues lines (both solid and dashed lines) do not merge with each other at stage k and $k+1$, and the recursion (12) holds. The LLR calculation in (13) shows that, with an arbitrary delay D and branch metric computation in (10), the RS-SOVE can be reviewed as an MLM equalizer, but with a full forward recursion and D-step back recursion at each stage.

Next we introduce the proposed optimal FOM shortener design that cooperates with decision-feedback in the RS-SOVE which has been introduced in this section.

3 The Optimal FOM Channel Shortener Design for RS-SOVE

As the HOM shortener is a static and heuristic approach, it neither takes the noise power nor the quality of feedback $\hat{\boldsymbol{x}}$ into account when designing $\boldsymbol{w}_{\mathrm{hom}}$. Consequently, the detection performance is often inferior to the UBM shortener[26]. Moreover, the UBM

shortener also suffers from performance losses in middle and high SNR regimes. The reason is that, as mentioned earlier, the channel tails are truncated and the RS-SOVE does not cooperate with feedback. On the other hand, with high SNR the hard decisions are sufficiently good along the ML path, which can be exploited to cancel the signal part corresponding to the channel tails, instead of direct truncating which ends up with a transmission-energy loss.

Since we are dealing with ISI channels, the FOM receiver filters are designed assuming a large K, in which case we can let $\boldsymbol{H}$ represent the $K\times K$ circular convolution matrix instead of the normal convolution[4]. Such an approximation has no impact on the information rate as $K\to\infty$, see e.g., [31] for a rigorous information-theoretic treatment. From Szegö's eigenvalue distribution theorem[32, 33], the eigenvalues of Toeplitz matrices converge to the Fourier transforms of the sequences that they induce. This implies that, we can equivalently work with the Fourier transforms of all involved Toeplitz matrices, or the vectors that specify them.

Denote the discrete-time Fourier transform (DTFT) of vector $\boldsymbol{h}$ and the inverse operation (IDTFT) as

$$H(\omega) = \sum_{\ell=0}^{L-1} h_\ell \exp(j\omega\ell), \tag{14}$$

$$h_\ell = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\omega)\exp(-j\omega\ell)\mathrm{d}\omega. \tag{15}$$

respectively. Next, we elaborate the optimal FOM shortener design. Although we adopt the same approach as MILB-maximization, the FOM shortener is different from the previous designs[21, 26], which are based on Ungerbeock model and take no feedback into consideration. In [34], the authors extend the UBM shortener to deal with soft feedback and with turbo iterations. However, with RS-SOVE, there are no turbo iterations and the UBM shortener is not applicable.

3.1 The FOM Channel Shortener Design with Feedback

Consider the Forney detection model with feedback,

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) = \exp\left(-\|\boldsymbol{W}\boldsymbol{y}-\boldsymbol{F}\boldsymbol{x}-\boldsymbol{B}\hat{\boldsymbol{x}}\|^2\right), \tag{16}$$

[4]Another conceptually simple way to interpret this is to replace the first $L-1$ symbols in $\boldsymbol{x}$ with its last $L-1$ symbols, i.e., inserting CP. But, here we make such an approximation on $\boldsymbol{H}$ is solely for the sake of designing optimal parameters of the channel shortener. We do not insert CP in the transmit blocks when evaluating the detection performance later.

where $\boldsymbol{W}$, $\boldsymbol{F}$ and $\boldsymbol{B}$ are $K\times K$ convolution matrices generated from $\boldsymbol{w}$, $\boldsymbol{f}$ and $\boldsymbol{b}$, respectively, and $\hat{\boldsymbol{x}}$ is the feedback. There is no constraint on $\boldsymbol{w}$, and $\boldsymbol{f}$, $\boldsymbol{b}$ are as below[5],

$$\boldsymbol{f} = \big(f_0, f_1, \cdots, f_\nu\big), \tag{17}$$

$$\boldsymbol{b} = \big(\underbrace{0, \cdots, 0}_{\nu+1}, b_0, b_1, \cdots, b_{L-\nu-2}\big). \tag{18}$$

The receiver filters ($\boldsymbol{w}$, $\boldsymbol{f}$, $\boldsymbol{b}$) are optimized through maximizing the MILB, which is defined as

$$I_{\mathrm{LB}} = \lim_{K\to\infty} \frac{1}{K}\Big(\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\big[\ln \tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}})\big] - \mathbb{E}_{\boldsymbol{y}}\big[\ln \tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}})\big]\Big), \tag{19}$$

where the expectations are taken over the true channel statistics[6] expression of I_{LB} and with $\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}})$ in (16),

$$\tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}) = \int_{\boldsymbol{x}} \tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

The quality of feedback $\hat{\boldsymbol{x}}$, which impacts the rate I_{LB}, is measured by two parameters,

$$\begin{aligned} \eta &= \frac{1}{K}\mathbb{E}[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\dagger], \\ \sigma &= \frac{1}{K}\mathbb{E}[\hat{\boldsymbol{x}}\boldsymbol{x}^\dagger]. \end{aligned} \tag{20}$$

In RS-SOVE, $\hat{\boldsymbol{x}}$ are hard symbols and we have $\eta = 1$. With soft symbols feedback, η can be calculated from the variance of the estimates, i.e., $\eta = 1 - \mathrm{var}(\hat{\boldsymbol{x}})$. With optimal ($\boldsymbol{w}$, $\boldsymbol{f}$, $\boldsymbol{b}$), and denoting $\tilde{\boldsymbol{y}}$ as the received samples after filtering by $\boldsymbol{w}$, the branch metric $\gamma_k^{i,j}$ in (10) is calculated as

$$\gamma_k^{i,j} = \left| \tilde{y}_k - \sum_{\ell=0}^{\nu} f_\ell x_{k-\ell} - \sum_{\ell=0}^{L-\nu-2} b_\ell \hat{x}_{k-\ell-\nu-1} \right|^2. \tag{21}$$

Before optimizing ($\boldsymbol{w}$, $\boldsymbol{f}$, $\boldsymbol{b}$), we introduce the following notations. Following (14), we denote the DTFT of $\boldsymbol{w}$, $\boldsymbol{f}$, and $\boldsymbol{b}$ as $W(\omega)$, $F(\omega)$, and $B(\omega)$, respectively. Then, we let

$$M(\omega) = -\frac{N_0}{N_0 + |H(\omega)|^2}, \tag{22}$$

$$\tilde{M}(\omega) = \sigma^2\,(1 + M(\omega)) - \sigma, \tag{23}$$

[5]Although with arbitrary $\boldsymbol{w}$, the feedback filter $\boldsymbol{b}$ can be arbitrary long, we make such constraints to align the complexity of decision-feedback detection in the RS-SOVE corresponding to the HOM shortener.

[6]In order to obtain a tractable problem[21], we make the assumption that $\boldsymbol{x}$ comprises IID complex Gaussian variables when calculating I_{LB}.

and

$$\boldsymbol{\phi}(\omega) = \left[\exp\left(j\omega(\nu+1)\right)\ \exp\left(j\omega(\nu+2)\right)\ \ldots\ \exp\left(j\omega(L-1)\right)\right]^{\mathrm{T}}. \tag{24}$$

Further, denote $(L-\nu-1)\times 1$ vector $\boldsymbol{\varepsilon}_1$, and $(L-\nu-1)\times(L-\nu-1)$ Hermitian matrix $\boldsymbol{\varepsilon}_2$ as

$$\boldsymbol{\varepsilon}_1 = \frac{\sigma}{2\pi}\int_{-\pi}^{\pi} M(\omega)F^*(\omega)\boldsymbol{\phi}(\omega)\mathrm{d}\omega, \tag{25}$$

$$\boldsymbol{\varepsilon}_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{M}(\omega)|F(\omega)|^2\boldsymbol{\phi}(\omega)\boldsymbol{\phi}(\omega)^{\dagger}}{1+|F(\omega)|^2}\mathrm{d}\omega. \tag{26}$$

With definitions in (22)-(26), we have the below lemma that states the closed-form MILB.

Lemma 1. *The MILB in (19) equals*

$$\begin{aligned} I_{\mathrm{LB}} \;=\; & \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log\left(1+|F(\omega)|^2\right)-|F(\omega)|^2-\frac{L(\omega)}{1+|F(\omega)|^2}\right)\mathrm{d}\omega \\ & +\frac{1}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\left\{F^*(\omega)\left(W(\omega)H(\omega)-\sigma B(\omega)\right)\right\}\mathrm{d}\omega, \end{aligned} \tag{27}$$

where

$$\begin{aligned} L(\omega) \;=\; & |F(\omega)W(\omega)|^2\left(N_0+|H(\omega)|^2\right)+\sigma|F(\omega)B(\omega)|^2 \\ & -2\sigma|F(\omega)|^2\mathcal{R}\{H(\omega)W(\omega)B^*(\omega)\}. \end{aligned}$$

Proof. In [20, eq.(5)-(6)], the generalized mutual information I_{GMI} is derived for any $K\times K$ linear multi-input and multi-output (MIMO) channel. For ISI channels, which can be viewed as special cases of MIMO channel, it holds that

$$I_{\mathrm{LB}} = \lim_{K\to\infty}\frac{1}{K}I_{\mathrm{GMI}}.$$

By applying Szegö's theorem [33] and after some manipulations, (27) follows. □

With I_{LB} stated in (27), the optimal $W(\omega)$ and $B(\omega)$ that maximize I_{LB} are in Theorem 1.

Theorem 1. *The optimal $W(\omega)$ that maximizes I_{LB} equals,*

$$W_{\mathrm{opt}}(\omega) = \frac{H^*(\omega)\left(1+|F(\omega)|^2+\sigma F(\omega)B^*_{\mathrm{opt}}(\omega)\right)}{F^*(\omega)(N_0+|H(\omega)|^2)}, \tag{28}$$

and when $\sigma > 0$, the optimal $B(\omega)$ reads,

$$B_{\text{opt}}(\omega) = -\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\phi}(\omega). \tag{29}$$

With $W_{\text{opt}}(\omega)$ and $B_{\text{opt}}(\omega)$, I_{LB} equals,

$$I_{\text{LB}} = \begin{cases} \mathcal{J}\big(F(\omega)\big), & \sigma = 0, \\ \mathcal{J}\big(F(\omega)\big) - \boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1, & 0 < \sigma \leq 1, \end{cases} \tag{30}$$

where $\mathcal{J}\big(F(\omega)\big)$ reads

$$\mathcal{J}\big(F(\omega)\big) = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big) + M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega. \tag{31}$$

Proof. See Appendix A. □

In (30), the term $-\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1$ is the information rate increment due to the feedback $\hat{\boldsymbol{x}}$. From Theorem 1, $W(\omega)$, $B(\omega)$ are in closed forms, and $\boldsymbol{w}$, $\boldsymbol{b}$ can be obtained through IDTFT operations. But for $F(\omega)$ and $\boldsymbol{f}$, a closed form solution can not be reached. Hence, we use a gradient-ascending based optimization, with the updating at each iteration defined as

$$\boldsymbol{f}^i = \boldsymbol{f}^{i-1} + \nabla_{\boldsymbol{f}^*} I_{\text{LB}}. \tag{32}$$

As the DTFT of $\boldsymbol{f}$ reads

$$F(\omega) = \sum_{k=0}^{\nu} f_k \exp\big(jk\omega\big),$$

the first-order derivatives of $\mathcal{J}(F(\omega))$ and $\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1$ in (30) with respect to f_k read

$$\frac{\partial \mathcal{J}}{\partial f_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(M(\omega) + \frac{1}{1+|F(\omega)|^2}\right)F^*(\omega)\exp\big(jk\omega\big)\mathrm{d}\omega,$$

$$\frac{\partial \boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1}{\partial f_k} = -\frac{\partial \boldsymbol{\varepsilon}_1^{\dagger}}{\partial f_k}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\frac{\partial \boldsymbol{\varepsilon}_2}{\partial f_k}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1,$$

respectively, and

$$\frac{\partial \boldsymbol{\varepsilon}_1^{\dagger}}{\partial f_k} = \frac{\sigma}{2\pi}\int_{-\pi}^{\pi} M(\omega)\boldsymbol{\phi}(\omega)^{\dagger}\exp(jk\omega)\mathrm{d}\omega,$$

$$\frac{\partial \boldsymbol{\varepsilon}_2}{\partial f_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2\boldsymbol{\phi}(\omega)\boldsymbol{\phi}(\omega)^{\dagger}}{\big(1+|F(\omega)|^2\big)^2}F^*(\omega)\exp\big(jk\omega\big)\mathrm{d}\omega.$$

Although due to the non-concaveness of I_{LB} in (30), the optimization may converge to a local maximum, such an optimization over $\boldsymbol{f}$ is still meaningful, in the sense that the MILB is increased even with a local maximum attained. We initialize $\boldsymbol{f}$ in (32) with $\boldsymbol{h}_{\mathrm{f}}$ obtained from the HOM shortener. When N_0 decreases and with $\sigma = 1$, such an initialization is asymptotically close to the maximum point as the HOM shortener performs close the the FOM shortener, due to the perfect feedback.

3.2 The UBM Channel Shortener Design without Feedback

Next, we introduce the UBM shortener design. By identifying $\boldsymbol{V} = \boldsymbol{F}^{\dagger}\boldsymbol{W}$, $\boldsymbol{R} = \boldsymbol{F}^{\dagger}\boldsymbol{B}$ and $\boldsymbol{G} = \boldsymbol{F}^{\dagger}\boldsymbol{F}$, the model (16) can be rewritten as

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\Big(2\mathcal{R}\{\boldsymbol{x}^{\dagger}(\boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}})\} - \boldsymbol{x}^{\dagger}\boldsymbol{G}\boldsymbol{x} + \vartheta\Big), \tag{33}$$

where $\vartheta = -\|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{B}\hat{\boldsymbol{x}}\|^2$. In the design of the UBM shortener, $\boldsymbol{G}$ is an arbitrary Hermitian matrix and can be non-positive definite[21]. In the RS-SOVE, the term ϑ is calculated with the survival path on each state. In order to calculate ϑ, we need to decompose $\boldsymbol{G} = \boldsymbol{F}^{\dagger}\boldsymbol{F}$, which requires $\boldsymbol{G}$ to be positive definite. In such a case, (33) is identical to (16), that is, the UBM shortener becomes the FOM shortener. This dilemma makes the Ungerboeck model not suitable for decision-feedback detection. But with turbo iterations, as $\hat{\boldsymbol{x}}$ is known before the RS-SOVE, it is the same for all states and ϑ can be removed from (33). However, as we are designing a channel shortener with no turbo iterations, we assume no feedback and (33) changes to

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\Big(2\mathcal{R}\{\boldsymbol{x}^{\dagger}\boldsymbol{V}\boldsymbol{y}\} - \boldsymbol{x}^{\dagger}\boldsymbol{G}\boldsymbol{x}\Big). \tag{34}$$

The $K \times K$ convolution matrix $\boldsymbol{V}$ generated from vector $\boldsymbol{v}$ has the same structure as $\boldsymbol{W}$, while the $K \times K$ Toeplitz matrix $\boldsymbol{G}$ is Hermitian and band-shaped, with only the middle $2\nu+1$ diagonals can take non-zero values. Denote the vector comprising the first $(\nu+1)$ elements in the first column of $\boldsymbol{G}$ as

$$\boldsymbol{g} = \big(g_0, g_1, \cdots, g_\nu\big).$$

With optimal $(\boldsymbol{v}, \boldsymbol{g})$, and denoting $\tilde{\boldsymbol{y}}$ as the received samples after filtering by $\boldsymbol{v}$, the branch metric $\gamma_k^{i,j}$ is calculated as

$$\gamma_k^{i,j} = g_0|x_k|^2 - 2\mathcal{R}\Big\{x_k^*\Big(\tilde{y}_k - \sum_{\ell=1}^{\nu} g_\ell x_{k-\ell}\Big)\Big\}. \tag{35}$$

The model (34) has been considered in earlier literatures such as [1, 21]. The optimal solutions of $(\boldsymbol{v}, \boldsymbol{g})$ can be found in [21], which can also be deduced from Theorem 1 directly.

By setting $\sigma=0$, the optimal $V(\omega)$ for (34) reads

$$V_{\text{opt}}(\omega) = \frac{H^*(\omega)}{N_0 + |H(\omega)|^2}\big(1 + G(\omega)\big), \tag{36}$$

and the optimal $G(\omega)$ is the unique solution that maximizes I_{LB} in (19), which is evaluated based on (34) and equals

$$I_{\text{LB}} = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1 + G(\omega)\big) + M(\omega)\big(1 + G(\omega)\big)\Big)\mathrm{d}\omega. \tag{37}$$

Comparing (37) to (31), the only difference is that $|F(\omega)|^2$ in (31) is replaced by $G(\omega)$. Therefore, the UBM shortener is more general than the FOM shortener under the case that $\sigma = 0$. We point out the fact that, both the FOM and UBM shorteners are invariant under the minimum-phase transforming of the original channel $\boldsymbol{h}$. This is because, the homomorphic filter $\boldsymbol{w}_{\text{hom}}$ is an all-pass filter, which has no impact on the noise statistical properties, and then the convolution matrix generated from the all-pass filter will be absorbed by the prefilters $\boldsymbol{W}$ and $\boldsymbol{V}$, respectively. Hence, with the FOM and UBM shorteners, it is no need to transform $\boldsymbol{h}$ into an minimum-phase equivalent form prior to prefiltering.

3.3 Design the Optimal σ for the FOM Channel Shortener

In Theorem 1, the optimal $W(\omega)$ and $F(\omega)$ are related to the feedback quality parameter σ. However, according to the expectation in (20), σ is hard to find at the design stage. Moreover, it is not necessarily optimal to use the σ calculated with (20). Therefore, it is a free optimization parameter. In the next, we analyze the optimal design of σ.

With higher-order modulations, we assume that when a symbol error occurs on the ML path, the hard decision $\hat{x}_k$ and the transmit symbol x_k are independent. Then,

$$\begin{aligned}\sigma &= (1 - P_{\text{e}}) \cdot \mathbb{E}[|x_k|^2] + P_{\text{e}} \cdot \mathbb{E}[\hat{x}_k x_k^*],\\ &\approx 1 - P_{\text{e}},\end{aligned} \tag{38}$$

where P_{e} is the symbol error rate (SER) of the RS-SOVE. As LMMSE detection is a special case of MILB detection with $\nu=0$, when $\nu>0$ the FOM shortener with the RS-SOVE is superior to the LMMSE detector and renders a lower SER [20]. That is, denoting $\hat{\boldsymbol{x}}^{\text{LMMSE}}$ as the LMMSE detector and $P_{\text{e}}^{\text{LMMSE}}$ as the corresponding SER, respectively, it holds that

$$\begin{aligned}P_{\text{e}} &\le P_{\text{e}}^{\text{LMMSE}}\\ &\overset{(a)}{\le} \mathbb{E}\left[\left(\boldsymbol{x} - \hat{\boldsymbol{x}}^{\text{LMMSE}}\right)\left(\boldsymbol{x} - \hat{\boldsymbol{x}}^{\text{LMMSE}}\right)^{\dagger}\right]\Big/2\\ &\overset{(b)}{=} \delta_{\text{mse}}/2.\end{aligned} \tag{39}$$

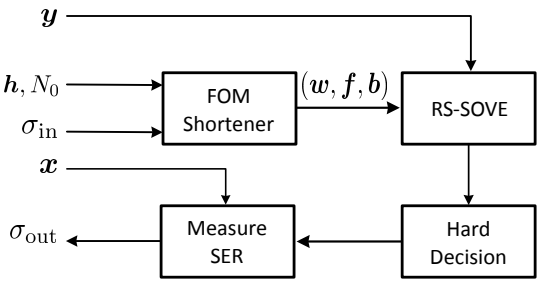


Figure 4: The diagram of evaluating the optimal σ for the FOM shortener. For each input $\sigma_{\rm in}$, the optimal ($\boldsymbol{w}$, $\boldsymbol{f}$, $\boldsymbol{b}$) are calculated with Theorem 1, and the detection uses the FOM shortener followed by the RS-SOVE. The output $\sigma_{\rm out} = 1 - P_{\rm e}$ is measured based on the hard decisions output from the RS-SOVE.

where

$$\delta_{\rm mse} = -\frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\mathrm{d}\omega,$$

which is the MSE of the LMMSE estimate. The inequality (a) is proved in Appendix B, and the equality (b) is from Szegö's eigenvalue distribution theorem. Hence, from (38) and (39),

$$\sigma \geq 1 - \delta_{\rm mse}/2. \tag{40}$$

The inequality (40) provides some insight about designing σ for the FOM shortener. As we are expecting that, the RS-SOVE with decision-feedback shall outperform itself without feedback, i.e., $\sigma = 0$, the input σ to design $(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{b})$ should be set to, at least larger than $1 - \delta_{\rm mse}/2$, and we can let $\sigma = 1$ when $\delta_{\rm mse}$ is sufficient small.

From Theorem 1, an input σ determines the optimal channel shortening parameters ($\boldsymbol{w}$, $\boldsymbol{f}$, $\boldsymbol{b}$), which in turn affects the quality of the decision-feedback $\hat{\boldsymbol{x}}$ in the RS-SOVE. Therefore, there is a mismatch between the input σ, and the practical output σ measured by the outputs of the RS-SOVE generated by such a designed σ. We therefore use Monte Carlo simulation under the EPR-4 and Proakis-C ISI channels to exploit the relationships between the input and output σ. The test set-up is depicted in Fig. 4 and the results are depicted in Fig. 5. The output $\sigma_{\rm out}$ is measured with (20), under each input $\sigma_{\rm in}$ which is utilized to generate the optimal parameters of the FOM shortener.

Under both channels, $\sigma_{\rm in}$ is increased from 0 to 1, and we have two interesting observations. The first observation is that, with FOM shortener, the RS-SOVE can only benefit from the hard decisions when the quality of feedback is above a certain threshold, otherwise, setting $\sigma_{\rm in} = 0$, i.e., utilizing no feedback in the RS-SOVE (such as the UBM shortener) is close to optimal. The second observation is that, when the RS-SOVE can benefit from the feedback, setting $\sigma_{\rm in} = 1$ is close to optimal, which is aligned with the analysis leading to (40). The reason behind this phenomenon is that, when SNR is low, implying that the quality of $\hat{\boldsymbol{x}}$ in the RS-SOVE is fairly poor, it is better to truncate the channel tails to

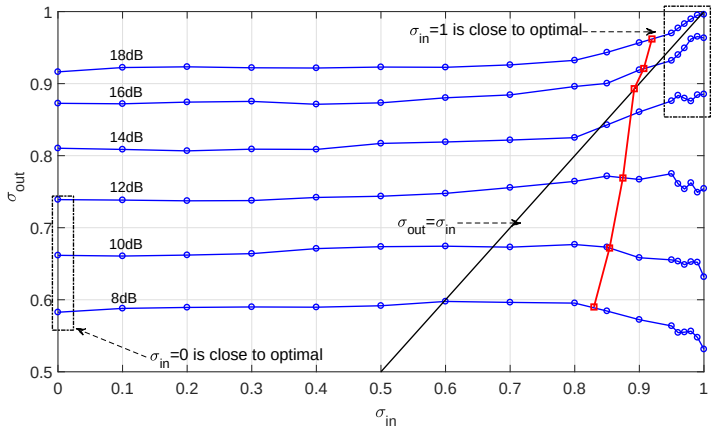


Figure 5: The optimal σ investigation under EPR-4 [35] channel with $\boldsymbol{h}=0.5+0.5^{-1}-0.5z^{-2}-0.5z^{-3}$ and 8PSK modulation. The value $\sigma_{\rm out}$ is measured according to eq. (38). The red line presents the results of $\sigma_{\rm out}$ by setting $\sigma_{\rm in}=1-\delta_{\rm mse}/2$ as in eq. (40). The simulation results under Proakis-C[36] channel with $\boldsymbol{h}=0.227+0.46z^{-1}+0.668z^{-2}+0.46z^{-3}+0.227z^{-4}$ are similar as those for EPR-4.

prevent error-propagation. However, when SNR is above a certain threshold, the feedback quality improves and the FOM shortener will benefit from $\hat{\boldsymbol{x}}$, in which case we can let $\sigma_{\rm in}=1$. Nevertheless, setting $\sigma_{\rm in}=1-\delta_{\rm mse}/2$ according to (40) also provides a good estimation of the optimal input σ, especially when there is no priori information about the required SNR in relation to the successful decoding.

As the UBM shortener is a more general and has better performance than the FOM shortener when $\sigma=0$[34], when designing the optimal channel shortener, it is sufficient to consider either the UBM shortener (34), or the FOM shortener (16) with $\sigma_{\rm in}=1$. The remaining issue is the criterion for choosing between these two shorteners. Such a criterion is difficult to find theoretically, but as we show later through numerical results, it can be designed based on the code-rate of the considered SC systems. This is so, since the decoder at medium and high code-rates need better quality inputs generate from the detector for successive decoding, therefore, the $P_{\rm e}$ is small and the quality of the feedbacks in the RS-SOVE is relatively good, in which case the proposed FOM channel shortener is superior to the UBM shortener.

With the HOM, FOM and UBM channel shorteners introduced in Sec. II-A, Sec. III-A, and Sec. III-B, respectively, next we analyze the mutual information (MI) characteristics. We show that the FOM shortener is superior to the HOM shortener in general, and better than the UBM shortener when the feedback $\hat{\boldsymbol{x}}$ are fairly good.

4 Theoretical Information Rates of the Channel Shorteners

For simplicity, we denote the optimal I_{LB} of the FOM and UBM shorteners as I_{FOM} and I_{UBM}, calculated in (30) and (37), respectively. Further, we denote I_{FOM} computed with $\sigma = 0$ and 1 as I^0_{FOM} and I^1_{FOM}. Similarly, we let I_{HOM} denote the information rate reached by the HOM shortener. Firstly, we state the below property.

Property 1. *Denote $H_{\mathrm{f}}(\omega)$ and $H_{\mathrm{b}}(\omega)$ as the DTFTs of $\boldsymbol{h}_{\mathrm{f}}$ and $\boldsymbol{h}_{\mathrm{b}}$ in (6) and (7), respectively. Then, it holds that*

$$I^{\mathrm{L}}_{\mathrm{HOM}} \leq I_{\mathrm{HOM}} \leq I^{\mathrm{U}}_{\mathrm{HOM}}, \tag{41}$$

where

$$I^{\mathrm{L}}_{\mathrm{HOM}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(1 + \frac{|H_{\mathrm{f}}(\omega)|^2}{1+|H_{\mathrm{b}}(\omega)|^2}\right) \mathrm{d}\omega, \tag{42}$$

$$I^{\mathrm{U}}_{\mathrm{HOM}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(1 + |H_{\mathrm{f}}(\omega)|^2\right) \mathrm{d}\omega. \tag{43}$$

Proof. As $\boldsymbol{h} = \boldsymbol{h}_{\mathrm{f}} + \boldsymbol{h}_{\mathrm{b}}$, the received signal model (5) can be rewritten as

$$\tilde{\boldsymbol{y}} \ = \boldsymbol{h}_{\mathrm{f}} \star \boldsymbol{x} + \boldsymbol{h}_{\mathrm{b}} \star \boldsymbol{x} + \tilde{\boldsymbol{n}}.$$

The lower bound of I_{HOM} is achieved when feedback $\hat{\boldsymbol{x}}$ acts as noise, while the upper bound is achieved when $\hat{\boldsymbol{x}}$ is perfect. Therefore, the inequality (41) holds. □

More discussions about the properties of truncated channel response $H_{\mathrm{f}}(\omega)$ can be found in, e.g., [37]. Here we mention the fact that, the upper bound $I^{\mathrm{U}}_{\mathrm{HOM}}$ can be higher than Shannon capacity $\mathcal{C}$, due to the perfect feedback. Secondly, we state Property 2.

Property 2. *The below inequalities hold,*

$$I_{\mathrm{HOM}} \leq I^0_{\mathrm{FOM}} \leq I_{\mathrm{UBM}} \leq \mathcal{C}. \tag{44}$$

Proof. See Appendix C. □

From Property 2, when there is no feedback, the FOM shortener is lower-bounded by the HOM shortener and upper-bounded by the UBM shortener. Further, all of them are bounded by $\mathcal{C}$. However, when $\sigma > 0$, the rate of the FOM shortener can be higher than, both the UBM shortener and $\mathcal{C}$, due to the presence of feedback.

Lastly, we build the relationship between $I^{\mathrm{U}}_{\mathrm{HOM}}$ and I^1_{FOM}, which is stated in Property 3. Although with $\sigma = 1$, the feedback $\hat{\boldsymbol{x}}$ is perfect, the symbol detection is still utilizing the

received samples $\tilde{\boldsymbol{y}}$, which are not perfect. Therefore, such a comparison is meaningful and shows that, when there are no errors in $\hat{\boldsymbol{x}}$, the FOM shortener is superior to the HOM shortener.

Property 3. *The below inequality holds,*

$$I^{\mathrm{U}}_{\mathrm{HOM}} \leq I^{1}_{\mathrm{FOM}}. \tag{45}$$

Proof. By setting $(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{b}) = (\boldsymbol{w}_{\mathrm{hom}}, \boldsymbol{h}_{\mathrm{f}}, \boldsymbol{h}_{\mathrm{b}})$, the FOM shortener is identical to the HOM shortener. And with $\sigma = 1$, I_{LB} in this case equals $I^{\mathrm{U}}_{\mathrm{HOM}}$. As I^{1}_{FOM} maximizes I_{LB}, (45) holds. □

We summarize the above discussions in the below theorem.

Theorem 2. *The below equalities of theoretical information rates hold with $\sigma = 0$,*

$$I^{\mathrm{L}}_{\mathrm{HOM}} \leq I_{\mathrm{HOM}} \leq I^{0}_{\mathrm{FOM}} \leq I_{\mathrm{UBM}} \leq \mathcal{C}, \tag{46}$$

while with $\sigma = 1$, the below inequalities hold,

$$I_{\mathrm{HOM}} \leq I^{\mathrm{U}}_{\mathrm{HOM}} \leq I^{1}_{\mathrm{FOM}}. \tag{47}$$

Proof. Combing Properties 1-3 yields Theorem 2. □

Theorem 2 shows that, when the quality of feedback is poor, i.e., $\sigma = 0$, the UBM shortener has best performance compared to both the FOM and HOM shorteners, while when the feedback is perfect, the FOM shortener outperforms both the HOM and UBM shorteners. Moreover, as we showed earlier, the optimal σ for the FOM shortener is either 0 or 1, hence, one can design a system that switches between the UBM shortener, and the FOM shortener designed for $\sigma = 1$, to achieve the best performance under all cases.

Note that with $\boldsymbol{f} = \boldsymbol{h}_{\mathrm{f}}$ and the optimal $\boldsymbol{w}_{\mathrm{opt}}, \boldsymbol{b}_{\mathrm{opt}}$ calculated in (28) and (29), I_{LB} equals

$$I_{\mathrm{LB}} = 1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} \Big(\log\big(1 + |H_{\mathrm{f}}(\omega)|^2\big) + M(\omega)\big(1 + |H_f(\omega)|^2\big) \Big) \mathrm{d}\omega - \boldsymbol{\varepsilon}_1^{\dagger} \boldsymbol{\varepsilon}_2^{-1} \boldsymbol{\varepsilon}_1. \tag{48}$$

By definition, I_{LB} in (48) is no less than I_{LB} computed with $(\boldsymbol{w}_{\mathrm{hom}}, \boldsymbol{h}_{\mathrm{f}}, \boldsymbol{h}_{\mathrm{b}})$, which equals $I^{\mathrm{U}}_{\mathrm{HOM}}$. From (43) and (48), we have an interesting corollary below that shows the relation between $\boldsymbol{h}$ and $\boldsymbol{h}_{\mathrm{f}}$ for any ISI channels, and reveals the fact that, with the same target response $\boldsymbol{f}$ but optimized $\boldsymbol{w}$, $\boldsymbol{b}$, the FOM shortener outperforms the HOM shortener.

Corollary 1. *For any ISI channel $\boldsymbol{h}$ and the target response $\boldsymbol{h}_{\mathrm{f}}$ defined in (6), the inequality*

$$\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\left(1+|H_{\mathrm{f}}(\omega)|^2\right)\mathrm{d}\omega \leq 1,$$

holds, where $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_1$ are defined in (25) and (26) with $F(\omega)=H_{\mathrm{f}}(\omega)$, and $M(\omega)$, $\tilde{M}(\omega)$ are defined in (22) and (23) with $\sigma=1$, respectively.

5 Detection Complexity

In this section, we analyze the detection complexity of the HOM, UBM, and FOM detectors. The detection complexity contains three parts: computing the channel shortening filters that are listed in Table 1, prefiltering and feedback filtering operations, and the trellis-search in RS-SOVE.

We first discuss the complexity involved in the computation of the channel shortening filters. It was shown in [26, 38] that, the number of complex multiplications (CM) to compute the channel shortening filters is $\mathcal{O}\left(N\log_2 N\right)$ for both the HOM and MILB detectors, where we assume that one CM is equivalent to 4 real multiplications and N is the fast Fourier transforms (FFT) size for frequency domain computations (with a typical value 128). Moreover, although sharing the same complexity order, the MILB detectors are shown to require only around half the complexity of the HOM detector in [26]. With FOM detector, the preprocessing complexity is slightly higher as a numerical optimization of $\boldsymbol{f}$ is needed. However, the complexity increment is neglectable compared to the complexity of trellis-search in RS-SOVE due to the fast convergence of the optimization. Nevertheless, the majority part of the detection complexity lies in the the filtering operations and trellis-search as they are per-symbol based operations, especially with a large block length K and high-order modulation size $|\mathcal{X}|$.

Next we analyze the complexity involved in prefiltering, feedback filtering, and the trellis-search, which are summarized in Table 2 in terms of the number of CM per symbol. Without loss of generality, the filtering operations are in time domain and the lengths of prefilters are set to the size of FFT. In RS-SOVE, the number of states is $|\mathcal{X}|^{\nu}$ and the branching factor equals $|\mathcal{X}|$ for all three detectors. Computing the branch metric $\gamma_k^{i,j}$ for each state transition requires $\left(\nu+\frac{3}{2}\right)$ CM operations (without the feedback filtering). As can be seen from Table 2, the FOM and HOM have the same complexities, while the UBM has less complexity due to the fact that there is no feedback filtering operations.

Table 2: Complexity-Performance Comparison between the Detectors per Symbol-detection.

Detector	Prefiltering	Feedback filtering	Number of states	Branching factor	Compute path metric $\gamma_k^{i,j}$	The total number of CM operations per symbol
FOM	N	$L-\nu-1$	$\lvert\mathcal{X}\rvert^{\nu}$	$\lvert\mathcal{X}\rvert$	$\nu+\frac{3}{2}$	$N+\lvert\left(L+\frac{1}{2}\right)\mathcal{X}\rvert^{\nu+1}$
UBM	N	n.a.	$\lvert\mathcal{X}\rvert^{\nu}$	$\lvert\mathcal{X}\rvert$	$\nu+\frac{3}{2}$	$N+\lvert\left(\nu+\frac{3}{2}\right)\mathcal{X}\rvert^{\nu+1}$
HOM	N	$L-\nu-1$	$\lvert\mathcal{X}\rvert^{\nu}$	$\lvert\mathcal{X}\rvert$	$\nu+\frac{3}{2}$	$N+\lvert\left(L+\frac{1}{2}\right)\mathcal{X}\rvert^{\nu+1}$

6 Empirical Results

In this section, we provide empirical results to show the information rates and detection performance of the proposed FOM channel shortener with RS-SOVE, which is compared to the UBM and HOM shorteners. Throughout all tests, without explicitly pointing out we assume that the memory length $\nu=1$ after channel shortening to achieve a low-complexity receiver design.

For each transmit symbol vector $\boldsymbol{x}$, with different channel shorteners, the bit LLRs $L(x_{k,n})$ are calculated in (8) based on different branch metric computations as in (10), (21), and (35), respectively. As the transmit bits $x_{k,n}$ are independent, the logarithm of the conditional probability of each symbol $x'_k \in \mathcal{X}$ for a given transmit symbol x_k, i.e., $p(x'_k|x_k)$, can be computed as

$$\begin{aligned}\log p(x'_k|x_k) &= \sum_{n=0}^{\log_2|\mathcal{X}|-1} \log p(x'_{k,n}|x_{k,n}) \\ &= \sum_{n=0}^{\log_2|\mathcal{X}|-1} \Big(\frac{(1+x'_{k,n})L(x_{k,n})}{2} - \log\big(1+\exp\left(L(x_{k,n})\right)\big)\Big).\end{aligned}$$

Then, the measured MI is calculated as

$$I(\boldsymbol{y};\boldsymbol{x}) = \log|\mathcal{X}| - \mathbb{E}_{x_k,x'_k\in\mathcal{X}}\big[\log p(x'_k|x_k)\big].$$

6.1 The Impact of Decision-Delay D in RS-SOVE

First, we evaluate the normalized MI measured for the EPR-4 channel, and investigate the impact of decision-delay D for different modulation schemes in RS-SOVE. The HOM shortener is tested with D set to $L-1$, $L+2$ and $L+20$, respectively. As can be seen in Fig. 6, with 16QAM modulation, enlarging D from $L-1$ to $L+2$ has around an SNR gain of 0.4 dB in terms of the normalized MI. However, further increasing D up to $L+20$ only has marginal SNR gain. Since a larger delay increases process latency in the RS-SOVE, in the remaining tests we set $D=L+2$ in the RS-SOVE for all channel shorteners.

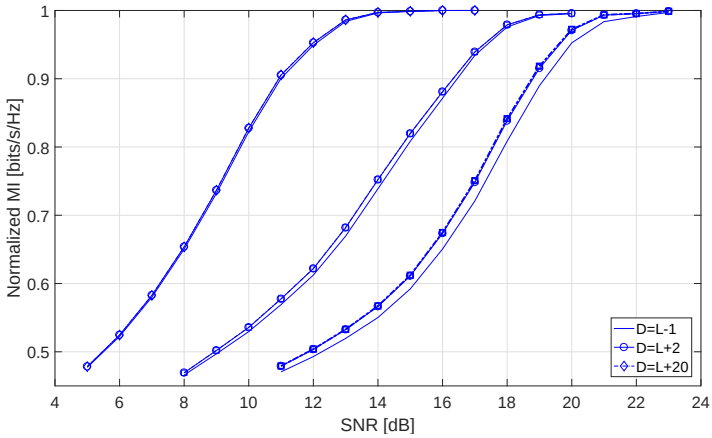


Figure 6: Performance evaluation of the HOM shortener with different delays and modulation schemes. From left to right, the modulation schemes are quadrature-phase-shift keying (QPSK), 8PSK and 16QAM. The normalized MI is measured with the output from the RS-SOVE.

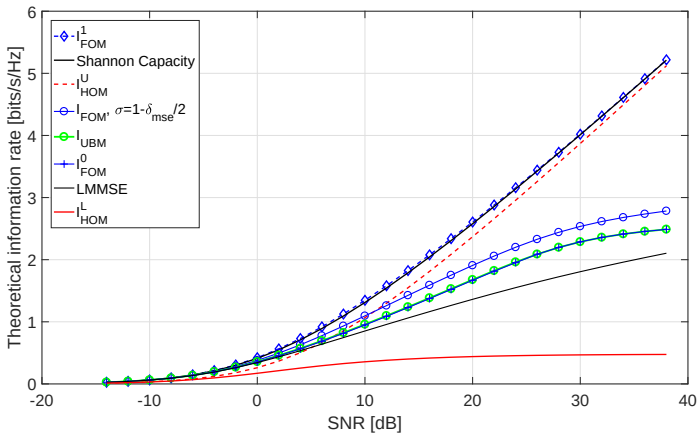


Figure 7: Theoretical information rates under EPR-4 channel (the legend is ordered from the top curve to the bottom curve).

6.2 Theoretical Information Rates

Next, we simulated the theoretical information rates that have been discussed in Sec. IV under EPR-4 channel (the results for the Proakis-C channel is similar). In comparison, we also add the rates of I_{FOM} with $\sigma = 1 - \delta_{\mathrm{mse}}/2$. The rates of LMMSE detection and Shannon capacity $\mathcal{C}$ are also presented. As can be seen in Fig. 7, with $\sigma = 0$ the information rates of the FOM shortener (I^0_{FOM}) and the UBM shortener (I_{UBM}) are quite close. In the low SNR regime, the UBM shortener is superior, while in the high SNR regime, the information rates of the FOM shortener with $\sigma = 1$ (I^1_{FOM}) are the best. As SNR increases, $I^{\mathrm{U}}_{\mathrm{HOM}}$ asymptotically approaches the rate I^1_{FOM}, and both curves almost overlap with the capacity $\mathcal{C}$.

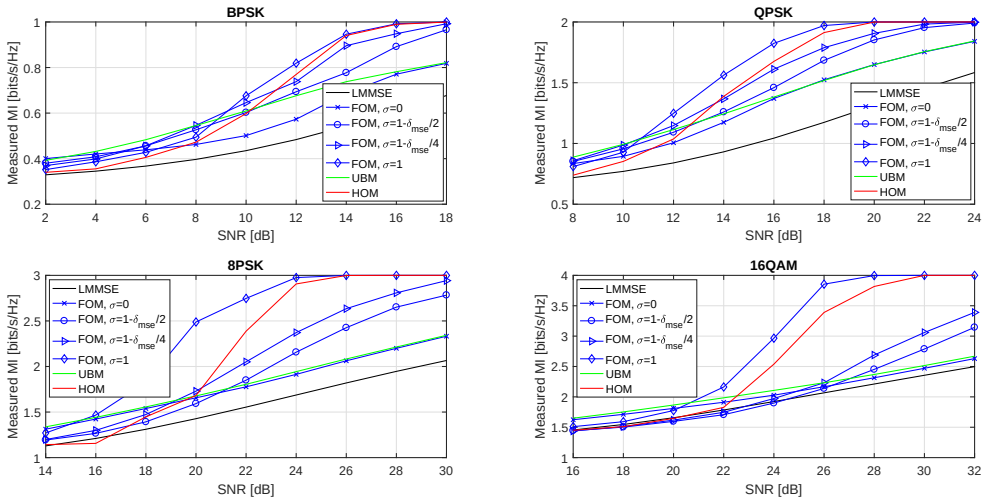


Figure 8: Measured MI under Proakis-C channel and with different modulation schemes. The UBM shortener provides the best performances when the normalized MI is lower than around 1/2, while the FOM shortener with $\sigma = 1$ is the best when the normalized MI is above 1/2. The conventional HOM shortener is in general interior, except in the high SNR regime where approaches the rates of the FOM.

6.3 Measured MI

To verify the practical performance, we measure the MI achieved by the three shorteners with different modulation schemes under Proakis-C channel. As can be seen in Fig. 8, the measured MI results are well aligned with the theoretical analysis illustrated in Fig. 7. The UBM shortener outperforms both the FOM and HOM shorteners in the low SNR regime. But as SNR increases, the FOM shortener becomes superior to the other two shorteners. The HOM shortener is in general inferior to the FOM shortener, and in the high SNR regime the HOM shortener performs close to the FOM shortener. We also add the information rates of the FOM shortener with both $\sigma = 1-\delta_{\text{mse}}/2$ and $\sigma = 1-\delta_{\text{mse}}/4$, which are inferior to the rates of the FOM shortener with $\sigma = 1$ in the high SNR regime.

Most interestingly, the cross points of the FOM shortener with $\sigma = 1$ and the UBM shortener are around 1/2 in terms of the normalized MI, which indicates that, the switching criterion of the FOM and UBM shorteners can be based on the output MI of the RS-SOVE, or equivalently, the input MI to the outer-decoder. As for error-correcting codes, the input MI of the LLRs sent to the decoders shall be no less than the code-rate for successfully decoding. Hence, we can use the code-rate as a criterion: when the code-rate is above 1/2, the FOM shortener will provide better performance, otherwise we switch to the UBM shortener. This is also due to the fact that the FOM shortener is superior to the UBM only when the feedback quality is fairly good.

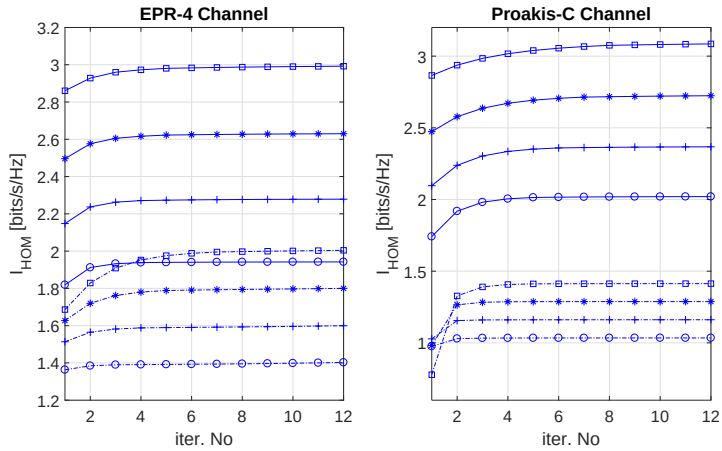


Figure 9: The convergence speed of the FOM shortener. The dashed lines are with $\sigma = 1/2$ while the solid lines are with $\sigma = 1$. In both cases and from bottom to up, the SNR equals 10dB, 12dB, 14dB, and 16dB, respectively. With larger σ, the optimization need more steps to converge. However, as can be seen in both figures, the optimization process converges in 4-8 iterations.

6.4 Parameter Optimization of the FOM Channel Shortener

Next, we evaluate the parameter optimization of the FOM channel shortener. As stated in Theorem 1, the optimal prefilters $\boldsymbol{w}$ and $\boldsymbol{b}$ are in closed forms, while the optimal $\boldsymbol{f}$ has to be found through an optimization process. In Fig. 9, we plot the convergence speed under EPR-4 and Proakis-C channels at different SNR points. We test with $\sigma = 1/2$ and $\sigma = 1$, respectively. As can be seen, the optimization converges very fast in a few number of iterations.

6.5 Performance Evaluation with Turbo Codes

At last, we evaluate the coded BER performance with turbo codes from LTE standard [39] and under different code-rates and modulation schemes. The number of information bits is set to $K = 1064$ for all tests. We make no attempts to optimize σ and directly set $\sigma = 1$ for testing with EPR-4 and Proakis-C channels.

In Fig. 10, we evaluate the coded BER under EPR-4 channel with 8PSK modulation. As expected, the UBM shortener performs the best at code-rates 1/3 and 1/2. At higher code-rates 2/3 and 3/4, the UBM shortener becomes inferior to the proposed FOM shortener. In all cases, the FOM shortener is superior to the HOM shortener.

In Fig. 11, we evaluate the coded BER under Proakis-C channel with 16QAM modulation. In this case, the UBM shortener outperforms the other two channel shorteners at code-rate 1/3 only. At higher code-rates, the UBM shortener perform poorly. However, the FOM shortener is still around 1-2 dB better than the HOM shortener at all code-rates in terms of SNR. The results are also aligned with Fig. 8 where we show that the UBM shortener

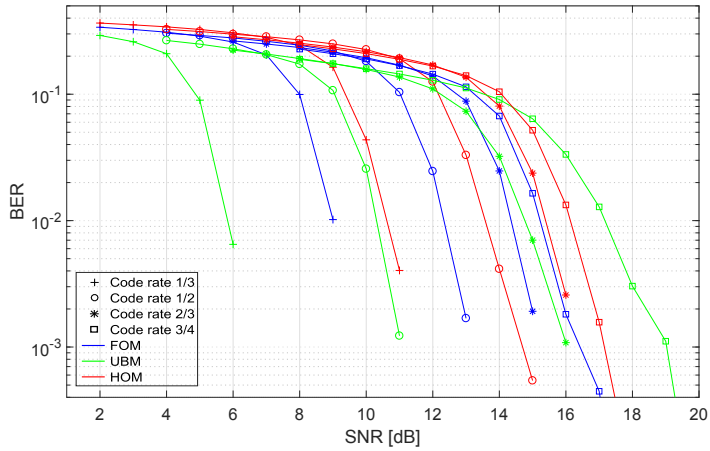


Figure 10: The coded BER with different code-rates under EPR-4 channel. At code-rate 2/3 and 3/4, the FOM shortener is superior to the UBM shortener, while at all code-rates, the FOM shortener is better than the HOM shortener.

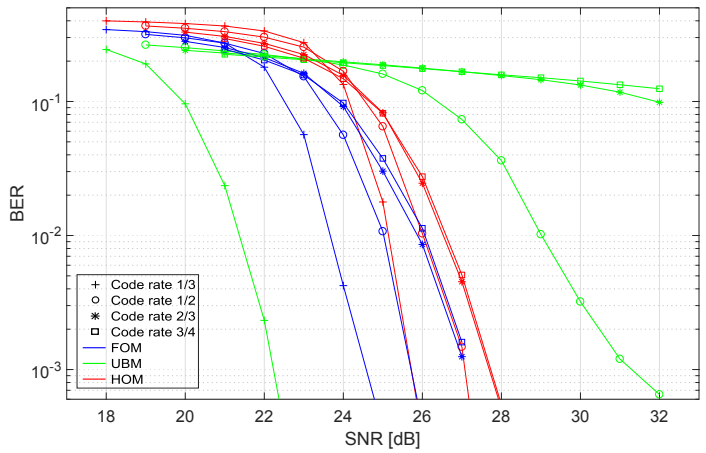


Figure 11: The coded BER with different code-rates under Proakis-C channel. Truncating the channel tails in the UBM shortener renders significantly performance losses at code-rates higher than 1/2. The FOM shortener is better than the HOM shortener for all code-rates.

outperforms the FOM shortener only when the normalized MI is less than 1/2, and with higher MI the UBM is inferior.

6.6 Performance Evaluation with Statistical Channel

In Fig. 12, we evaluate the coded BER under 8-tap IID Gaussian channels, where each channel-tap is a real Gaussian variable with zero-mean and variance 1/8, that is, the total power of the ISI channel is 1. We test with 8PSK modulation and a code-rate 2/3. As the ISI channel is statistical, we set $\sigma = 1 - \delta_{\text{mse}}/2$ according to (40) for the parameter design of the FOM shortener. As can be seen from Fig. 12, with $\nu = 1$ and 2, the FOM shortener

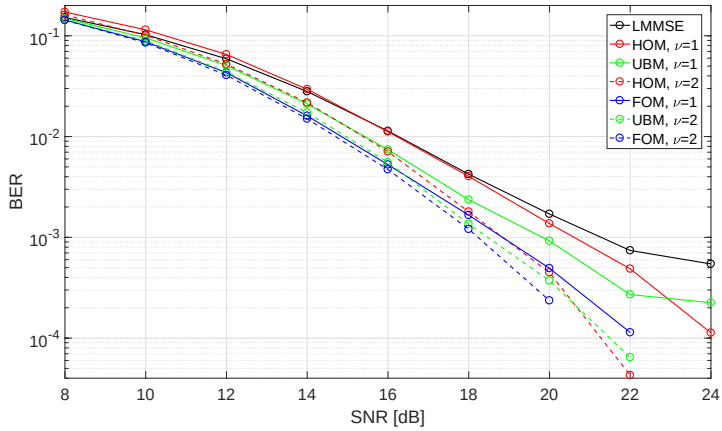


Figure 12: The coded BER with a code-rate 2/3 under 8-tap IID Gaussian channels with 8PSK modulation and $\nu = 1$ and 2.

outperforms the UBM and HOM shorteners more than one 1 dB and 0.5 dB in terms of SNR at $0.1/\%$ BER. Further, the UBM shortener has an error-floor at the high SNR with $\nu = 1$ due to the truncation of the channel tails.

Next, we evaluate the coded BER with ISI channels where the power-delay-profile (PDP) obeys different exponentiation decays. We also test under 8-tap IID Gaussian channels, but now each channel-tap is a complex Gaussian variable with zero-mean and the power for the nth $(0 \leq n \leq 7)$ tap equals

$$P_n = \exp(-n\theta) \,/\, \sum_{n=0}^{7} \exp(-n\theta) \,.$$

With $\theta = 0$, the PDP is uniformly distributed with identical power 1/8 for each channel-tap, while with setting θ to be sufficiently large, we obtain a single-tap complex Gaussian channel without ISI. The PDP for different values of θ is depicted in Fig. 13.

In Fig. 14, we evaluate the required SNR for a target coded BER $0.1/\%$ for different setting of θ as in Fig. 13, and with 8PSK modulation and a code-rate 2/3. As it can been see that, the UBM shortener has a large SNR gap compared to the other shorteners due to the high code-rate, while the FOM shortener is superior to the other shorteners in a wide range of θ. The SNR gain of the FOM shortener compared to the UBM and HOM shorteners is around 4 dB and 2dB when θ is relatively small, that is, severe ISI are introduced by the channel. Further, as it can be seen that when θ is smaller than 1, the HOM shortener becomes even inferior to the LMMSE because of the error propagations in the decision feedbacks. As θ increases, all shorteners performs close as the channel energy is focused on the first few channel taps. With the largest values of θ, all shorteners are aligned with LMMSE since now the channel becomes a single-tap channel.

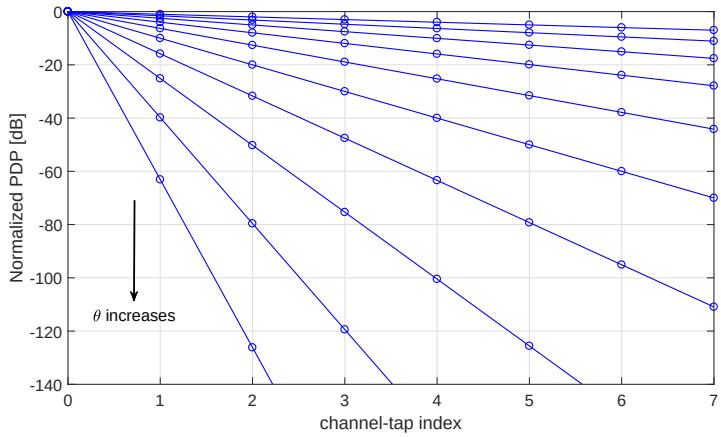


Figure 13: The PDP of the evaluated 8-tap IID complex Gaussian channel with different ISI levels that specified by the values $\theta =$ [0.1000, 0.1585, 0.2512, 0.3981, 0.6310, 1.0000, 1.5849, 2.5119, 3.9811, 6.3096]. The PDP is normalized by the first channel-tap. In general, a small θ introduces strong ISI in the system, and vice visa.

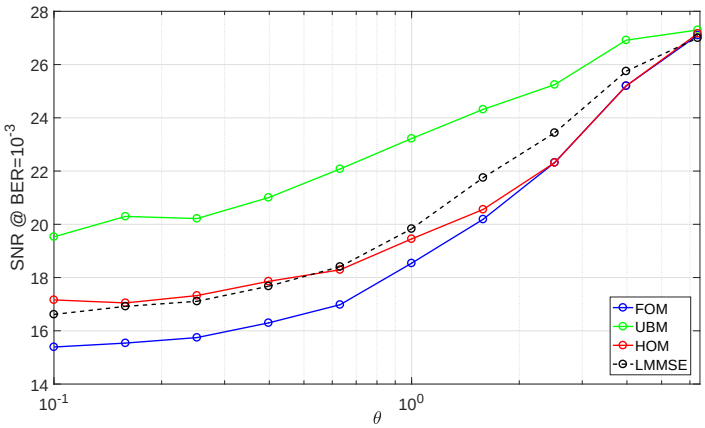


Figure 14: The required SNR for $0.1/\%$ BER with different values of θ.

7 Summary

In this paper, we consider the mutual information lower bound (MILB) based channel shortener design that used in conjunction with the reduced-state soft-output Viterbi equalizer (RS-SOVE), namely, the FOM shortener. We show that the FOM channel shortener cooperating with the RS-SOVE has major gains over the Ungerboeck detection model based channel shortener, namely, the UBM shortener, at medium and high code-rates. Due to the lack of probabilistic meaning, the UBM shortener truncates the channel tails and utilizes no decision-feedback detection. Both the FOM and UBM shorteners significantly outperform the conventional homomorphic filtering based channel shortener, namely, the HOM shortener. We also analyze the theoretical information rates of the proposed FOM channel shortener in relation to the Shannon capacity and the previous channel shortener

designs. In addition, we extend the RS-SOVE to an arbitrary delay that can be larger than the duration of the intersymbol interference (ISI) channel, and we show that, the trellis search process is equivalent to a full forward recursion and a backward recursion with a depth that equals the delay.

Appendix A: The Proof of Theorem 1

The DTFT of $\boldsymbol{w}$ reads

$$W(\omega) = \sum_{k=-\infty}^{\infty} w_k \exp(jk\omega)\,,$$

and the differential of I_{LB} in (27) with respect to w_k is

$$\begin{aligned}\frac{\partial I_{\mathrm{LB}}}{\partial w_k} &= -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|F(\omega)|^2\big(N_0+|H(\omega)|^2\big)W^*(\omega)}{1+|F(\omega)|^2}\exp(jk\omega)\,\mathrm{d}\omega \\ &\quad +\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(F^*(\omega)H(\omega)+\frac{\sigma|F(\omega)|^2H(\omega)B^*(\omega)}{1+|F(\omega)|^2}\Big)\exp\big(jk\omega\big)\mathrm{d}\omega. \qquad (49)\end{aligned}$$

As (49) shall equal zero for all k, the optimal $W(\omega)$ is in (28). Inserting $W_{\mathrm{opt}}(\omega)$ back into (27) yields

$$\begin{aligned}I_{\mathrm{LB}} = 1+\frac{\sigma}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)B(\omega)M(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big) \\ +\frac{\tilde{M}(\omega)|B(\omega)F(\omega)|^2}{1+|F(\omega)|^2}+M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega. \qquad (50)\end{aligned}$$

Setting $\sigma=0$, I_{LB} in (50) is then equal to (31). With $0<\sigma\leq 1$, the terms related to $B(\omega)$ in (50) are

$$\mathcal{F}(B(\omega)) = \frac{\sigma}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)B(\omega)M(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|B(\omega)F(\omega)|^2}{1+|F(\omega)|^2}\mathrm{d}\omega. \qquad (51)$$

With $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$ defined in (25) and (26), (51) can be rewritten as

$$\mathcal{F}(B(\omega)) = \boldsymbol{b}\boldsymbol{\varepsilon}_2\boldsymbol{b}^{\dagger}+2\mathcal{R}\big\{\boldsymbol{b}\boldsymbol{\varepsilon}_1\big\}. \qquad (52)$$

Optimizing (52) directly yields $\boldsymbol{b}_{\mathrm{opt}} = -\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}$. Then the optimal $B_{\mathrm{opt}}(\omega)$ is given in (29). Inserting $B_{\mathrm{opt}}(\omega)$ back into (50) and after some manipulations, I_{LB} with the optimal $W_{\mathrm{opt}}(\omega)$ and $B_{\mathrm{opt}}(\omega)$ is then in (30).

Appendix B: Proof of Inequality (a) in (39)

Let $\boldsymbol{e} = \hat{\boldsymbol{x}}^{\mathrm{LMMSE}} - \hat{\boldsymbol{x}}$, where $\hat{\boldsymbol{x}}$ is the hard decision corresponding to estimate $\hat{\boldsymbol{x}}^{\mathrm{LMMSE}}$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left(\boldsymbol{x}-\hat{\boldsymbol{x}}^{\mathrm{LMMSE}}\right)\left(\boldsymbol{x}-\hat{\boldsymbol{x}}^{\mathrm{LMMSE}}\right)^{\dagger}\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}-\hat{\boldsymbol{x}}-\boldsymbol{e})(\boldsymbol{x}-\hat{\boldsymbol{x}}-\boldsymbol{e})^{\dagger}\right] \\
&= \left(1-P_e^{\mathrm{LMMSE}}\right)\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\dagger}\right] + P_e^{\mathrm{LMMSE}}\left(\mathbb{E}\left[(\boldsymbol{x}-\hat{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{x}})^{\dagger}\right] + \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\dagger}\right]\right) \\
&\geq P_e^{\mathrm{LMMSE}}\mathbb{E}\left[(\boldsymbol{x}-\hat{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{x}})^{\dagger}\right].
\end{aligned}$$

Assuming $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are independent for higher-order modulations, it holds that

$$\mathbb{E}\left[(\boldsymbol{x}-\hat{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{x}})^{\dagger}\right] = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\dagger}\right] + \mathbb{E}\left[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\dagger}\right] = 2.$$

Therefore, the below inequality holds,

$$P_{\mathrm{e}}^{\mathrm{LMMSE}} \leq \mathbb{E}\left[\left(\boldsymbol{x}-\hat{\boldsymbol{x}}^{\mathrm{LMMSE}}\right)\left(\boldsymbol{x}-\hat{\boldsymbol{x}}^{\mathrm{LMMSE}}\right)^{\dagger}\right] \Big/ 2.$$

Appendix C: Proof of Property 2

As the HOM shortener is a special case of the FOM shortener, by definition $I_{\mathrm{HOM}} \leq I_{\mathrm{FOM}}^{0}$ holds. With $\sigma = 0$ and from Theorem 1, by identifying $G(\omega) = |F(\omega)|^2$, I_{LB} can be written in the same form as in (37). As the UBM shortener maximizes (37) under constraint that $1+G(\omega) \geq 0$ for all ω, which is also true for stetting $G(\omega) = |F(\omega)|^2$, therefore, $I_{\mathrm{FOM}}^{0} \leq I_{\mathrm{UBM}}$ holds.

Next, we prove $I_{\mathrm{UBM}} \leq \mathcal{C}$. Note that,

$$G(\omega) = 2\mathcal{R}\left\{g_0 + \sum_{k=1}^{\nu} g_k \exp\left(jk\omega\right)\right\}.$$

Taking the differential of I_{UBM} in (37) with respect to g_k and g_k^*results in

$$\int_{-\pi}^{\pi} \frac{\exp\left(jk\omega\right)}{1+G(\omega)}\mathrm{d}\omega = -\int_{-\pi}^{\pi} M(\omega)\exp\left(jk\omega\right)\mathrm{d}\omega, \ -\nu \leq k \leq \nu.$$

Hence, the below equality holds with the optimal $G(\omega)$, which we denote as $G_0(\omega)$,

$$\frac{1}{1+G_0(\omega)} + M(\omega) = 2\mathcal{R}\left\{\sum_{|k|>\nu} \tau_k \exp\big(jk\omega\big)\right\}, \tag{53}$$

for some constants τ_k. On the other hand, as

$$G_0(\omega) = 2\mathcal{R}\left\{\hat{g}_0 + \sum_{k=1}^{\nu} \hat{g}_k \exp\big(jk\omega\big)\right\}, \tag{54}$$

for some $\hat{\boldsymbol{g}} = (\hat{g}_0, \hat{g}_1, \cdots, \hat{g}_\nu)$, multiplying both sides in (53) with $(1 + G_0(\omega))$ results in

$$1+M(\omega)(1+G_0(\omega)) = -2\,(1+G_0(\omega))\mathcal{R}\left\{\sum_{k>\nu} \tau_k \exp\big(jk\omega\big)\right\}. \tag{55}$$

Integrating (55) over ω in $[-\pi, \pi)$ and utilizing (54) lead to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\,(1+G_0(\omega))\,\mathrm{d}\omega = -1. \tag{56}$$

Therefore, with $G_0(\omega)$, I_{LB} in (37) equals

$$I_{\mathrm{UBM}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(1+G_0(\omega))\mathrm{d}\omega.$$

As the logarithm function is concave, from the definition of $M(\omega)$ in (22) and utilizing (56),

$$\begin{aligned} I_{\mathrm{UBM}} - \mathcal{C} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(1+G_0(\omega))\mathrm{d}\omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(1+\frac{|H(\omega)|^2}{N_0}\right)\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\big(-M(\omega)\,(1+G_0(\omega))\big)\mathrm{d}\omega \\ &\leq \log\left(-\frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\,(1+G_0(\omega))\,\mathrm{d}\omega\right) \\ &= 0. \end{aligned}$$

Therefore, $I_{\mathrm{UBM}} \leq \mathcal{C}$ holds which completes the proof.

References

[1] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "A Low-complexity channel shortening receiver with diversity support for evolved 2G device," IEEE Int. Conf. Commun. (ICC), Kuala Lumpur, Malaysia, May 2016.

[2] 3GPP, TS 36.201, *Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer*, Release 13, Jun. 2016.

[3] Ericsson, White Paper, "Cellular networks for massive IoT," Jan. 2016.

[4] G. Colavolpe, A. Modenini, and F. Rusek, "Channel shortening for nonlinear satellite channels," *IEEE Commun. Lett.*, vol. 16, no. 12, pp. 1929-1932, Dec. 2012.

[5] Y. Chen and L. M. Davis, "Single carrier filtering system architecture for flexible frequency domain multiplexing uplink," Int. Conf. Comm. Workshop (ICCW), Jun. 2015, pp. 1048-1053.

[6] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Commun. Mag.*, vol. 40, no. 4, pp. 58-66, Aug. 2002.

[7] A. Duel-Hallen and C. Heegard, "Delayed decision-feedback sequence estimation," *IEEE Trans. Commun.*, vol. 37, no. 5, pp. 428-436, May 1989.

[8] F. Sun and T. Zhang, "Quasi-reduced-state soft-output Viterbi detector for magnetic recording read channel," *IEEE Trans. Mag.,* vol. 43, no. 10, pp. 3921-3924, Oct. 2007.

[9] L. Reggiani, G. Tartara, and G. M. Maggio, "A reduced-state soft input soft output algorithm based on state partitioning," IEEE Global Telecommun. Conf. (GLOBECOM), Nov. 2001, pp. 906-910.

[10] X. Chen and K. M. Chugg, "Reduced-state soft-input/soft-output algorithms for complexity reduction in iterative and non-iterative data detection," *Proc.* IEEE Int. Conf. Commun. (ICC), New Orleans, U.S.A., Jun. 2000, pp. 6-10.

[11] M. V. Eyuboglu, S. U. H. Qureshi, "Reduced-state sequence estimation with set partitioning and decision feedback," *IEEE Trans. Commun.*, vol. 36, no. 1, pp. 13-20, Jan. 1988.

[12] T. Hashimoto, "A list-type reduced-constraint generalization of the Viterbi algorithm," *IEEE Trans. Inf. Theory,* vol. 33, no. 6, pp. 866-876, 1987.

[13] B. Penther, D. Castelain, and H. Kubo, "A modified turbo detector for long delay spread channels," in In. Symp. Turbo Codes, Brest, France, Sep. 2000.

[14] G. D. Forney Jr., "Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 363-378, May 1972.

[15] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo Codes," *Proc.* IEEE Int. Conf. Commun. (ICC), Geneva, Switzerland, May 1993, pp. 1064-1070.

[16] R. G. Gallager, *Low-density parity check codes over GF(q),* MIT press, Cambridge, MA, 1962.

[17] J. Hagenauer and P. Hoeher, "A Viterbi algorithm with soft-decision outputs and its applications," IEEE Global Telecommun. Conf. (GLOBECOM), Dallas, Texas, U.S.A., Nov. 1989, pp. 1680-1686.

[18] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 284-287, Mar. 1974.

[19] W. Koch and A. Baier, "Optimum and sub-optimum detection of intersymbol interference," IEEE Global Telecommun. Conf. (GLOBECOM), vol. 3, Dec. 1990, pp. 1679-1984.

[20] S. Hu and F. Rusek, "On the design of reduced state demodulators with interference cancellation for iterative receivers," IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC), Hongkong, Sep. 2015, pp. 981-985.

[21] F. Rusek and A. Prlja, "Optimal channel shortening of MIMO and ISI channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 810-818, Feb. 2012.

[22] A. V. Oppenheim and R. W. Schafer, *Digital signal processing*, the first edition, Englewood Cliffs, Prentice-Hall, 1989.

[23] W. H. Gerstacker, F. Obernosterer, R. Meyer, and J. B. Huber, "On prefilter computation for reduced-state equalization," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 793-800, Oct. 2002.

[24] N. Al-Dhahir, "FIR channel-shortening equalizers for MIMO ISI channels," *IEEE Trans. Commun.*, vol. 49, no. 2, pp. 213-218, Feb. 2001.

[25] C. Benkeser, S. Zwicky, H. Kröll, J. Widmer, and Q. Huang, "Efficient channel shortening for higher order modulation: Algorithm and architecture," *Proc.* IEEE Int. Symp. Circuits and Syst (ISCAS), May 2012, pp. 2377-2380.

[26] H. Kröll, S. Altorfer, T. Willi, A. Burg, and Q. Huang, "Channel shortening and equalization based on information rate maximization for evolved GSM/EDGE," IEEE Workshop on Signal Process. Syst., Oct. 2015, pp. 1-6.

[27] M. Loncar and F. Rusek, "On reduced-complexity equalization based on Ungerboeck and Forney observation models," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3784-3789, Aug. 2008.

[28] F. Rusek, G. Colavolpe, and C. W. Sundberg, "40 Years with the Ungerboeck model: a look at its potentialities [lecture notes]," *IEEE Signal Process. Mag.,* vol. 32, no. 3, pp. 156-161, May 2015.

[29] T. K. Moon, *Error correction coding: mathematical methods and algorithms*, New York, U.S.A., Wiley, 2005.

[30] G. Bauch and V. Franz, "A comparison of soft-in/soft-out algorithms for 'turbo detection'," *Proc.* IEEE Int. Conf. Telecomm., Jun. 1998, pp. 259-263.

[31] W. Hirt, *Capacity and information rates of discrete-time channels with memory*, Ph.D thesis, no. ETH 8671, Inst. Signal and Inf. Process., Swiss Federal Inst. Technol., Zürich, 1988.

[32] U. Grenander and G. Szegö, *Toeplitz forms and their applications*, University of Calif. Press, 2001.

[33] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Commun. and Inf. Theory*, vol. 2, no. 3, pp. 155-239, 2006.

[34] F. Rusek, N. Al-Dhahir, and A. Gomaa, "A rate-maximizing channel-shortening detector with soft feedback side information," IEEE Global Telecommun. Conf. (GLOBECOM), Anaheim, CA, pp. 2256-2261, Dec. 2012.

[35] P. Kabal and S. Pasupathy, "Partial-response signaling," *IEEE Trans. Commun.,* vol. 23, no. 9, pp. 921-934, Sep. 1975.

[36] J. G. Proakis and M. Salehi, *Digital communications*, the fifth edition, McGraw-Hill, 2008.

[37] A. Said and J. B. Anderson, "Bandwidth-efficient coded modulation with optimized linear partial-response signals," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 701-713, Mar. 1998.

[38] C. Benkeser, S. Zwicky, H. Kröll, J. Widmer, and Q. Huang, "Efficient channel shortening for higher order modulation: Algorithm and architecture," IEEE Int. Symp. Circuits and Syst. (ISCAS), 2012, May 2012, pp. 2377-2380.

[39] 3GPP, TS 36.212, *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding*, Release 12, Mar. 2015.

Paper III

A Soft-Output MIMO Detector with Achievable Information Rate based Partial Marginalization

In this paper, we propose a soft-output detector for multi-input multi-output (MIMO) channels that utilizes achievable information rate (AIR) based partial marginalization (PM). The proposed AIR based PM (AIR-PM) detector has superior performance compared to previously proposed PM designs and other soft-output detectors such as K-best, while at the same time yielding lower computational complexity, a detection latency that is independent of the number of transmit layers, and straightforward inclusion of soft input information. Using a tree representation of the MIMO signal, the key property of the AIR-PM is that the connections among all child layers are broken. Therefore, least-square (LS) estimates used for marginalization are obtained independently and in parallel, which have better quality than the zero-forcing decision feedback (ZF-DF) estimates used in previous PM designs. Such a property of the AIR-PM detector is designed via a mismatched detection model that maximizes the AIR. Furthermore, we show that the chain rule holds for the AIR calculation, which facilitates an information theoretic characterization of the AIR-PM detector.

S. Hu and F. Rusek,
"Soft-output MIMO detector with achievable information rate based partial marginalization,"
IEEE Trans. Signal Process., vol. 65, no. 6, pp. 1622-1637, Mar. 2017.

1 Introduction

As specified by 3GPP standardization, user equipment (UE) connected to an LTE/LTE-A system nowadays needs to support downlink multi-input multi-output (MIMO) transmissions with up to 8 transmit antennas. In the foreseeable future, many communication systems will be based upon full-dimension MIMO or massive MIMO that immerse in 5G [1], and the number of transmit antennas will increase to 16 or more. Meanwhile, the number of antennas in the UE has also rapidly increased [2] to facilitate downlink transmissions with a large number of spatial layers. This requires efficient design of low-complexity and hardware-friendly MIMO detectors.

A review of state-of-the-art MIMO detection techniques can be found in, e.g., [3]. Although maximum-likelihood (ML) detection [7] provides optimal performance, the complexity, however, increases exponentially in the number of transmit antennas N (for simplicity, we assume that the number of transmit layers equals the number of transmit antennas) and the cardinality of the signal constellation $\mathcal{X}$. Therefore, ML is only applicable for settings with small N, and low-order modulation such as 4 quadrature amplitude modulation (4-QAM). For UEs that support MIMO with large N and 16-QAM or higher modulation orders, the computational cost of ML is prohibitive. On the other hand, linear detectors such as zero-forcing (ZF) and linear minimum mean square error (LMMSE), have low computational cost but suffer from rate losses, especially when the MIMO channel is spatially correlated.

Other detectors, such as sphere decoding (SD) [8, 9] and list decoding [10], usually have random detection complexities. Fixed-complexity sphere decoding (FCSD) [11], and breadth-first SD [12] variants such as the QR-decomposition based M-algorithm[13] and the K-best algorithm[14], have been developed to overcome such an obstacle. However, these detectors are efficient in finding the ML path in the tree-search which generates the hard decisions, but not the best competing paths. It can occur that no counter-hypothesis is available in the set of survival paths for calculating the log-likelihood ratio (LLR) of a certain bit. Then, approaches such as using the difference between the best and worst metrics among all the paths as the soft information[15], flipping the desired bit of the ML path[16], and assigning a predefined-value, are adopted to approximate the LLR. However, such approximations degrade the quality of LLRs sent to the error-correcting decoder.

Another approach to reduce the detection complexity is through partial marginalization (PM) [17, 18], which carefully selects ν layers (which we refer to as parent layers) out of N layers, and marginalizes over the remaining layers (which we refer to as child layers) using ZF with decision feedback (ZF-DF) estimates. The advantages of PM are that it has fixed complexity, and that the detection is parallelizable for parent layers. However, there are also drawbacks. Firstly, as the ZF-DF process based on QR-decomposition is

needed for each bit assumption of the child layers, the process is heavy for large N and the quality of the ZF-DF estimates also degrades. Secondly, as ZF-DF utilizes a sequential detection, it is not parallelizable for child layers, rendering a detection latency that is linear in the number of child layers. Lastly, when dealing with soft information, such as with an iterative detection and decoding receiver, the ZF-DF process needs to solve a nonlinear equation at each iteration [18].

In order to address such issues and further reduce the detection complexity, it is of interest to improve the ZF-DF process in PM. An intuitive idea is to break the connections among child layers through a pre-process such that, sub-optimal estimates are easy to find while all advantages of PM are still preserved. This brings us to design a detection model based on maximization of the achievable information rate (AIR). The AIR, which is a generalized mutual information[19, 20] that the transceiver system can achieve with a mismatched channel model at the receiver, is first considered in [21] for designing a reduced-state detector in inter-symbol interference channels. In [22] and successively [23–26], the authors extend the AIR based reduced-state detection to any linear vector channel with a closed form optimization procedure. The advantages of the framework in [22] are that the detection model can be chosen at will, and that the parameters are easy to optimize.

In light of [17] and [22], we develop the AIR-PM detector in this paper. After a pre-process, all connections among child layers are removed in the effective channel and results in a special structure. This structure, which will be made precise in forthcoming sections, significantly reduces the computational cost of finding estimates for the child layers. The AIR-PM detector also provides a simpler and fully parallel hardware structure both for parent and child layers. Another advantage of the AIR-PM detector is that the order among the child layers and the parent layers is irrelevant. Hence, finding the optimal ordering with AIR-PM collapses into the much simpler problem of selecting the best ν parent layers out of N layers.

The main contribution of the paper is that we propose a two-step soft-output AIR-PM MIMO detector. The first step is pre-processing. With a user-defined parameter ν, ν layers that maximize the AIR are selected as parent layers, with the remaining layers being child layers. The parameters to describe the new detection model are solved in closed forms. Then, in the second step, a full tree description of the signal space is maintained for parent layers, and the marginalization over child layers utilizes least-square (LS) estimates. Different from the PM detector, as there are no connections among child layers, obtaining the LS estimates is straightforward, and the marginalization through AIR-maximization guarantees that an optimal AIR is attained. Furthermore, we show that the chain rule holds for the AIR calculation of the AIR-PM detector, and analyze the properties of ergodic AIR in correlated Rayleigh fading channels.

The rest of the paper is organized as follows. In Sec. II, the MIMO signal model and a

review of previous related detectors are given. In Sec. III, the proposed AIR-PM detector is explained in detail. In Sec. IV, the chain rule of the AIR is proved, and the ergodic AIR in correlated Rayleigh fading channels is analyzed. In Sec. V, we extend the AIR-PM detector to multiple detection branches, imperfect channel estimation, soft inputs, and provide Monte-Carlo simulations with finite constellations. In Sec. VI, empirical results on AIR, ergodic AIR, and frame error rate (FER) are presented for a variety of setups. Finally, Sec. VII concludes the paper.

Notation

Throughout this paper, boldface letters indicate vectors and boldface uppercase letters designate matrices. Superscripts $(\cdot)^{-1}$, $(\cdot)^{1/2}$, $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^\dagger$ stand for the inverse, matrix square root, complex conjugate, transpose, and Hermitian transpose, respectively. Furthermore, $\mathbb{E}[\cdot]$ is the expectation operator, $\mathcal{R}\{\cdot\}$ takes the real part, and $\mathcal{I}\{\cdot\}$ takes the imaginary part. The expression $A \propto B$ means that $(A-B)$ equals some constant, and $\boldsymbol{A} \succ \boldsymbol{B}$ means that $(\boldsymbol{A}-\boldsymbol{B})$ is positive-definite. The complex normal distribution is denoted as $\boldsymbol{h} \sim \mathcal{CN}(\boldsymbol{a}, \boldsymbol{R})$, with $\boldsymbol{a}$ and $\boldsymbol{R}$ being the mean and covariance matrix of $\boldsymbol{h}$, respectively. We also reserve $a_{m,n}$ to denote the element at the mth row and nth column of matrix $\boldsymbol{A}$, a_m to denote the mth element of vector $\boldsymbol{a}$, and $\boldsymbol{I}$ to represent the identity matrix. In addition, $\mathrm{vec}\,(\boldsymbol{A})$ stacks the columns of $\boldsymbol{A}$ on top of each other, $\boldsymbol{A}^{\mathrm{diag}}$ denotes a diagonal matrix whose diagonal elements are identical to $\boldsymbol{A}$, and $|\mathcal{X}|$ is the cardinality of the symbol constellation $\mathcal{X}$. Furthermore, we define an operator "$\ominus$" such that

$$(n, N) \ominus \nu = \max(n, N - \nu).$$

2 Signal Model and review of previous related work

Consider a signal model with N transmit antennas and K receiver antennas. The transmitted signal $\boldsymbol{x} \in \mathbb{C}^N$ comprises unit average energy information symbols that belong to a constellation $\mathcal{X}$, with each symbol mapped from M bits. The matrix $\boldsymbol{H} \in \mathbb{C}^{K \times N}$ represents the MIMO communication channel, and the received signal $\boldsymbol{y} \in \mathbb{C}^K$ is modeled as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

where the noise term $\boldsymbol{e} \sim \mathcal{CN}(\boldsymbol{0}, N_0\boldsymbol{I})$. Next, we review some state-of-the-art soft-output MIMO detectors.

2.1 MAP/ML Detection

The conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ according to model (1) reads

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(\pi N_0)^N} \exp\left(-\frac{1}{N_0}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2\right). \tag{2}$$

Denoting x_n^m (equals to 1 or -1) as the mth bit of the nth symbol x_n in $\boldsymbol{x}$, and given the observable $\boldsymbol{y}$ and prior distribution $p(\boldsymbol{x})$, the maximum a posteriori probability (MAP) detector generates the LLR of bit x_n^m as

$$\begin{aligned} L(x_n^m|\boldsymbol{y}) &= \ln \frac{p(x_n^m=1|\boldsymbol{y})}{p(x_n^m=-1|\boldsymbol{y})} \\ &= \ln \frac{\sum\limits_{\boldsymbol{x}:x_n^m=1} p\left(\boldsymbol{y}|\boldsymbol{x}\right) p(\boldsymbol{x})}{\sum\limits_{\boldsymbol{x}:x_n^m=-1} p\left(\boldsymbol{y}|\boldsymbol{x}\right) p(\boldsymbol{x})} \\ &= \ln \frac{\sum\limits_{\boldsymbol{x}:x_n^m=1} \exp\left(\mu(\boldsymbol{y}|\boldsymbol{x}) + \ln p(\boldsymbol{x})\right)}{\sum\limits_{\boldsymbol{x}:x_n^m=-1} \exp\left(\mu(\boldsymbol{y}|\boldsymbol{x}) + \ln p(\boldsymbol{x})\right)}, \end{aligned} \tag{3}$$

where the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ is defined as

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = -\frac{1}{N_0}\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2. \tag{4}$$

With sufficiently good interleaving, we can safely assume the transmit bits to be independent. Therefore, $\ln p(\boldsymbol{x}) = \sum\limits_{n=1}^{N} \ln p(x_n)$ and

$$\ln p(x_n) = \sum_{m=1}^{M} \ln p(x_n^m) = \sum_{m=1}^{M} \left(\frac{(1+x_n^m)\tilde{L}_n^m}{2} - \ln\left(1+\exp\left(\tilde{L}_n^m\right)\right)\right), \tag{5}$$

where $\tilde{L}_n^m$ is *a priori* LLR of bit x_n^m, such as the extrinsic information output from the outer decoder in an iterative detection and decoding receiver. With no prior information, i.e., $\tilde{L}_n^m = 0$, the LLR in (3) equals

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}:x_n^m=1} \exp\left(\mu(\boldsymbol{y}|\boldsymbol{x})\right)}{\sum\limits_{\boldsymbol{x}:x_n^m=-1} \exp\left(\mu(\boldsymbol{y}|\boldsymbol{x})\right)}. \tag{6}$$

The fundamental issue with computing (3) or (6) is that, the sum operation is taken over $\mathcal{O}(|\mathcal{X}|^N)$ terms, which makes direct evaluation infeasible for large $|\mathcal{X}|$ or N. To reduce the

computational complexity, the ML with Max-Log approximation (MLM) detector utilizes the Jacobian approximation [28] $\ln(e^a + b^a) \approx \max(a, b)$ to approximate the LLR calculation in (6) as

$$L(x_n^m|\boldsymbol{y}) = \max_{\boldsymbol{x}:x_n^m=1} \mu(\boldsymbol{y}|\boldsymbol{x}) - \max_{\boldsymbol{x}:x_n^m=-1} \mu(\boldsymbol{y}|\boldsymbol{x}). \tag{7}$$

However, finding the maximum of $\mu(\boldsymbol{y}|\boldsymbol{x})$ over $\boldsymbol{x}$ is an NP-hard optimization problem, and the MLM still suffers from intensive computations. An alternative approach, is to only keep a small number of survival paths in (7), such as in the K-best detector.

2.2 K-best Detector

The K-best detector is a variant of the breadth-first SD detector. After an optimal reordering of the columns in $\boldsymbol{H}$, the QR-decomposition $\boldsymbol{Q}^\dagger \boldsymbol{H} = \boldsymbol{R}$ is implemented, where $\boldsymbol{Q}$ is unitary and $\boldsymbol{R}$ is upper-triangular. Let $\tilde{\boldsymbol{y}} = \boldsymbol{Q}^\dagger \boldsymbol{y}$, the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (4) is computed as

$$\begin{aligned}\mu(\boldsymbol{y}|\boldsymbol{x}) &= -\frac{1}{N_0}\|\boldsymbol{Q}^\dagger \boldsymbol{y} - \boldsymbol{R}\boldsymbol{x}\|^2 \\ &= -\frac{1}{N_0}\left|\tilde{y}_N - r_{N,N}x_N\right|^2 - \sum_{n=N-1}^{1} \underbrace{\frac{1}{N_0}\left|\tilde{y}_n - r_{n,n}x_n - \sum_{k=N}^{n+1} r_{n,k}\hat{x}_k\right|^2}_{\mu(\boldsymbol{y}|x_n, x_{n+1}, \cdots, x_N)}.\end{aligned}$$

For the nth layer, when computing $\mu(\boldsymbol{y}|x_n, x_{n+1}, \cdots, x_N)$, the K best hard decisions of symbol vector $(\hat{x}_{n+1}, \hat{x}_{n+2}, \cdots, \hat{x}_N)$ are used. The advantages of K-best detector over other SD detectors are that, it has fixed complexity and is easy to implement in hardware.nThe drawback is, without the best counter-hypotheses found to maximize $\mu(\boldsymbol{y}|\boldsymbol{x})$ for each of the bit assumptions, the LLR values calculated through (7) will not comply with their true probabilities. At worst cases, the counter-hypotheses may be missing for some of the bit assumptions and the LLRs have to be saturated to some limit values. As a result, these large LLRs may become difficult to be corrected by channel decoder and degrade the decoding performance. Such an issue of 'LLR overestimating' is commonly encountered with suboptimal detectors such as the LMMSE and the soft LMMSE that cooperates *a priori* information into the filter [26, 27]. With PM detector, the LLRs might still be overestimated (due to the partial marginalization), but the counter-hypotheses for all bits are evaluated and the corresponding metrics are preserved, which yield better estimates of the LLRs than those produced by the K-best detector.

2.3 Partial Marginalization (PM) Detector

With PM detector, $\boldsymbol{x}$ (after re-ordering) is split into two parts $(\boldsymbol{x}_b, \boldsymbol{x}_a)$, which we denote as

$$\begin{aligned}\boldsymbol{x}_a &= (x_{N-\nu+1}, x_{N-\nu+2}, \cdots, x_N)\,,\\ \boldsymbol{x}_b &= (x_1, x_2, \cdots, x_{N-\nu})\,.\end{aligned}$$

The posterior density is marginalized exactly over signal part $\boldsymbol{x}_a$, while the density is approximately marginalized over signal part $\boldsymbol{x}_b$ using its hard decision. The metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (4) can equivalently be written as

$$\mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a) = -\frac{1}{N_0}\|\boldsymbol{y} - \boldsymbol{H}_b\boldsymbol{x}_b - \boldsymbol{H}_a\boldsymbol{x}_a\|^2,$$

where $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$ are the sub-channels corresponding to signals $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$,

$$\begin{aligned}\boldsymbol{H}_a &= (\boldsymbol{h}_{N-\nu+1}, \boldsymbol{h}_{N-\nu+2}, \cdots, \boldsymbol{h}_N)\,,\\ \boldsymbol{H}_b &= (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{N-\nu})\,.\end{aligned}$$

With partial marginalization over $\boldsymbol{x}_b$, the LLRs corresponding to bit x_n^m in $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are approximated [17] as

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}_a: x_n^m = 1} \exp\left(\max\limits_{\boldsymbol{x}_b} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}{\sum\limits_{\boldsymbol{x}_a: x_n^m = -1} \exp\left(\max\limits_{\boldsymbol{x}_b} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}, \tag{8}$$

and

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}_a} \exp\left(\max\limits_{\boldsymbol{x}_b: x_n^m = 1} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}{\sum\limits_{\boldsymbol{x}_a} \exp\left(\max\limits_{\boldsymbol{x}_b: x_n^m = -1} \mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)\right)}, \tag{9}$$

respectively. In previous PM designs, maximizing $\mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)$ under each assumption of $\boldsymbol{x}_a$ utilizes a ZF-DF estimate of $\boldsymbol{x}_b$ in (8). However, in (9) the ZF-DF needs to be implemented for each bit assumption of $\boldsymbol{x}_b$, which requires $(N-\nu)(M+1)$ operations for each given $\boldsymbol{x}_a$.

To simplify the computational cost and improve the quality of ZF-DE estimates, we propose the AIR-PM detector, with the target to design a detection model such that, maximizing $\mu(\boldsymbol{y}|\boldsymbol{x}_b, \boldsymbol{x}_a)$ is independent over layers in $\boldsymbol{x}_b$, while an optimal AIR is attained under such a constraint. The LLR computing in AIR-PM detector will still use (8) and (9), but the computations are simplified. In the next section, we will elaborate on the AIR-PM detector in detail.

3 Soft MIMO Detector with AIR-Maximization based PM

The intention of the AIR-PM detector is to prefilter the received signal to: (a) break dependencies among symbols in $\boldsymbol{x}_b$, (b) preserve dependencies among symbols in $\boldsymbol{x}_a$, and (c) preserve dependencies between symbols in $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$. Then, we form a search tree where we fully exhaust all possibilities for $\boldsymbol{x}_a$ as parent layers, and for each assumption, symbols in $\boldsymbol{x}_b$ are detected in parallel with LS estimates, which is optimal for the given model since there are no connections among them. As $\boldsymbol{x}_b$ only connects to $\boldsymbol{x}_a$, the tree search is significantly simplified with AIR-PM.

3.1 AIR-Maximization Detection Model

The metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (4) can be expressed as

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{N_0}\left(2\mathcal{R}\{\boldsymbol{x}^\dagger \boldsymbol{H}^\dagger \boldsymbol{y}\} - \boldsymbol{x}^\dagger \boldsymbol{H}^\dagger \boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}^\dagger \boldsymbol{y}\right).$$

As the last term $\boldsymbol{y}^\dagger \boldsymbol{y}$ is irrelevant for detection, it can be removed. We generalize $\mu(\boldsymbol{y}|\boldsymbol{x})$ as

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = 2\mathcal{R}\{\boldsymbol{x}^\dagger \boldsymbol{H}_{\mathrm{r}}^\dagger \boldsymbol{y}\} - \boldsymbol{x}^\dagger \boldsymbol{G}_{\mathrm{r}} \boldsymbol{x}, \tag{10}$$

which corresponds to the detection model

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\left(2\mathcal{R}\{\boldsymbol{x}^\dagger \boldsymbol{H}_{\mathrm{r}}^\dagger \boldsymbol{y}\} - \boldsymbol{x}^\dagger \boldsymbol{G}_{\mathrm{r}} \boldsymbol{x}\right), \tag{11}$$

instead of the true conditional probability (2). Without loss of generality, the term N_0 is absorbed into both $\boldsymbol{H}_{\mathrm{r}}$ and $\boldsymbol{G}_{\mathrm{r}}$. With AIR-PM, we constrain $\boldsymbol{G}_{\mathrm{r}}$ to be Hermitian and with a shape illustrated in Fig. 1(a). That is, only the elements along the main diagonal, the last ν rows, and the last ν columns of $\boldsymbol{G}_{\mathrm{r}}$ can be non-zero. With such constraints on $\boldsymbol{G}_{\mathrm{r}}$, and by defining

$$\tilde{\boldsymbol{y}} = \boldsymbol{H}_{\mathrm{r}}^\dagger \boldsymbol{y}, \tag{12}$$

we have the below proposition of the metric decomposition for AIR-PM detector by expanding $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (10) directly.

Proposition 1. *With detection model (11), the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (10) can be rewritten as*

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = \sum_{n=1}^{N-\nu} \mu_1^n\left(\tilde{y}_n | x_n, \boldsymbol{x}_a\right) + \sum_{n=N-\nu+1}^{N} \mu_2^n\left(\tilde{y}_n | x_n, x_{n+1}, \cdots, x_N\right), \tag{13}$$

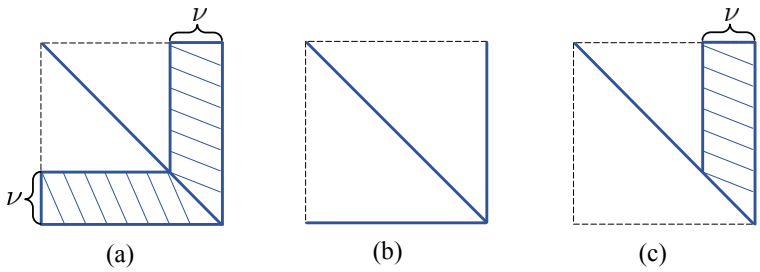


Figure 1: A general shape of $\boldsymbol{G}_{\mathrm{r}}$ is depicted in (a), and (b) is $\boldsymbol{G}_{\mathrm{r}}$ with $\nu=1$. The shape of $\boldsymbol{U}$, which satisfies the decomposition $\boldsymbol{I}+\boldsymbol{G}_{\mathrm{r}}=\boldsymbol{U}^{\dagger}\boldsymbol{U}$, is depicted in (c).

where $\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)$ *and* $\mu_2^n\left(\tilde{y}_n|x_n,x_{n+1},\cdots,x_N\right)$ *are defined as*

$$\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)=2\mathcal{R}\left\{\left(\tilde{y}_n-\sum_{k=N-\nu+1}^{N}g_{n,k}x_k\right)x_n^*\right\}-g_{n,n}|x_n|^2, \quad (14)$$

$$\mu_2^n\left(\tilde{y}_n|x_n,x_{n+1},\cdots,x_N\right)=2\mathcal{R}\left\{\left(\tilde{y}_n-\sum_{k=n+1}^{N}g_{n,k}x_k\right)x_n^*\right\};-g_{n,n}|x_n|^2. \quad (15)$$

Note that, in general the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (10) does not correspond to a Euclidean distance when $\boldsymbol{G}_{\mathrm{r}}$ is not positive definite, i.e., the Cholesky decomposition of $\boldsymbol{G}_{\mathrm{r}}$ is not available. However, the path metric can still be computed based on (14)-(15). From (14), $\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)$ can be computed in parallel for all symbols $x_n\in\boldsymbol{x}_b$ for a given $\boldsymbol{x}_a$. The detection complexity (measured by the total number of paths that have been visited in the tree search), in this case, is $\mathcal{O}(|\mathcal{X}|^{\nu})$. By setting $\nu=0$ and $\nu=N-1$, the model (11) becomes identical to LMMSE and ML detection, respectively. This shows trade-off between detection complexity and performance through parameter ν.

In [29], the authors have proposed a WL-decomposition (WLD) based detection, which utilizes subspace orthogonalization to decompose the channel as

$$\boldsymbol{W}^{\dagger}\boldsymbol{H}=\boldsymbol{L}. \quad (16)$$

The matrix $\boldsymbol{L}$ is a punctured version of the original channel $\boldsymbol{H}$ and the shape can be specified at will. As a special case, with setting $\nu=1$, the shape Fig. 1(c) is equivalent to the shapes depicted in Fig. 1(b)-(e) in [29] after permutations of the columns. However, there are some substantial differences between the WLD detector and the proposed AIR-PM detector.

Firstly, with (16) the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (10) becomes

$$\mu(\boldsymbol{y}|\boldsymbol{x})=-\frac{1}{N_0}\|\boldsymbol{W}^{\dagger}\boldsymbol{y}-\boldsymbol{L}\boldsymbol{x}\|^2=\frac{1}{N_0}\left(2\mathcal{R}\{\boldsymbol{x}^{\dagger}\boldsymbol{L}^{\dagger}\boldsymbol{W}^{\dagger}\boldsymbol{y}\}-\boldsymbol{x}^{\dagger}\boldsymbol{L}^{\dagger}\boldsymbol{L}\boldsymbol{x}-\boldsymbol{y}^{\dagger}\boldsymbol{y}\right). \quad (17)$$

Removing the constant term $\boldsymbol{y}^{\dagger}\boldsymbol{y}$ from (17) and comparing it with (10), the WLD detector is thus a special case of the AIR-PM detector with identifying $\boldsymbol{H}_{\mathrm{r}}=\boldsymbol{W}\boldsymbol{L}$ and $\boldsymbol{G}_{\mathrm{r}}=\boldsymbol{L}^{\dagger}\boldsymbol{L}$.

Therefore, from the point of view of maximizing the AIR, the WLD detector is inferior to the proposed AIR-PM detector.

Secondly, as $\boldsymbol{W}$ is not unitary with the WLD detector, after WL decomposition the noise will be colored since

$$\mathbb{E}\left[\boldsymbol{W}^{\dagger}\boldsymbol{e}\boldsymbol{e}^{\dagger}\boldsymbol{W}\right] = N_0\boldsymbol{W}^{\dagger}\boldsymbol{W}.$$

Although with normalizing the column vectors of $\boldsymbol{W}$ to have unit Frobenius norms, the noise power remains the same, but the colored noise will degrade the detection performance without taking it into consideration [30]. Especially under the case that $\boldsymbol{H}$ is ill-conditioned, the WL-decomposition results in significant noise-enhancement which needs to be taken care of. To overcome such an obstacle, the authors have proposed to recalculate the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ based on the original channel $\boldsymbol{H}$ with the obtained survival paths [31]. However, such an approach significantly increases the computational cost for large values of n_{t} and M. Although with the AIR-PM detector the noise is also colored by utilizing the model (11), but with the optimization over the parameters $\boldsymbol{G}_{\mathrm{r}}$ and $\boldsymbol{H}_{\mathrm{r}}$, the AIR-PM detector guarantees that an optimal AIR is attained with the detection model (11).

Lastly, with AIR-PM detector we calculate the metric for each of the bit assumptions of the child layers, while the WLD detector utilizes multiple detection branches to calculate the LLRs alternatively for all layer and fetch the minimal metrics $\mu(\boldsymbol{y}|\boldsymbol{x})$ among all branches. Although with the same values of ν and E in AIR-PM and WLD detectors, respectively, the WLD detector has slightly less complexity than the AIR-PM detector with multiple branches, but the AIR-PM detector with a single detection branch is simpler than the WLD detector.

Next, we describe the design of the optimal parameters $\boldsymbol{G}_{\mathrm{r}}$ and $\boldsymbol{H}_{\mathrm{r}}$ based on AIR-maximization.

3.2 Parameter Optimization

With model (11), the AIR is defined as

$$I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\ln\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{y}}\left[\ln\tilde{p}(\boldsymbol{y})\right], \tag{18}$$

where the expectations are taken over the true channel statistics. Following the approach in [22], and under the assumption that $\boldsymbol{x}$ is complex Gaussian distributed[1], a closed form for $I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right)$ can be reached. Optimizing (18) over $K\times N$ prefilter matrix $\boldsymbol{H}_{\mathrm{r}}$ yields

$$\boldsymbol{H}_{\mathrm{r}} = \boldsymbol{W}^{\dagger}\left(\boldsymbol{I} + \boldsymbol{G}_{\mathrm{r}}\right), \tag{19}$$

[1]The complex Gaussian assumption on $\boldsymbol{x}$ is made to derive a closed form $I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right)$, and based on that, $\boldsymbol{G}_{\mathrm{r}}$ and $\boldsymbol{H}_{\mathrm{r}}$ are optimized. We show later that, the parameters optimized for Gaussian inputs work well for finite constellations.

where $\boldsymbol{W}$ is the LMMSE filter

$$\boldsymbol{W} = \boldsymbol{H}^{\dagger}\left(\boldsymbol{H}\boldsymbol{H}^{\dagger} + N_0\boldsymbol{I}\right)^{-1}, \tag{20}$$

and $\boldsymbol{B}$ is the mean-square-error (MSE) matrix

$$\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{H}. \tag{21}$$

The resulting $I_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x})$ equals

$$I_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x}) = N + \ln\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{r}}) - \mathrm{Tr}\big(\boldsymbol{B}(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{r}})\big), \tag{22}$$

where $\boldsymbol{G}_{\mathrm{r}}$ is chosen such that $(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{r}}) \succ \boldsymbol{0}$, and obtained through maximizing (22) under the constraints stated earlier. This optimization is now treated in detail.

First we denote the principle submatrix at the lower right corner of $\boldsymbol{B}$ generated by removing the first n rows and columns as ($\tilde{\boldsymbol{B}}_n = \boldsymbol{0}$ if $n = N$)

$$\tilde{\boldsymbol{B}}_n = \begin{bmatrix} b_{n+1,n+1} & b_{n+1,n+2} & \cdots & b_{n+1,N} \\ b_{n+2,n+1} & b_{n+2,n+2} & \cdots & b_{n+2,N} \\ \vdots & \vdots & \vdots & \vdots \\ b_{N,n+1} & b_{N,n+2} & \cdots & b_{N,N} \end{bmatrix}. \tag{23}$$

Make the decomposition $\boldsymbol{I}+\boldsymbol{G}_{\mathrm{r}} = \boldsymbol{U}^{\dagger}\boldsymbol{U}$; $\boldsymbol{U}$ is upper-triangular as depicted in Fig. 2(c), with only elements along the main diagonal and the last ν columns being non-zero. That is, $u_{n,k} = 0$, for any $k \neq n$ and $k \leq N-\nu$. Letting

$$\boldsymbol{u}_n = \left(u_{n,(n,N)\ominus\nu+1}, u_{n,(n,N)\ominus\nu+2}, \ldots, u_{n,N}\right), \tag{24}$$

$$\boldsymbol{b}_n = \left(b_{n,(n,N)\ominus\nu+1}, b_{n,(n,N)\ominus\nu+2}, \ldots, b_{n,N}\right), \tag{25}$$

we then have the optimal $\boldsymbol{G}_{\mathrm{r}}$ stated in Proposition 2, whose proof is deferred to Appendix A.

Proposition 2. *The non-zero elements of $\boldsymbol{U}$ that maximize the AIR in (22) are calculated as*

$$u_{n,n} = \begin{cases} \sqrt{\left(b_{n,n} - \boldsymbol{b}_n\tilde{\boldsymbol{B}}^{-1}_{(n,N)\ominus\nu}\boldsymbol{b}_n^{\dagger}\right)^{-1}}, & 1 \leq n < N, \\ 1/\sqrt{b_{N,N}}, & n = N. \end{cases} \tag{26}$$

$$\boldsymbol{u}_n = -u_{n,n}\boldsymbol{b}_n\tilde{\boldsymbol{B}}^{-1}_{(n,N)\ominus\nu}. \tag{27}$$

The optimal $\boldsymbol{G}_{\mathrm{r}}$ is then obtained through $\boldsymbol{G}_{\mathrm{r}} = \boldsymbol{U}^{\dagger}\boldsymbol{U} - \boldsymbol{I}$, and the AIR reads

$$I_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x}) = 2\sum_{n=1}^{N}\ln u_{n,n}. \tag{28}$$

From Proposition 2, the optimal AIR in (28) only depends on the values of $u_{n,n}$, which are calculated through (26). For the first $N-\nu$ layers, only the connections with the last ν layers are preserved through $\boldsymbol{b}_n$ and $\tilde{\boldsymbol{B}}_{(n,N)\ominus\nu}$. For instance, if setting $\nu=1$, all the first $N-1$ layers are only connected to the last layer, and we have $\boldsymbol{b}_n=b_{n,N}$ and $\tilde{\boldsymbol{B}}_{(n,N)\ominus\nu}=b_{N,N}$ for $1\leq n<N$. Then, $u_{n,n}$ $(1\leq n<N)$ equals

$$u_{n,n}=\sqrt{\left(b_{n,n}-\frac{|b_{n,N}|^2}{b_{N,N}}\right)^{-1}}.$$

Under the two extreme cases $\nu=0$ and $\nu=N-1$, we next show that, the AIR-PM detector is identical to the LMMSE and ML detectors, respectively.

For the case $\nu=0$, from Proposition 2 it holds that,

$$\begin{aligned}\boldsymbol{G}_{\mathrm{r}}&=\left(\boldsymbol{B}^{\mathrm{diag}}\right)^{-1}-\boldsymbol{I},\\ \boldsymbol{H}_{\mathrm{r}}&=\boldsymbol{W}^{\dagger}\left(\boldsymbol{B}^{\mathrm{diag}}\right)^{-1}.\end{aligned}$$

Equivalently, we can set $\boldsymbol{G}_{\mathrm{r}}=\boldsymbol{I}$ and

$$\boldsymbol{H}_{\mathrm{r}}=\boldsymbol{W}^{\dagger}\left(\boldsymbol{B}^{\mathrm{diag}}\right)^{-1}\left(\left(\boldsymbol{B}^{\mathrm{diag}}\right)^{-1}-\boldsymbol{I}\right)^{-1}=\boldsymbol{W}^{\dagger}\left((\boldsymbol{W}\boldsymbol{H})^{\mathrm{diag}}\right)^{-1},\tag{29}$$

which is the normalized LMMSE filter[27]. Then from (10), maximizing the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ is equivalent to LMMSE detection.

On the other hand, with $\nu=N-1$, it can be shown from Proposition 2 (or directly taking the differential of I_{AIR} in (22) with respect to $\boldsymbol{G}_{\mathrm{r}}$) that

$$\boldsymbol{G}_{\mathrm{r}}=\boldsymbol{B}^{-1}-\boldsymbol{I}.$$

From (19)-(21) and utilizing the matrix inversion lemma [32], it can be shown that

$$\begin{aligned}\boldsymbol{G}_{\mathrm{r}}&=\frac{1}{N_0}\boldsymbol{H}^{\dagger}\boldsymbol{H},\\ \boldsymbol{H}_{\mathrm{r}}&=\frac{1}{N_0}\boldsymbol{H},\end{aligned}\tag{30}$$

which shows that the AIR-PM detector is equivalent to the ML detector from (10).

Moreover, with Proposition 2 we can obtain Corollary 1, whose proof is in Appendix B. However, we point out the fact that, in general, $\boldsymbol{G}_{\mathrm{r}}\succ\boldsymbol{0}$ does not hold.

Corollary 1. *For $0<N_0<\infty$, the diagonal elements of $\boldsymbol{G}_{\mathrm{r}}$ are positive values.*

3.3 Bit LLR Calculation

The LLR calculation of AIR-PM detector uses (8) and (9), and marginalization over $\boldsymbol{x}_b$ needs to search for an x_n in $\mathcal{X}$ that maximizes $\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)$ in (14). Let

$$\hat{\mu}_1^n\left(\tilde{y}_n|\boldsymbol{x}_a\right)=\max_{x_n}\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right). \tag{31}$$

By taking the derivative of $\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)$ with respect to x_n, and noticing that $g_{n,n}$ is positive as shown in Corollary 1, the LS estimate reads

$$\hat{x}_n=\left[\frac{g_{n,N}x_N-\tilde{y}_n}{g_{n,n}}\right],$$

where '[ ]' rounds the estimate to the nearest symbol in $\mathcal{X}$. In a similar way, we can maximize $\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)$ over bit assumption 1 and -1 for x_n^m, and denote

$$\begin{aligned}\hat{\mu}_{1,+}^{n,m}\left(\tilde{y}_n|\boldsymbol{x}_\nu\right)&=\max_{x_n:x_n^m=1}\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right),\\ \hat{\mu}_{1,-}^{n,m}\left(\tilde{y}_n|\boldsymbol{x}_a\right)&=\max_{x_n:x_n^m=-1}\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right),\end{aligned}$$

respectively. Then, the exact LLR computation of (6) with detection model (11) is in (32), and the approximations of (32) for symbols in $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are in (33) and (34), respectively.

4 Characterization of the AIR with AIR-PM Detector

4.1 Chain Rule of the AIR

With the design of AIR-PM detector and the detection process introduced, in this section we prove that chain rule of AIR holds, and then analyze properties of the ergodic AIR for different selections of $\boldsymbol{x}_a$. We first introduce the below lemma.

Lemma 1. *With $u_{n,n}$ defined in Proposition 2, we have the below equalities,*

$$I\left(\boldsymbol{y};x_n|\boldsymbol{x}_a\right)=2\ln u_{n,n},\ 1\le n\le N-\nu, \tag{35}$$

$$I\left(\boldsymbol{y};\boldsymbol{x}_a\right)=2\sum_{n=N-\nu+1}^{N}\ln u_{n,n},\ N-\nu<n\le N, \tag{36}$$

where $I\left(\boldsymbol{y};x_n|\boldsymbol{x}_a\right)$ and $I\left(\boldsymbol{y};\boldsymbol{x}_a\right)$ are the conditional mutual information and mutual information, respectively.

The proof is in Appendix C. With Lemma 1 and Proposition 2, and also using the chain rule of mutual information, we have the chain rule of the AIR in Proposition 3.

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}:x_n^m=1} \exp\left(\sum\limits_{n=1}^{N-\nu} \mu_1^n\left(y_n|x_n,\boldsymbol{x}_a\right) + \sum\limits_{n=N-\nu+1}^{N} \mu_2^n\left(y_n|x_n,x_{n+1},\cdots,x_N\right)\right)}{\sum\limits_{\boldsymbol{x}:x_n^m=-1} \exp\left(\sum\limits_{n=1}^{N-\nu} \mu_1^n\left(y_n|x_n,\boldsymbol{x}_a\right) + \sum\limits_{n=N-\nu+1}^{N} \mu_2^n\left(y_n|x_n,x_{n+1},\cdots,x_N\right)\right)}. \tag{32}$$

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}_a,x_n^m=1} \exp\left(\sum\limits_{n=1}^{N-\nu} \hat{\mu}_1^n\left(y_n|\boldsymbol{x}_a\right) + \sum\limits_{n=N-\nu+1}^{N} \mu_2^n\left(y_n|x_n,x_{n+1},\cdots,x_N\right)\right)}{\sum\limits_{\boldsymbol{x}_a,x_n^m=-1} \exp\left(\sum\limits_{n=1}^{N-\nu} \hat{\mu}_1^n\left(y_n|\boldsymbol{x}_a\right) + \sum\limits_{n=N-\nu+1}^{N} \mu_2^n\left(y_n|x_n,x_{n+1},\cdots,x_N\right)\right)}. \tag{33}$$

$$L(x_n^m|\boldsymbol{y}) = \ln \frac{\sum\limits_{\boldsymbol{x}_a} \exp\left(\hat{\mu}_{1,+}^{n,m}\left(y_n|\boldsymbol{x}_a\right) + \sum\limits_{k=1,k\neq n}^{N-\nu} \hat{\mu}_1^n\left(y_k|\boldsymbol{x}_a\right) + \sum\limits_{n=N-\nu+1}^{N} \mu_2^n\left(y_n|x_n,x_{n+1},\cdots,x_N\right)\right)}{\sum\limits_{\boldsymbol{x}_a} \exp\left(\hat{\mu}_{1,-}^{n,m}\left(y_n|\boldsymbol{x}_a\right) + \sum\limits_{k=1,k\neq n}^{N-\nu} \hat{\mu}_1^n\left(y_k|\boldsymbol{x}_a\right) + \sum\limits_{n=N-\nu+1}^{N} \mu_2^n\left(y_n|x_n,x_{n+1},\cdots,x_N\right)\right)}. \tag{34}$$

Proposition 3. *With detection model (11), the following chain rule of the AIR holds,*

$$\begin{aligned} I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right) &= I\left(\boldsymbol{y};\boldsymbol{x}_a\right) + \sum_{n=1}^{N-\nu} I\left(\boldsymbol{y};x_n|\boldsymbol{x}_a\right) \\ &= I(\boldsymbol{y};x_N) + \sum_{n=N-\nu+1}^{N-1} I(\boldsymbol{y};x_n|x_{n+1},x_{n+2},\cdots,x_N) + \sum_{n=1}^{N-\nu} I\left(\boldsymbol{y};x_n|\boldsymbol{x}_a\right). \end{aligned} \tag{37}$$

Proposition 3 reveals an interesting property of the AIR-PM detector, that is, the AIR is the sum of the rate corresponding to signal part $\boldsymbol{x}_a$ with the optimal detection, and the rate corresponding to signal part $\boldsymbol{x}_b$ with LMMSE detection and with no interference from $\boldsymbol{x}_a$. Hence, the AIR with AIR-PM detector for $0<\nu<N-1$ is lower and upper bounded by the AIRs of LMMSE and ML, respectively. Thus, we have the following corollary.

Corollary 2. *For the case $\nu=0$ and $\nu=N-1$, the AIR reads*

$$\begin{aligned} I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right) &= \sum_{n=1}^{N} I\left(\boldsymbol{y};x_n\right), \\ I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right) &= \ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}^{\dagger}\boldsymbol{H}}{N_0}\right), \end{aligned}$$

i.e., the rate with LMMSE detection and channel capacity, respectively.

Note that, the mutual information $I\left(\boldsymbol{y};x_a\right)$ can be written as

$$\begin{aligned} I\left(\boldsymbol{y};x_a\right) &= \ln\det\left(\boldsymbol{I}_a+\boldsymbol{H}_a^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger}+N_0\boldsymbol{I}\right)^{-1}\boldsymbol{H}_a\right) \\ &= \ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}^{\dagger}\boldsymbol{H}}{N_0}\right) - \ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_b^{\dagger}\boldsymbol{H}_b}{N_0}\right). \end{aligned} \tag{38}$$

Similarly, the mutual information $I\left(\boldsymbol{y};x_n|\boldsymbol{x}_a\right)$ equals

$$I\left(\boldsymbol{y};x_n|\boldsymbol{x}_a\right) = \ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_b^{\dagger}\boldsymbol{H}_b}{N_0}\right) - \ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_{b,n}^{\dagger}\boldsymbol{H}_{b,n}}{N_0}\right), \tag{39}$$

where $\boldsymbol{H}_{b,n}$ is obtained by removing the nth column of $\boldsymbol{H}_b$. Utilizing (38), (39) and from Proposition 3, the AIR can also be written in the equivalent form

$$\begin{aligned} I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right) = \ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}^{\dagger}\boldsymbol{H}}{N_0}\right) + (N-\nu-1)\ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_b^{\dagger}\boldsymbol{H}_b}{N_0}\right) \\ - \sum_{n=1}^{N-\nu}\ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_{b,n}^{\dagger}\boldsymbol{H}_{b,n}}{N_0}\right). \end{aligned} \tag{40}$$

Below we give an example of 3×3 complex MIMO channel to illustrate Proposition 2 and 3.

Example 1. *Assume that* $\nu=1$, $N_0=1$, $i=\sqrt{-1}$, *and*

$$\boldsymbol{H}=\begin{bmatrix}1+i & 2+2i & 1+3i\\ 3+3i & 1+3i & 2+2i\\ 2+3i & 2+i & 2+2i\end{bmatrix}.$$

With Proposition 2, the optimal $\boldsymbol{U}$ and $\boldsymbol{G}_{\mathrm{r}}$ are given at the bottom of the next page. The AIR reads $I_{\mathrm{AIR}}=5.7751$ (nats/s/Hz), and $\boldsymbol{H}_{\mathrm{r}}$ is obtained via (19). To verify Proposition 3, we assume that the last column $\boldsymbol{h}_3$ in $\boldsymbol{H}$ corresponds to the single parent layer in $\boldsymbol{x}_a$. Then,

$$\begin{aligned}
I\left(\boldsymbol{y};x_3\right)&=\ln\left(1+\boldsymbol{h}_3^{\dagger}\left(\boldsymbol{h}_1\boldsymbol{h}_1^{\dagger}+\boldsymbol{h}_2\boldsymbol{h}_2^{\dagger}+N_0\boldsymbol{I}\right)^{-1}\boldsymbol{h}_3\right)=1.1887,\\
I\left(\boldsymbol{y};x_2|x_3\right)&=\ln\left(1+\boldsymbol{h}_2^{\dagger}\left(\boldsymbol{h}_1\boldsymbol{h}_1^{\dagger}+N_0\boldsymbol{I}\right)^{-1}\boldsymbol{h}_2\right)=2.1191,\\
I\left(\boldsymbol{y};x_1|x_3\right)&=\ln\left(1+\boldsymbol{h}_1^{\dagger}\left(\boldsymbol{h}_2\boldsymbol{h}_2^{\dagger}+N_0\boldsymbol{I}\right)^{-1}\boldsymbol{h}_1\right)=2.4674,
\end{aligned}$$

which shows that the chain rule for the AIR holds since

$$I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right)=I\left(\boldsymbol{y};x_3\right)+\sum_{n=1}^{2}I\left(\boldsymbol{y};x_n|x_3\right).$$

4.2 Maximizing the Ergodic AIR

With the chain rule for the AIR stated, we next analyze the selections of $\boldsymbol{x}_a$ in connection with the ergodic AIR. We consider a Rayleigh channel with Kronecker correlation model $\mathrm{vec}\left(\boldsymbol{H}\right)\sim\mathcal{CN}\left(\boldsymbol{0},\frac{1}{N}\left(\boldsymbol{R}_{\mathrm{t}}\otimes\boldsymbol{R}_{\mathrm{r}}\right)\right)$, where $\boldsymbol{R}_{\mathrm{t}}$ and $\boldsymbol{R}_{\mathrm{r}}$ represent the transmit and receive

$$\boldsymbol{U}=\begin{bmatrix}3.4339 & 0 & 1.4803-1.0921i\\ 0 & 2.8851 & 1.5292+1.2233i\\ 0 & 0 & 1.8118\end{bmatrix},$$

$$\boldsymbol{G}_{\mathrm{r}}=\begin{bmatrix}10.7917 & 0 & 5.0833-3.7500i\\ 0 & 7.3235 & 4.4118+3.5294i\\ 5.0833+3.7500i & 4.4118-3.5294i & 9.5016\end{bmatrix}.$$

correlations, respectively. Assume decompositions $\boldsymbol{R}_{\mathrm{r}} = \boldsymbol{U}_{\mathrm{r}}\boldsymbol{\Sigma}\boldsymbol{U}_{\mathrm{r}}^{\dagger}$ and $\boldsymbol{R}_{\mathrm{t}} = \boldsymbol{U}_{\mathrm{t}}\boldsymbol{\Delta}\boldsymbol{U}_{\mathrm{t}}^{\dagger}$, where $\boldsymbol{U}_{\mathrm{t}}$, $\boldsymbol{U}_{\mathrm{r}}$ are unitary, and $\boldsymbol{\Sigma}$, $\boldsymbol{\Delta}$ are diagonal with the nth diagonal element denoted as σ_n and δ_n, respectively. The correlated Rayleigh channel can then be modeled as

$$\boldsymbol{H} = \boldsymbol{U}_{\mathrm{r}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{i.i.d.}}\boldsymbol{\Delta}^{1/2}\boldsymbol{U}_{\mathrm{t}}^{\dagger}, \tag{41}$$

with $\mathrm{vec}\left(\boldsymbol{H}_{\mathrm{i.i.d.}}\right) \sim \mathcal{CN}\left(0, \frac{1}{N}\boldsymbol{I}\right)$. Inserting (41) back into signal model (1), we have

$$\boldsymbol{y} = \boldsymbol{U}_{\mathrm{r}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{i.i.d.}}\boldsymbol{\Delta}^{1/2}\boldsymbol{U}_{\mathrm{t}}^{\dagger}\boldsymbol{x} + \boldsymbol{e},$$

or equivalently,

$$\boldsymbol{U}_{\mathrm{r}}^{\dagger}\boldsymbol{y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{i.i.d.}}\boldsymbol{\Delta}^{1/2}\boldsymbol{U}_{\mathrm{t}}^{\dagger}\boldsymbol{x} + \boldsymbol{U}_{\mathrm{r}}^{\dagger}\boldsymbol{e}. \tag{42}$$

As $\boldsymbol{y}$, $\boldsymbol{x}$ and $\boldsymbol{e}$ comprise i.i.d. complex Gaussian entries, multiplying them with a unitary matrix will not change their statistical properties. Therefore, the AIR of signal model (42) is equivalent to the model

$$\boldsymbol{y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{i.i.d.}}\boldsymbol{\Delta}^{1/2}\boldsymbol{x} + \boldsymbol{e}. \tag{43}$$

Hence, for the ergodic AIR analysis in correlated Rayleigh fading channels, it is sufficient to consider the channel model

$$\boldsymbol{H} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{i.i.d.}}\boldsymbol{\Delta}^{1/2}. \tag{44}$$

Proposition 4. *Consider $\boldsymbol{H}$ modeled in (44), and assume that, the diagonal elements δ_n are positive and sorted in ascending order,*

$$0 < \delta_1 \leq \delta_2 \leq \cdots \leq \delta_N.$$

Then, selecting the last ν layers for $\boldsymbol{x}_a$ will maximize the ergodic AIR

$$I_{\mathrm{AIR}}^{\mathrm{e}}\left(\boldsymbol{y};\boldsymbol{x}\right) = \mathbb{E}_{\boldsymbol{H}}\left[I_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right)\right]. \tag{45}$$

The proof is in Appendix D. Proposition 4 shows that, with AIR-PM detector, the layer with higher transmit power shall be chosen as parent layer, regardless of the receive correlation. This is somewhat contradictory to the PM detector, where layers with lower SNRs are chosen as parent layers [17]. Note that, here the selection is on a long-term, and not an instantaneous, basis. For a given channel realization $\boldsymbol{H}$, the selection of $\boldsymbol{x}_a$ is to maximize the AIR (28), and the correlation matrices are irrelevant of such a selection.

Next, we analyze the asymptotic property of the ergodic AIR, and consider the case $N = K$. Denote $\mathcal{L}\left(\boldsymbol{H}, N_0\right)$ as the ergodic capacity

$$\mathcal{L}\left(\boldsymbol{H}, N_0\right) = \mathbb{E}_{\boldsymbol{H}}\left[\ln\det\left(\boldsymbol{I} + \frac{\boldsymbol{H}^{\dagger}\boldsymbol{H}}{N_0}\right)\right].$$

Then, from (40) the ergodic AIR in (45) can be calculated as

$$I^{\mathrm{e}}_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right) = \mathcal{L}\left(\boldsymbol{H},N_0\right) + (N-\nu-1)\mathcal{L}\left(\boldsymbol{H}_b,N_0\right) - \sum_{n=1}^{N-\nu}\mathcal{L}\left(\boldsymbol{H}_{b,n},N_0\right), \quad (46)$$

which shows the relationship between the ergodic AIR and the ergodic capacity. For general cases, the ergodic AIR analysis can also be based on (46). However, the discussions of the ergodic capacity are fairly long, see e.g., [33, 34], therefore, we do not discuss them here.

In the low SNR regime, it is known that LMMSE detection is close to optimal, hence, the AIR-PM detector will have similar performance as LMMSE, and $I^{\mathrm{e}}_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right)$ will not depend much on the selection of $\boldsymbol{x}_a$. On the other hand, in the high SNR regime, by utilizing high SNR expansion [34]

$$\mathcal{L}\left(\boldsymbol{H},N_0\right) \approx N\big(-\ln N_0 - \mathcal{L}^{\infty}_{\boldsymbol{H}}\left(\boldsymbol{H},N_0\right)\big), \quad (47)$$

where $\mathcal{L}^{\infty}_{\boldsymbol{H}}\left(\boldsymbol{H},N_0\right)$ is the high SNR power-offset, we can also show that, $I^{\mathrm{e}}_{\mathrm{AIR}}\left(\boldsymbol{y};\boldsymbol{x}\right)$ does not depend on the selection of $\boldsymbol{x}_a$. These results are summarized in Proposition 5, and the proof is in Appendix E.

Proposition 5. *With signal model (43) and $N = K$, as N_0 goes to infinity or 0, the differences of the ergodic AIR with different selections of $\boldsymbol{x}_a$ asymptotically equal zero.*

In the next section, we will discuss some useful extensions of the proposed AIR-PM detector.

5 Extensions

5.1 Parallel Detection with Multiple Branches

The AIR-PM detector can be extended to have multiple detection branches in parallel, where each individual detection branch is represented by the same detection model (11), but the parent layers are different from each other. All detection branches share the same constraint on $\boldsymbol{G}_{\mathrm{r}}$, and as more layers are selected as parent layers, the LLR quality is boosted. Next, we show that, the overheads introduced by the pre-processing of $\boldsymbol{H}$ is shared for all branches, including the matrix inverse operation to calculate $\boldsymbol{W}$ which is the most computationally demanding part of the pre-processing.

Consider a detection branch with $\boldsymbol{x}_a$ comprising ν_ℓ arbitrary layers from $\boldsymbol{x}$. Define a permutation matrix $\boldsymbol{P}_\ell$ that permutes the ν_ℓ columns in $\boldsymbol{H}$ corresponding to layers in $\boldsymbol{x}_a$ to be the last ν_ℓ columns, regardless of the order of the other columns, and denote the channel

after permutation as $\boldsymbol{H}_\ell = \boldsymbol{H}\boldsymbol{P}_\ell$. From (20) and (21), it holds that, $\boldsymbol{W}_\ell = \boldsymbol{P}_\ell^{\mathrm{T}}\boldsymbol{W}$ and $\boldsymbol{B}_\ell = \boldsymbol{P}_\ell^{\mathrm{T}}\boldsymbol{B}\boldsymbol{P}_\ell$. That is, $\boldsymbol{W}_\ell$ and $\boldsymbol{B}_\ell$ are just permutations of $\boldsymbol{W}$ and $\boldsymbol{B}$ corresponding to the original $\boldsymbol{H}$, respectively. Then, based on $\boldsymbol{B}_\ell$, $\boldsymbol{G}_{\mathrm{r}}$ is obtained through Proposition 2, and $\boldsymbol{H}_{\mathrm{r}}$ is calculated in (19).

For the sake of complexity saving, in general we assume that each layer can appear as parent layer at most once among all detection branches. The LLR calculations of parent layers then use (33). For a layer x_n that never appears as a parent layer in any detection branch, we choose the branch that has the largest $I\left(\boldsymbol{y}; x_n|\boldsymbol{x}_a\right)$ for LLR calculation of x_n based on (34).

5.2 Detection with A Priori Information

When there is *a priori* information $\tilde{L}_n^m$ available, the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ in (10) changes to

$$\mu(\boldsymbol{y}|\boldsymbol{x}) = 2\mathcal{R}\{\boldsymbol{x}^\dagger\tilde{\boldsymbol{y}}\} - \boldsymbol{x}^\dagger\boldsymbol{G}_{\mathrm{r}}\boldsymbol{x} + \ln p(\boldsymbol{x}),$$

where $\ln p(x_n)$ is calculated based on updated $\tilde{L}_n^m$ as in (5). The priori information $\tilde{L}_n^m$ is updated in each iteration and the term $\ln p(\boldsymbol{x})$ is changed accordingly. To cooperate the priori information, by replacing the metrics $\mu_1^n\left(\tilde{y}_n|x_n, \boldsymbol{x}_a\right)$ and $\mu_2^n\left(\tilde{y}_n|x_n, x_{n+1}, \cdots, x_N\right)$ in (14) and (15) with

$$\mu_1^n\left(\tilde{y}_n|x_n, \boldsymbol{x}_a\right) \leftarrow \mu_1^n\left(\tilde{y}_n|x_n, \boldsymbol{x}_a\right) + \ln p(x_n), \tag{48}$$

$$\mu_2^n\left(\tilde{y}_n|x_n, x_{n+1}, \cdots, x_N\right) \leftarrow \mu_2^n\left(\tilde{y}_n|x_n, x_{n+1}, \cdots, x_N\right) + \ln p(x_n), \tag{49}$$

respectively, Proposition 2 still holds and the detection process follows Sec. III. The LS estimate of $\boldsymbol{x}_b$ is now obtained through

$$\hat{\mu}_1^n\left(\tilde{y}_n|\boldsymbol{x}_a\right) = \max_{x_n}\left(\mu_n\left(\tilde{y}_n|x_n, \boldsymbol{x}_a\right) + \ln p(x_n)\right), \tag{50}$$

which is nonlinear in x_n and a search over $\mathcal{X}$ is needed. However, for the cases that x_n can be written in a linear function[2] of bit vector $\boldsymbol{s}_n = (x_n^1, x_n^2, \cdots, x_n^M)$ such that

$$x_n = \boldsymbol{s}_n\boldsymbol{\gamma},$$

the metric $\hat{\mu}_1^n\left(y_n|\boldsymbol{x}_a\right)$ in (50) equals the maximum (over variable $\boldsymbol{s}_n$) of

$$\mathcal{J} = 2\mathcal{R}\left\{\left(\tilde{y}_n - \sum_{k=N-\nu+1}^{N} g_{n,k}x_N\right)(\boldsymbol{s}_n\boldsymbol{\gamma})^*\right\} - g_{n,n}|\boldsymbol{s}_n\boldsymbol{\gamma}|^2 + \sum_{m=1}^{M}\left(\frac{(1+x_n^m)\tilde{L}_n^m}{2} - \ln\left(1+\exp\left(\tilde{L}_n^m\right)\right)\right).$$

[2]Such a property usually holds, for example, for a 4-PAM modulation with [-1,-1], [-1,1], [1,-1], [1,1] mapped to $\frac{-3}{\sqrt{10}}$, $\frac{-1}{\sqrt{10}}$, $\frac{1}{\sqrt{10}}$, and $\frac{3}{\sqrt{10}}$, respectively, we have $\boldsymbol{\gamma} = \frac{1}{\sqrt{10}}(2,1)^{\mathrm{T}}$.

Since $\mathcal{J}$ is quadratic in $\boldsymbol{s}_n$, the maximization can be solved via LS.

The PM detector with soft inputs has also been discussed in [18], where a transform of the detection model with soft information, into an extended model that shares the same form of (8) is made. However, as $\tilde{L}_n^m$ changes over iterations, the QR-decomposition according to the extended model is required at each iteration. With AIR-PM detector, only a slight modification of the metrics is needed as in (48) and (49).

5.3 Detection with Imperfect Channel Estimate

In practical systems, the channel estimate is imperfect, and it is beneficial for the detector to take the estimation error into account. We model the imperfect channel $\mathrm{vec}(\tilde{\boldsymbol{H}}) = \mathrm{vec}(\boldsymbol{H}) + \boldsymbol{e}_h$, with $\mathrm{vec}(\boldsymbol{H}) \sim \mathcal{CN}\left(0, \frac{1}{N}\boldsymbol{I}\right)$ and the error vector $\boldsymbol{e}_h \sim \mathcal{CN}(0, \beta\boldsymbol{I})$. Following a similar discussion in [17], to cope with channel estimation error, we can modify model (2) to

$$p(\boldsymbol{y}|\tilde{\boldsymbol{H}}, \boldsymbol{x}) = \frac{1}{\left(\pi\tilde{N}_0\right)^N} \exp\left(-\frac{1}{\tilde{N}_0}\|\boldsymbol{y} - \frac{1}{1+\beta N}\tilde{\boldsymbol{H}}\boldsymbol{x}\|^2\right), \tag{51}$$

where the updated noise power

$$\tilde{N}_0 = \frac{1}{N}\mathrm{Tr}\left(\left(\boldsymbol{y} - \frac{1}{1+\beta N}\tilde{\boldsymbol{H}}\boldsymbol{x}\right)\left(\boldsymbol{y} - \frac{1}{1+\beta N}\tilde{\boldsymbol{H}}\boldsymbol{x}\right)^\dagger\right) = \frac{\beta N}{1+\beta N} + N_0.$$

The design and process of AIR-PM detector for (51) are the same as for (2), but the performance is decreased as the SNR is decreased from $1/N_0$ to $1/\left((1+\beta N)^2\tilde{N}_0\right)$.

5.4 AIR Computation with Finite Constellation

To obtain a closed-form expression of the AIR for finite constellations is difficult [22]. Therefore, the parameter optimization over $\boldsymbol{H}_\mathrm{r}$ and $\boldsymbol{G}_\mathrm{r}$ are based on the assumption that the input vector $\boldsymbol{x}$ is Gaussian distributed. In order to evaluate the AIR for finite constellations with AIR-PM, we use Monte Carlo simulations. With AIR-PM, the complexity of the Monte Carlo simulations are also greatly simplified. With $\tilde{p}(\boldsymbol{y}|\boldsymbol{x})$ in (11), the AIR in (18) can be evaluated based on

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\ln\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right] &= \frac{1}{|\mathcal{X}|^N}\mathbb{E}_{\boldsymbol{y}}\left[\sum_{\boldsymbol{x}}\ln\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right],\\ \mathbb{E}_{\boldsymbol{y}}\left[\ln\tilde{p}(\boldsymbol{y})\right] &= \mathbb{E}_{\boldsymbol{y}}\left[\ln\left(\sum_{\boldsymbol{x}}\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right)\right] - N\ln|\mathcal{X}|.\end{aligned}$$

Utilizing property (13), the terms inside the expectation operators can be decomposed as in (52) and (53), respectively, where the number of summations is reduced to $\mathcal{O}(|\mathcal{X}|^{\nu+1})$.

Later we will show that, although the optimal $\boldsymbol{H}_{\mathrm{r}}$ and $\boldsymbol{G}_{\mathrm{r}}$ are optimized for the Gaussian constellation, such a design of the AIR-PM detector also work well for finite constellations.

6 Receiver Structure and Complexity Analysis

6.1 Receiver Structure with the AIR-PM Detector

In Fig. 2 we depicted the receiver structure of the AIR-PM detector. The $\boldsymbol{G}_{\mathrm{r}}$ is constructed based on Proposition 2, and the metric calculation $\mu(\boldsymbol{y}|\boldsymbol{x})$ follows Proposition 1. The LLRs for $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are calculated based on (33) and (34), respectively. Next we analyze the complexity of the proposed AIR-PM detector and compare it with the other MIMO detectors.

6.2 Selection the Best Parent Layers

As the target of AIR-PM is to maximize AIR, the optimal selection of $\boldsymbol{x}_a$ shall maximize the AIR in (28). As $I_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x})$ can be rewritten in the equivalent form (40) and only depends on $\boldsymbol{H}$ and $\boldsymbol{H}_b$, changing the orders of the layers in $\boldsymbol{x}_a$, i.e., permuting the columns of $\boldsymbol{H}_a$ will no impact the AIR. On the other hand, permuting the columns of $\boldsymbol{H}_b$ and $\boldsymbol{H}_{b,n}$ will not change the determinants $\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_b^{\dagger}\boldsymbol{H}_b}{N_0}\right)$ and $\det\left(\boldsymbol{I}+\frac{\boldsymbol{H}_{b,n}^{\dagger}\boldsymbol{H}_{b,n}}{N_0}\right)$, respectively.

$$\begin{aligned}\sum_{\boldsymbol{x}}\ln\left(\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right)&=\sum_{\boldsymbol{x}}\left(\sum_{n=1}^{N-\nu}\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)+\sum_{n=N-\nu+1}^{N}\mu_2^n\left(\tilde{y}_n|x_n,x_{n+1},\cdots,x_N\right)\right)\\&=|\mathcal{X}|^{(N-\nu)}\sum_{\boldsymbol{x}_a}\left(\sum_{n=N-\nu+1}^{N}\mu_2^n\left(\tilde{y}_n|x_n,x_{n+1},\cdots,x_N\right)+\frac{1}{|\mathcal{X}|}\sum_{n=1}^{N-\nu}\sum_{x_n}\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)\right).\end{aligned}\tag{52}$$

$$\begin{aligned}\sum_{\boldsymbol{x}}\tilde{p}(\boldsymbol{y}|\boldsymbol{x})&=\sum_{\boldsymbol{x}}\left(\prod_{n=1}^{N-\nu}\exp\left(\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)\right)\cdot\prod_{n=N-\nu+1}^{N}\exp\left(\mu_2^n\left(\tilde{y}_n|x_n,x_{n+1},\cdots,x_N\right)\right)\right)\\&=\sum_{\boldsymbol{x}_a}\left(\prod_{n=N-\nu+1}^{N}\exp\left(\mu_2^n\left(\tilde{y}_n|x_n,x_{n+1},\cdots,x_N\right)\right)\cdot\prod_{n=1}^{N-\nu}\left(\sum_{x_n}\exp\left(\mu_1^n\left(\tilde{y}_n|x_n,\boldsymbol{x}_a\right)\right)\right)\right).\end{aligned}\tag{53}$$

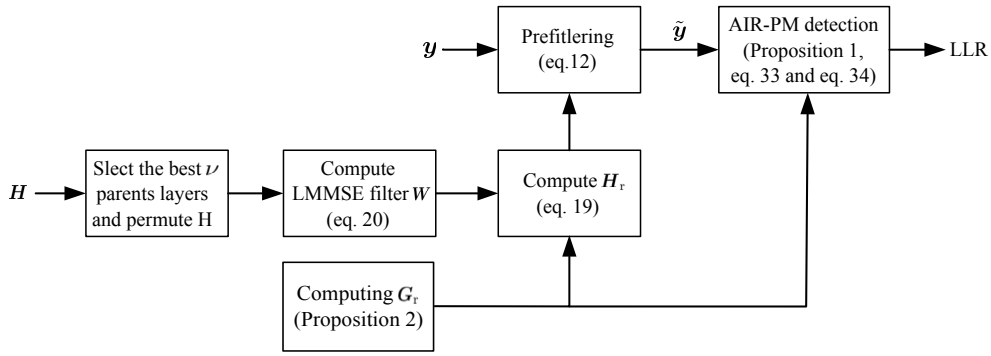


Figure 2: The receiver structure of the AIR-PM detector. With multiple branches, each detection branch has the same structure and the pre-processing can be shared as introduced in Sec. V-A.

Therefore, the ordering of the layers inside $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ will not change the AIR with the AIR-PM detector. Hence, unlike the PM detector, re-ordering of $\boldsymbol{x}$ is not needed in AIR-PM detector, and the complexity of such selection is less compared to finding the optimal ordering of $\boldsymbol{x}$. With $\nu = 1$, the task is just to select one single best layer for $\boldsymbol{x}_a$.

6.3 Detection Complexity

With AIR-PM detector, the pre-processing needs to compute $\boldsymbol{H}_{\mathrm{T}}$ which requires a matrix inversion of size $\min(K, N)$, and prefilter the data according to (12). However, with the other detections except for the ML detection, decompositions (QR, Cholesky and WL) of $\boldsymbol{H}$ are also needed to convert $\boldsymbol{H}$ into upper-triangular form followed by prefiltering the received signal. The remaining pre-processing of the AIR-PM detector is to construct $\boldsymbol{G}_{\mathrm{T}}$ based on Proposition 2, whose complexity is neglectable for a small ν. Therefore, the complexity of pre-processes among different detectors are similar.

For a large constellation size $|\mathcal{X}|$, the detection complexity is much more heavy compared to the complexity of the pre-processing. In Table I, we analyze the detection complexity of AIR-PM detector, and compare it to the state-of-the-art detectors. The detection complexity is measured in terms of the number of paths that have been visited in the search process, and the number of complex multiplications to calculate the metric $\mu(\boldsymbol{y}|\boldsymbol{x})$ of each path (for fair comparisons, the metric calculations are simplified through possible channel decompositions). We also list the look-up-table (LUT) operation for exponential and logarithm operations. With PM and AIR-PM detectors, the number of evaluated paths are calculated for $\boldsymbol{x}_a$. However, the number of complex multiplications for each path includes the computations of all the counter-hypotheses of $\boldsymbol{x}_b$ based on the ZF-DF and LS estimates, respectively. As can be seen in the table, the AIR-PM detector has lower complexity than the other detectors, yet with a fixed complexity and fully parallelizable structure for all layers.

Table 1: Complexity-Performance Comparison of the MIMO Detectors.

Detector	Parameter	Number of evaluated paths	Complex multiplications per path	LUT per bit	Channel decomposition	Fixed complexity?	Fully parallelizable?
exact ML	n.a.	$\lvert\mathcal{X}\rvert^N$	$N(N+3)/2$	$\frac{1}{NM}\lvert\mathcal{X}\rvert^N+2$	n.a.	yes	yes
SD[10]	n.a.	$\lvert\mathcal{X}\rvert^{\alpha N}$ $(0<\alpha<1)$	$N(N+3)/2$	$\frac{1}{NM}\lvert\mathcal{X}\rvert^{\alpha N}+2$	$\boldsymbol{L}^\dagger\boldsymbol{L}=\boldsymbol{H}^\dagger\boldsymbol{H}$	no	no
K-best[14]	K	$\min(K^2, K\lvert\mathcal{X}\rvert)$	$N(N+3)/2$	$\frac{K}{NM}+2$	$\boldsymbol{H}=\boldsymbol{Q}\boldsymbol{R}$	yes	no
PM[17]	ν	$\lvert\mathcal{X}\rvert^\nu$	$\nu(\nu+3)/2$ $+\nu(N-\nu)\left(\frac{(N-\nu+1)M}{2}+1\right)$	$\frac{1}{NM}\lvert\mathcal{X}\rvert^\nu+2$	$\boldsymbol{H}=\boldsymbol{Q}\boldsymbol{R}$	yes	no
WLD [29]	E	$\frac{N}{E}\lvert\mathcal{X}\rvert^E$	$E(E+3)/2$	n.a.	$\boldsymbol{H}=\boldsymbol{W}\boldsymbol{L}$	yes	no
AIR-PM	ν	$\lvert\mathcal{X}\rvert^\nu$	$\nu(\nu+3)/2+(N-\nu)(M+1)$	$\frac{1}{NM}\lvert\mathcal{X}\rvert^\nu+2$	$\boldsymbol{H}_\mathrm{r}=\boldsymbol{W}^\dagger(\boldsymbol{I}+\boldsymbol{G}_\mathrm{r})$	yes	yes

7 Empirical performance evaluation

In this section, we provide numerical results with slow-fading Rayleigh MIMO channels, and each channel realization spreads across the entire transmit data block. The Rayleigh channel $\boldsymbol{H}$ is modeled according to (44) and is perfectly known, since the process of dealing with the imperfect channel estimate is the same and the latter one only suffers from an SNR loss. In all tests, except for the ergodic AIR evaluation, $\boldsymbol{H}$ has no correlations, i.e., $\boldsymbol{\Sigma}=\boldsymbol{\Delta}=\boldsymbol{I}$.

The FER results are evaluated with a rate-$1/2$, (1032, 2064) turbo code, which is from LTE [2] standard and the decoder uses 8 internal iterations. The simulations are executed until 1000 frame errors are encountered. In addition, as the complex-valued model (1) can be rewritten as

$$\boldsymbol{y}=\begin{bmatrix}\mathcal{R}\{\boldsymbol{y}\}\\ \mathcal{I}\{\boldsymbol{y}\}\end{bmatrix},\quad \boldsymbol{x}=\begin{bmatrix}\mathcal{R}\{\boldsymbol{x}\}\\ \mathcal{I}\{\boldsymbol{x}\}\end{bmatrix},\quad \boldsymbol{n}=\begin{bmatrix}\mathcal{R}\{\boldsymbol{n}\}\\ \mathcal{I}\{\boldsymbol{n}\}\end{bmatrix}, \tag{54}$$

and

$$\boldsymbol{H}=\begin{bmatrix}\mathcal{R}\{\boldsymbol{H}\} & -\mathcal{I}\{\boldsymbol{H}\}\\ \mathcal{I}\{\boldsymbol{H}\} & \mathcal{R}\{\boldsymbol{H}\}\end{bmatrix}, \tag{55}$$

we use the real-valued signal model (54) and (55) for detection in all FER simulations, which has a higher degrees of freedom than the complex-valued model.

7.1 AIR with Optimal Selection

In Fig. 3, we consider the instantaneous selection of $\boldsymbol{x}_a$, and compare the AIR with the optimal selection with the average AIR over all possible selections. We test 4×4 complex MIMO, and as expected, the AIR with optimal selection outperforms the average AIR. For $\nu=1$, the SNR gain is more than 0.5 dB. The channel capacity and the rate of LMMSE detector (which is equivalent to AIR-PM with $\nu=0$) are also presented. As can be seen, the AIR-PM detector with $\nu=1$ significantly improves the AIR over LMMSE detector, and the AIR-PM detector with $\nu=2$ is quite close to the channel capacity (which is equivalent to AIR-PM with $\nu=3$).

7.2 AIR with Finite Constellation

In Fig. 4, we evaluate the AIR for finite constellations with AIR-PM detector via Monte Carlo simulation. We test 4×4 complex MIMO and set $\nu=1$, that is, we have 4 different choices of $\boldsymbol{x}_a$. The AIR with the optimal selection of $\boldsymbol{x}_a$ is compared to the averaged AIR

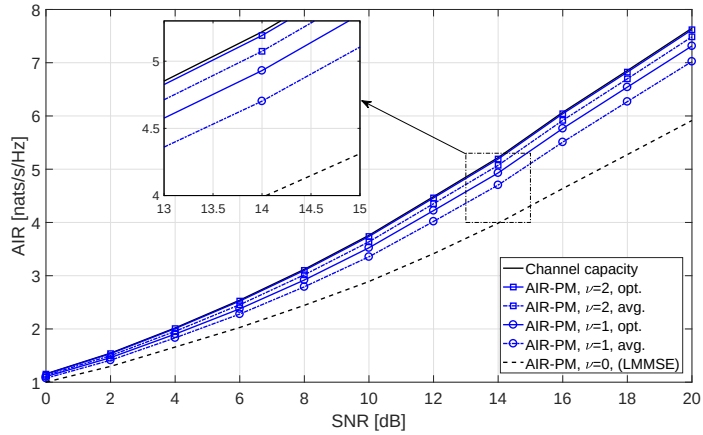


Figure 3: The AIR of AIR-PM detector with $\nu = 0, 1$, and 2 under 4×4 complex MIMO channels. The AIR with the optimal selection of $\boldsymbol{x}_a$ that maximizes (28) is compared to the averaged AIR of all possible selections.

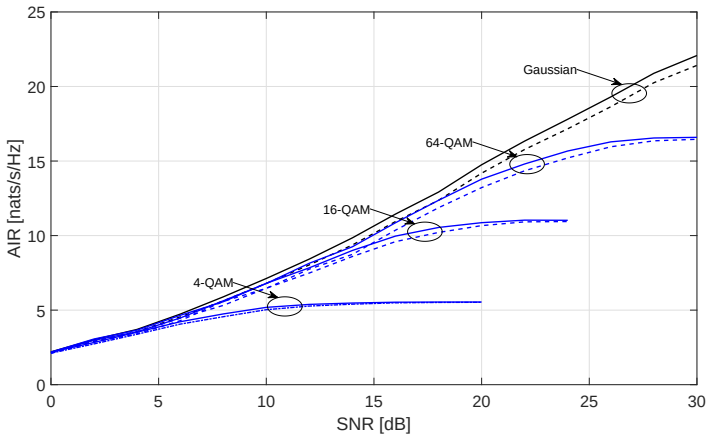


Figure 4: The Monte Carlo simulation of the AIR with 4×4 complex MIMO channels of the AIR-PM detector. The solid lines are the AIR with the optimal selection of $\boldsymbol{x}_a$, while the dashed lines are the averaged AIR of all possible selections. The finite constellations 4-QAM, 16-QAM and 64-QAM reuse the optimal index of parent layer chosen from the Gaussian inputs.

of all 4 possible choices. For finite constellations, not only the parameters $\boldsymbol{H}_{\mathrm{r}}$, $\boldsymbol{G}_{\mathrm{r}}$ are designed based on Gaussian inputs, the optimal selections of $\boldsymbol{x}_a$ also reuse the schemes for Gaussian inputs. The results in Fig. 4 show that, although the AIR-PM detection model (11) is designed for Gaussian inputs, it works well for finite constellations.

7.3 Ergodic AIR with Correlated Channel

In both Fig. 5 and Fig. 6, we simulate the ergodic AIR with AIR-PM detector under 8×8 complex MIMO and $\nu = 1$. We alternatively select $\boldsymbol{x}_a$ to be one of the 8 different layers,

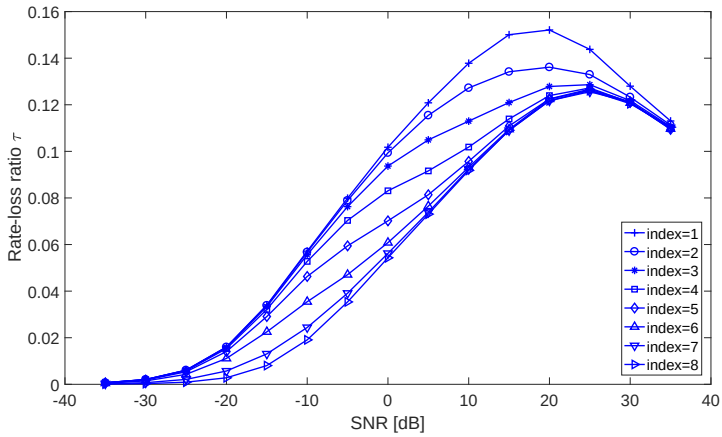


Figure 5: The ergodic AIR of AIR-PM detector with $\nu = 1$ under 8×8 complex MIMO channels, with only transmit correlation. The transmit powers are in ascending order from the first layer to the last layer.

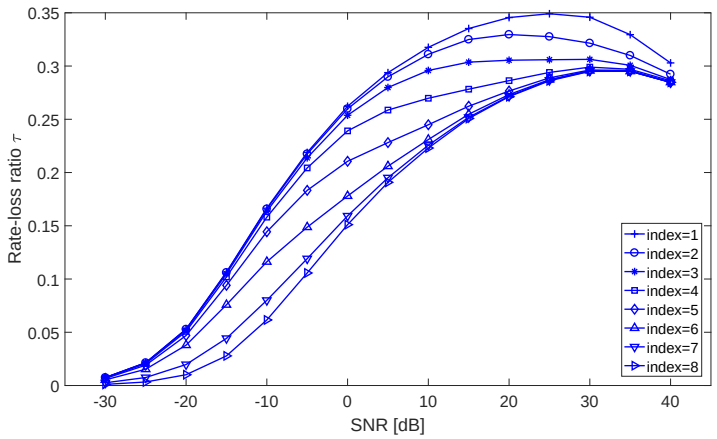


Figure 6: Repeat the test in Fig. 5, but with the equal transmit and receive correlations.

and measure the rate-loss ratio τ of $I^{\mathrm{e}}_{\mathrm{AIR}}$ compared to ergodic capacity $\mathcal{L}(\boldsymbol{H}, N_0)$,

$$\tau = 1 - I^{\mathrm{e}}_{\mathrm{AIR}}(\boldsymbol{y}; \boldsymbol{x}) / \mathcal{L}(\boldsymbol{H}, N_0).$$

In Fig. 5, we test with transmit correlation only, and set $\delta_n = \exp(n) / \sum_{n=1}^{8} \exp(n)$. As can be seen, as δ_n decreases (from index 8 to 1), the rate losses become larger. Furthermore, at very low and high SNRs, the rate losses of different selections converge. These results show that layers corresponding to stronger channels should be chosen as parent layers, and that in the low and high SNR regimes, $I^{\mathrm{e}}_{\mathrm{AIR}}(\boldsymbol{y}; \boldsymbol{x})$ are asymptotically equal for different selections of $\boldsymbol{x}_a$. These observations are well aligned with Propositions 4 and 5.

In Fig. 6, we repeat the test in Fig. 5, but with both transmit and receive correlations, and we set $\sigma_n = \delta_n$. The conclusions are the same as drawn from Fig. 5, and due to correlations at both sides, the rate losses are larger than those in Fig. 5.

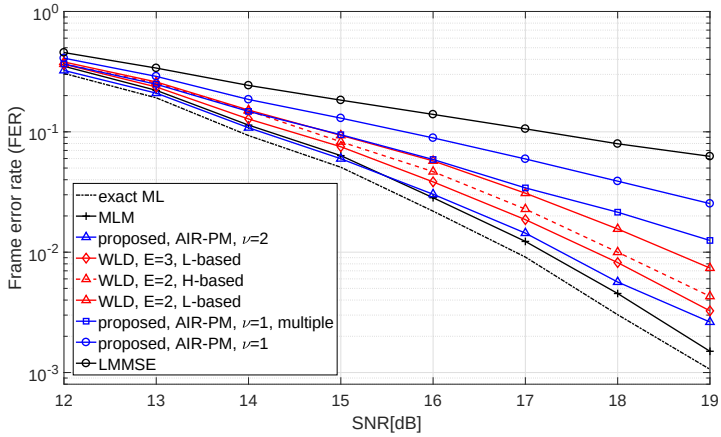


Figure 7: The FER performance of 6×6 real (3×3 complex) MIMO channels with 4-PAM (16-QAM) modulation.

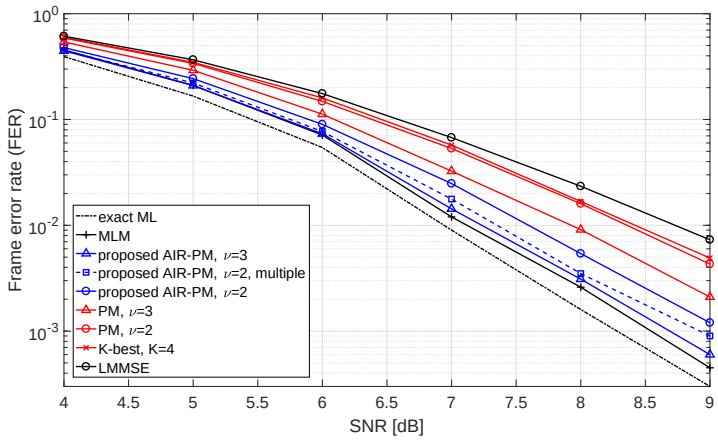


Figure 8: The FER performance of 8×8 real (4×4 complex) MIMO channels with 2-PAM (4-QAM) modulation.

7.4 Frame Error Rate with Turbo Code

Next, we show FER performance. With the real-valued model, the number of transmit layers is $2N$ and the complex QAM symbols are split to real PAM symbols. The LLR calculation of exact ML and MLM are based on (6)-(7), while the PM detector is based on (8)-(9), and the AIR-PM detector is based on (33)-(34). The K-best detector is from [14], and when there is counter-hypothesis missing, we use the difference between the best and worst metrics among all the paths to approximate the LLR. The WLD detector follows [29, 31]. We also simulate the AIR-PM detector with N parallel detection branches, with $\boldsymbol{x}_a$ in each branch containing $\nu=2$ non-intersecting layers, and the selections of $\boldsymbol{x}_a$ is to maximize the sum of the AIRs for all branches. We only calculate the LLRs for $\boldsymbol{x}_a$ in each branch, and then combine them from all branches.

In Fig. 7, we compared the FER performance between the proposed AIR-PM detector and

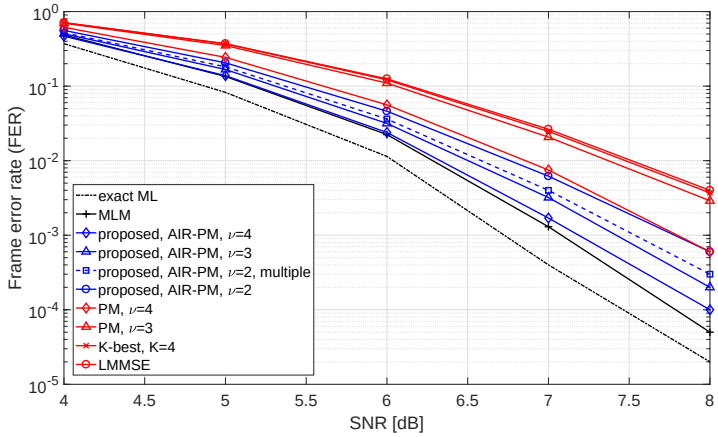


Figure 9: The FER performance of with 12×12 real (6×6 complex) MIMO channels with 2-PAM (4-QAM) modulation.

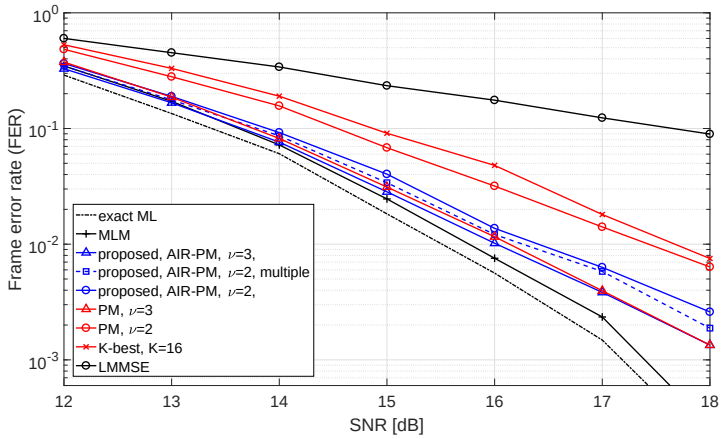


Figure 10: The FER performance of 8×8 real (4×4 complex) MIMO channels with 4-PAM (16-QAM) modulation.

the WLD detector under 6×6 real (3×3 complex) MIMO channels with 16-QAM modulation. Note that, the WLD detector utilizes N/E parallel detection branches. As can be seen that, the WLD detector with $E=2$ (both for L-based and H-based metric computations) is inferior to the AIR-PM detector with $\nu=2$, which utilizes a single detection branch and performs fairly close to the MLM detector.

In Fig. 8 and Fig. 9, we show the FER under 8×8 real (4×4 complex) and 12×12 real (6×6 complex) MIMO channels with 2-PAM (4-QAM for complex symbols) modulation, respectively. As can be seen, the proposed AIR-PM detector outperforms the other detectors. With the same setting of ν, the AIR-PM detector is 0.5 dB better than the PM detector in terms of SNR. Further, the AIR-PM detector with $\nu=3$ is close to MLM. The SNR gains of the AIR-PM detector over the PM detector becomes larger in Fig. 9 than in Fig. 8. This is because that, the quality of ZF-DF estimates degrades, due to error propagation when N increases.

In Fig. 10, we repeat the tests in Fig. 8 with 4-PAM (16-QAM for complex symbols) modulation. As can been seen, the FER performance of the AIR-PM detector with $\nu = 2$ outperforms the PM detector around 1dB, due to a higher modulation scheme is used. With $\nu = 3$, both the AIR-PM and PM detectors perform close to the MLM, while the AIR-PM detector has less computational cost than the PM detector.

7.5 Cooperating with A Priori Information

In Fig. 11, we test the EXIT chart [36, 37] of the proposed AIR-PM detector and compare it with the optimal MAP and the linear LMMSE with parallel interference cancellation (LMMSE-PIC) [26, 27] detectors under 6×6 complex MIMO channels with 4-QAM modulation. The AIR-PM detector uses the soft inputs as described in Sec. V-B. As can be seen, with the input mutual information I_{A} increases, the output extrinsic I_{E} measured with the LLRs generated from the detectors also increases, and the AIR-PM detector significantly outperforms the LMMSE-PIC detector when I_{A} is small. As I_{A} increases, the LMMSE-PIC becomes superior to the AIR-PM and is close to the MAP, due to the fact that, the priori information based interference cancellation becomes perfect in the LMMSE-PIC and there is no interference cancellation utilized in AIR-PM. The AIR-PM detector can also be modified to cooperate with interference cancellation and will outperform the LMMSE-PIC [26], however, such an approach needs to introduce an interference cancellation term in (10) and a gradient-based optimization of the parameters is needed. Nevertheless, as it can be seen that, with $\nu = 2$ the AIR-PM detector is already superior to LMMSE-PIC detector for input I_{A} less than 0.75, and with $\nu = 3$ the AIR-PM performs close to MAP.

In Fig. 12, we test the AIR-PM detector in an iterative detection and decoding scheme with $\nu = 3$ under 6×6 real (3×3 complex) MIMO channels with 4-PAM (16-QAM) modulation. With up to 3 global iterations (including the initial iteration), more than 1 dB SNR gain can be observed at 1% FER both for MAP and the proposed AIR-PM detector. Moreover, as can be seen, the SNR gain of the AIR-PM detector over iterations is similar to the MAP detector. That is, with the pre-processing in AIR-PM detector, the iteration gain is not reduced compare to the generic MIMO detector without pre-processing.

8 Summary

In this paper, we considered soft-output detector design for MIMO channels. We proposed a detector with partial marginalization based on maximizing the achievable information rate. The parameters of the AIR-PM detector are optimized in closed forms via AIR-maximizing for a mismatched detection model. Since the AIR is maximized, the least

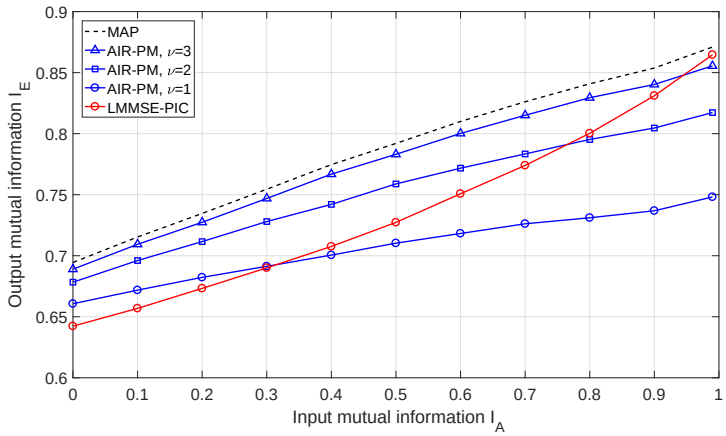


Figure 11: EXIT chart of the proposed AIR-PM detector under 6×6 complex MIMO channels with 4-QAM modulation.

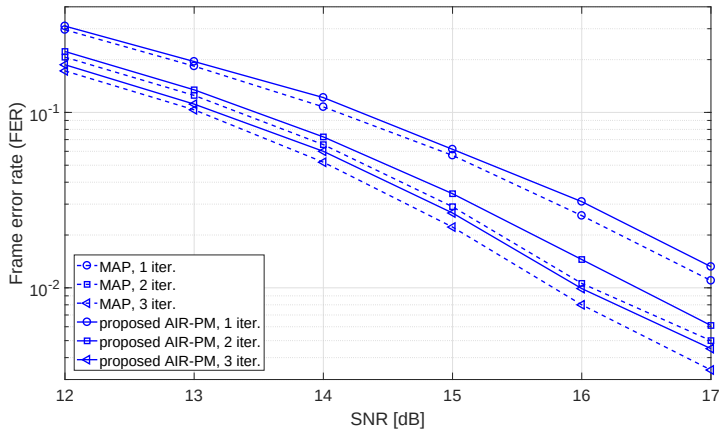


Figure 12: The FER performance of iterative detection and decoding under 6×6 real (3×3 complex) MIMO channels with 4-PAM (16-QAM) modulation. The metric of the AIR-PM detector with soft outputs is updated according to (48) and (49).

square estimates are of good quality. We also show that the chain rule holds for the AIR computation of the AIR-PM detector, and the ergodic AIR with correlated Rayleigh fading channel is maximized by the layers with higher transmit power. As the AIR-PM detector has low complexity and the detection process is fully parallelized for all layers, it results in flexible and efficient hardware design. In addition, numerical simulations show that, the AIR-PM detector has superior frame error rate performance than the state-of-the-art detectors.

Appendix A: Proof of Proposition 2

First, we show that, when $0 < N_0 < \infty$,

$$b_{n,n} - \boldsymbol{b}_n \tilde{\boldsymbol{B}}^{-1}_{(n,N)\ominus\nu} \boldsymbol{b}_n^\dagger > 0, \quad 1 \le n \le N. \tag{56}$$

As $\boldsymbol{B}$ is the MSE matrix, when $0 < N_0 < \infty$, $\boldsymbol{0} \prec \boldsymbol{B} \prec \boldsymbol{I}$ holds. Rewrite $\boldsymbol{B}$ in block form as

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 \\ \boldsymbol{B}_1^\dagger & \boldsymbol{B}_2 \end{pmatrix},$$

where submatrices $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ are of dimensions $(N-\nu)\times(N-\nu)$, $(N-\nu)\times\nu$, and $\nu\times\nu$, respectively. When $1 \le n \le N-\nu$, $\tilde{\boldsymbol{B}}_{(n,N)\ominus\nu} = \boldsymbol{B}_2$. As $\boldsymbol{B} \succ \boldsymbol{0}$, the Schur complement of $\boldsymbol{B}_2$, which equals $\left(\boldsymbol{B}_0 - \boldsymbol{B}_1^\dagger \boldsymbol{B}_2^{-1} \boldsymbol{B}_1\right)$, is also positive definite. Therefore, the diagonal elements of $\left(\boldsymbol{B}_0 - \boldsymbol{B}_1^\dagger \boldsymbol{B}_2^{-1} \boldsymbol{B}_1\right)$, with the nth element being $\left(b_{n,n} - \boldsymbol{b}_n \boldsymbol{B}_2^{-1} \boldsymbol{b}_n^\dagger\right)$, are positive.

On the other hand, when $N-\nu < n \le N$, $\tilde{\boldsymbol{B}}_{(n,N)\ominus\nu} = \tilde{\boldsymbol{B}}_n$. As $\begin{pmatrix} b_{n,n} & \boldsymbol{b}_n \\ \boldsymbol{b}_n^\dagger & \tilde{\boldsymbol{B}}_n \end{pmatrix}$ is a principle submatrix of $\boldsymbol{B}$, it is positive definite. Hence, the Schur complement of $\tilde{\boldsymbol{B}}_n$, which equals $\left(b_{n,n} - \boldsymbol{b}_n \tilde{\boldsymbol{B}}_n^{-1} \boldsymbol{b}_n^\dagger\right)$, is also positive. Therefore, (56) holds for all n.

With the decomposition $\boldsymbol{I} + \boldsymbol{G}_{\mathrm{r}} = \boldsymbol{U}^\dagger \boldsymbol{U}$, where $\boldsymbol{U}$ is upper-triangular with only the elements along the main diagonal and the last ν columns can take non-zero values. Furthermore, we constrain the diagonal elements to satisfy $u_{n,n} \ge 0$. Similar to $\boldsymbol{B}$, we also rewrite $\boldsymbol{U}$ in a block matrix form as

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_0 & \boldsymbol{U}_1 \\ \boldsymbol{0} & \boldsymbol{U}_2 \end{pmatrix},$$

where $\boldsymbol{U}_0$, $\boldsymbol{U}_1$, and $\boldsymbol{U}_2$ have the same sizes as $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, and $\boldsymbol{B}_2$, respectively. Then, we have

$$\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{U}^\dagger \boldsymbol{U}\right) = \mathrm{Tr}\big(\boldsymbol{U}_0 \boldsymbol{B}_0 \boldsymbol{U}_0^\dagger + \boldsymbol{U}_1 \boldsymbol{B}_1^\dagger \boldsymbol{U}_0^\dagger + \boldsymbol{U}_0 \boldsymbol{B}_1 \boldsymbol{U}_1^\dagger + \boldsymbol{U}_1 \boldsymbol{B}_2 \boldsymbol{U}_1^\dagger + \boldsymbol{U}_2 \boldsymbol{B}_2 \boldsymbol{U}_2^\dagger\big). \tag{57}$$

Taking the derivative of $\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{U}^\dagger \boldsymbol{U}\right)$ with respect to $\boldsymbol{U}_1$ and setting it equal to zero results in

$$\boldsymbol{U}_1 = -\boldsymbol{U}_0 \boldsymbol{B}_1 \boldsymbol{B}_2^{-1}. \tag{58}$$

Inserting (58) back into (57), we obtain

$$\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{U}^\dagger \boldsymbol{U}\right) \;=\; \mathrm{Tr}\left(\boldsymbol{U}_0 \big(\boldsymbol{B}_0 - \boldsymbol{B}_1 \boldsymbol{B}_2^{-1} \boldsymbol{B}_1^\dagger\big) \boldsymbol{U}_0^\dagger + \boldsymbol{U}_2 \boldsymbol{B}_2 \boldsymbol{U}_2^\dagger\right). \tag{59}$$

With (59), the AIR in (22) can be rewritten as

$$\begin{aligned} I_{\mathrm{AIR}} &= N + \sum_{n=1}^{N} \ln\det\left(\boldsymbol{U}^{\dagger}\boldsymbol{U}\right) - \mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{U}^{\dagger}\boldsymbol{U}\right) \\ &= I_1 + I_2, \end{aligned} \tag{60}$$

where

$$I_1 = N-\nu+2\sum_{n=1}^{N-\nu} \ln u_{n,n} - \mathrm{Tr}\left(\boldsymbol{U}_0\left(\boldsymbol{B}_0 - \boldsymbol{B}_1\boldsymbol{B}_2^{-1}\boldsymbol{B}_1^{\dagger}\right)\boldsymbol{U}_0^{\dagger}\right), \tag{61}$$

$$I_2 = \nu+2\sum_{n=N-\nu+1}^{N} \ln u_{n,n} - \mathrm{Tr}\left(\boldsymbol{U}_2\boldsymbol{B}_2^{\dagger}\boldsymbol{U}_2^{\dagger}\right). \tag{62}$$

Firstly, we consider $1 \le n \le N-\nu$, in which case $\tilde{\boldsymbol{B}}_{(n,N)\ominus\nu} = \boldsymbol{B}_2$ holds. As $\boldsymbol{U}_0$ is diagonal with $u_{n,n}$ being its nth diagonal element, taking the derivative of I_1 with respect to $u_{n,n}$ and setting it equal to zero results in the optimal solution

$$u_{n,n} = \sqrt{\left(b_{n,n} - \boldsymbol{b}_n\boldsymbol{B}_2^{-1}\boldsymbol{b}_n^{\dagger}\right)^{-1}}. \tag{63}$$

And from (58), the optimal $\boldsymbol{u}_n$ reads

$$\boldsymbol{u}_n = -u_{n,n}\boldsymbol{b}_n\boldsymbol{B}_2^{-1}. \tag{64}$$

Secondly, we consider $N-\nu < n \le N$. In this case, $\tilde{\boldsymbol{B}}_{(n,N)\ominus\nu} = \tilde{\boldsymbol{B}}_n$. As

$$\mathrm{Tr}\left(\boldsymbol{U}_2\boldsymbol{B}_2^{\dagger}\boldsymbol{U}_2^{\dagger}\right) = \sum_{n=N-\nu+1}^{N} \begin{pmatrix} u_{n,n} & \boldsymbol{u}_n \end{pmatrix} \begin{pmatrix} b_{n,n} & \boldsymbol{b}_n \\ \boldsymbol{b}_n^{\dagger} & \tilde{\boldsymbol{B}}_n \end{pmatrix} \begin{pmatrix} u_{n,n} & \boldsymbol{u}_n \end{pmatrix}^{\dagger}, \tag{65}$$

taking the derivative of $\mathrm{Tr}\left(\boldsymbol{U}_2\boldsymbol{B}_2^{\dagger}\boldsymbol{U}_2^{\dagger}\right)$ with respect to $\boldsymbol{u}_n$ and setting it equal to zero results in

$$\boldsymbol{u}_n = -u_{n,n}\boldsymbol{b}_n\tilde{\boldsymbol{B}}_n^{-1}. \tag{66}$$

Inserting (66) back into (65) and (62), I_2 can be rewritten as

$$I_2 = \nu + \sum_{n=N-\nu+1}^{N} \left(2\ln u_{n,n} - \left(b_{n,n} - \boldsymbol{b}_n\tilde{\boldsymbol{B}}_n^{-1}\boldsymbol{b}_n^{\dagger}\right)u_{n,n}^2\right).$$

Then, taking the derivative of I_2 with respect to $u_{n,n}$ and setting it equal to zero results in

$$u_{n,n} = \sqrt{\left(b_{n,n} - \boldsymbol{b}_n\tilde{\boldsymbol{B}}_n^{-1}\boldsymbol{b}_n^{\dagger}\right)^{-1}}. \tag{67}$$

Combining (63)-(64) and (66)-(67) proves the first part in Proposition 2.

Next we prove the second part in Proposition 2. From (63) it holds that,

$$\mathrm{Tr}\Big(\boldsymbol{U}_0\Big(\boldsymbol{B}_0-\boldsymbol{B}_1\boldsymbol{B}_2^{-1}\boldsymbol{B}_1^{\dagger}\Big)\boldsymbol{U}_0^{\dagger}\Big)=\sum_{n=1}^{N-\nu}u_{n,n}^2\left(b_{n,n}-\boldsymbol{b}_n\boldsymbol{B}_2^{-1}\boldsymbol{b}_n^{\dagger}\right)=N-\nu.$$

Hence, from (61) we have

$$I_1=2\sum_{n=1}^{N-\nu}\ln u_{n,n}. \tag{68}$$

On the other hand, inserting (66) into (65) results in

$$\mathrm{Tr}\Big(\boldsymbol{U}_2\boldsymbol{B}_2^{\dagger}\boldsymbol{U}_2^{\dagger}\Big)=\sum_{n=N-\nu+1}^{N}u_{n,n}^2\left(b_{n,n}-\boldsymbol{b}_n\tilde{\boldsymbol{B}}_n^{-1}\boldsymbol{b}_n^{\dagger}\right)=\nu,$$

where the last equality is from (67). Hence, from (62) it holds that

$$I_2=2\sum_{n=N-\nu+1}^{N}\ln u_{n,n}. \tag{69}$$

Combing (68) and (69) proves (28), which completes the proof.

Appendix B: Proof of Corollary 1

As $\mathbf{0}\prec\boldsymbol{B}\prec\boldsymbol{I}$, the diagonal element $0<b_{n,n}<1$. From Proposition 2, for $1\leq n\leq N$,

$$g_{n,n}\quad\geq\quad u_{n,n}^2-1>\frac{1}{b_{n,n}}-1>0.$$

Appendix C: Proof of Lemma 1

By definitions of submatrices $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ as in Appendix A, and utilizing the block matrix inversion lemma [32], after some manipulations, it holds that

$$\boldsymbol{B}_0-\boldsymbol{B}_1^{\dagger}\boldsymbol{B}_2^{-1}\boldsymbol{B}_1=\boldsymbol{I}-\boldsymbol{H}_b^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger}+N_0\boldsymbol{I}\right)^{-1}\boldsymbol{H}_b, \tag{70}$$

$$\boldsymbol{B}_2=\left(\boldsymbol{I}+\boldsymbol{H}_a^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger}+N_0\boldsymbol{I}\right)^{-1}\boldsymbol{H}_a\right)^{-1}. \tag{71}$$

By the definition of $u_{n,n}$ in Proposition 2 and from (70), when $1 \le n \le N-\nu$,

$$u_{n,n}^{-2} = 1 - \boldsymbol{h}_n^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger} + N_0\boldsymbol{I}\right)^{-1}\boldsymbol{h}_n. \tag{72}$$

On the other hand, as $I\left(\boldsymbol{y}; x_n|\boldsymbol{x}_a\right)$ equals

$$\begin{aligned} I\left(\boldsymbol{y}; x_n|\boldsymbol{x}_a\right) &= \ln\left(1 + \boldsymbol{h}_n^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger} - \boldsymbol{h}_n\boldsymbol{h}_n^{\dagger} + N_0\boldsymbol{I}\right)^{-1}\boldsymbol{h}_n\right) \\ &= -\ln\left(1 - \boldsymbol{h}_n^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger} + N_0\boldsymbol{I}\right)^{-1}\boldsymbol{h}_n\right), \end{aligned} \tag{73}$$

combing (72) and (73), (35) follows. To prove (36), as from (71),

$$I\left(\boldsymbol{y}; \boldsymbol{x}_a\right) = \ln\left(\boldsymbol{I} + \boldsymbol{H}_a^{\dagger}\left(\boldsymbol{H}_b\boldsymbol{H}_b^{\dagger} + N_0\boldsymbol{I}\right)^{-1}\boldsymbol{H}_a\right) = -\ln\det\boldsymbol{B}_2, \tag{74}$$

utilizing the matrix determinant lemma, it holds that

$$\det\boldsymbol{B}_2 = \det\tilde{\boldsymbol{B}}_{N-\nu+1} = u_{N-\nu+1,N-\nu+1}^{-2}\det\tilde{\boldsymbol{B}}_{N-\nu+2} = \prod_{n=N-\nu+1}^{N} u_{n,n}^{-2}. \tag{75}$$

Combining (74) and (75) proves (36).

Appendix D: Proof of Proposition 4

Without loss of generality, consider two selections of $\boldsymbol{x}_a$ that only differ at the first layer as

$$\begin{aligned} \boldsymbol{x}_a^1 &= \left(x_{n_1}, x_{N-\nu+2}, x_{N-\nu+3}, \cdots, x_N\right), \\ \boldsymbol{x}_a^2 &= \left(x_{n_2}, x_{N-\nu+2}, x_{N-\nu+3}, \cdots, x_N\right), \end{aligned}$$

where $n_1 \neq n_2$ and $\delta_{n_1} \le \delta_{n_2}$. From Proposition 3,

$$I_{\mathrm{AIR}}^{i}\left(\boldsymbol{y}; \boldsymbol{x}\right) = I\left(\boldsymbol{y}; \boldsymbol{x}_a^i\right) + \sum_{n=1, n\neq n_1}^{N-\nu+1} I\left(\boldsymbol{y}; x_n|\boldsymbol{x}_a^i\right),\ i = 1, 2, \tag{76}$$

and from the chain rule of mutual information,

$$\begin{aligned} I_{\mathrm{AIR}}\left(\boldsymbol{y}; \boldsymbol{x}_a^1, x_{n_2}\right) &= I_{\mathrm{AIR}}\left(\boldsymbol{y}; \boldsymbol{x}_a^2, x_{n_1}\right) \\ &= I\left(\boldsymbol{y}; \boldsymbol{x}_a^1\right) + I\left(\boldsymbol{y}; x_{n_2}|\boldsymbol{x}_a^1\right) \\ &= I\left(\boldsymbol{y}; \boldsymbol{x}_a^2\right) + I\left(\boldsymbol{y}; x_{n_1}|\boldsymbol{x}_a^2\right). \end{aligned} \tag{77}$$

Combining (76) and (77), it holds that

$$I^1_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x}) - I^2_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x}) = \sum_{n=1,n\neq n_1,n_2}^{N-\nu+1} \Big(I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^1_a\right) - I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^2_a\right) \Big). \tag{78}$$

Hence, to show $\mathbb{E}_{\boldsymbol{H}}\left[I^1_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x})\right] \leq \mathbb{E}_{\boldsymbol{H}}\left[I^2_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x})\right]$, it is sufficient to prove

$$\mathbb{E}_{\boldsymbol{H}}\left[I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^1_a\right)\right] \leq \mathbb{E}_{\boldsymbol{H}}\left[I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^2_a\right)\right]$$

for $1 \leq n \leq (N-\nu+1)$ and $n \neq n_1, n_2$.

On the other hand, as $\boldsymbol{h}_{n_i}$ can be modeled by $\boldsymbol{h}_{n_i} = \sqrt{\delta_{n_i}}\boldsymbol{h}$, where $\boldsymbol{h} \sim \mathcal{CN}\left(\boldsymbol{0}, \frac{1}{N}\boldsymbol{\Sigma}\right)$, replacing $\boldsymbol{h}_{n_i}$ with $\sqrt{\delta_{n_i}}\boldsymbol{h}$ will not change the ergodic AIR, that is,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{H}}\left[I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^{3-i}_a\right)\right] &= \mathbb{E}_{\boldsymbol{H}}\left[\ln\left(1+\boldsymbol{h}^\dagger_n\left(\boldsymbol{h}_{n_i}\boldsymbol{h}^\dagger_{n_i}+N_0\boldsymbol{I}+\sum_{k=1,n\neq n_1,n_2}^{N-\nu+1}\boldsymbol{h}_k\boldsymbol{h}^\dagger_k\right)^{-1}\boldsymbol{h}_n\right)\right] \\
&= \mathbb{E}_{\boldsymbol{H}}\left[\ln\left(1+\boldsymbol{h}^\dagger_n\left(\delta^2_{n_i}\boldsymbol{h}\boldsymbol{h}^\dagger+N_0\boldsymbol{I}+\sum_{k=1,n\neq n_1,n_2}^{N-\nu+1}\boldsymbol{h}_k\boldsymbol{h}^\dagger_k\right)^{-1}\boldsymbol{h}_n\right)\right].
\end{aligned}$$

As logarithmic function is monotonically increasing,

$$\mathbb{E}_{\boldsymbol{H}}\left[I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^1_a\right)\right] \leq \mathbb{E}_{\boldsymbol{H}}\left[I\left(\boldsymbol{y}; x_n|\boldsymbol{x}^2_a\right)\right]$$

follows from $\delta_{n_1} \leq \delta_{n_2}$. Hence, replacing a layer in $\boldsymbol{x}_a$ with another layer that has larger δ_n increases $I^{\mathrm{e}}_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x})$. By induction, the selection of $\boldsymbol{x}_a$ maximizing $I^{\mathrm{e}}_{\mathrm{AIR}}(\boldsymbol{y};\boldsymbol{x})$ comprises layers with ν largest δ_n, which are the last ν layers.

Appendix E: Proof of Proposition 5

Let $\big(i(1), i(2), \cdots, i(N)\big)$ be an arbitrary permutation of $(1, 2, \cdots, N)$. Denote $\eta_n = \sigma_n\delta_n$ and $\tilde{\eta}_n = \eta_{i(n)}$, and set

$$\begin{aligned}
\boldsymbol{x}_a &= \big(x_{i(N-\nu+1)}, x_{i(N-\nu+2)}, \cdots, x_{i(N)}\big), \\
\boldsymbol{x}_b &= \big(x_{i(1)}, x_{i(2)}, \cdots, x_{i(N-\nu)}\big).
\end{aligned}$$

In the low SNR regime, the ergodic capacity can be approximated as

$$\mathcal{L}(\boldsymbol{H}, N_0) = N\mathbb{E}_\lambda\left[\ln\left(1+\frac{\lambda}{N_0}\right)\right] \approx \frac{N}{N_0}\mathbb{E}_\lambda[\lambda], \tag{79}$$

where λ is the unordered eigenvalue of $\boldsymbol{H}_{\text{i.i.d.}}^{\dagger}\boldsymbol{H}_{\text{i.i.d.}}$, which obeys the complex Wishart distribution [38]. Therefore, it holds that

$$\mathbb{E}_{\lambda}\left[\lambda\right]=\frac{1}{N}\text{Tr}\left\{\mathbb{E}\left[\boldsymbol{H}^{\dagger}\boldsymbol{H}\right]\right\}=\frac{1}{N}\text{Tr}\left\{\mathbb{E}\left[\boldsymbol{\Delta}^{1/2}\boldsymbol{H}_{\text{i.i.d.}}^{\dagger}\boldsymbol{\Sigma}\boldsymbol{H}_{\text{i.i.d.}}\boldsymbol{\Delta}^{1/2}\right]\right\}=\frac{1}{N}\sum_{n=1}^{N}\eta_n. \tag{80}$$

Inserting (79) and (80) back into (46), it holds that

$$I_{\text{AIR}}^{\text{e}}\left(\boldsymbol{y};\boldsymbol{x}\right)=\frac{1}{N_0}\left(\sum_{n=1}^{N}\eta_n+(N-\nu-1)\sum_{n=1}^{N-\nu}\tilde{\eta}_n-\sum_{n=1}^{N-\nu}\sum_{k=1,k\neq n}^{N-\nu}\tilde{\eta}_k\right)=\frac{1}{N_0}\sum_{n=1}^{N}\eta_n,$$

which shows that, the ergodic AIR is independent of the selection of $\boldsymbol{x}_a$.

In the high SNR regime, as $\mathcal{L}^{\infty}\left(\boldsymbol{H},N_0\right)$ is the ergodic capacity which is irrelevant of the selection of $\boldsymbol{x}_a$, form (46) and (47), the optimal selection of $\boldsymbol{x}_a$ shall minimize

$$\tilde{\mathcal{L}}=(N-\nu-1)(N-\nu)\mathcal{L}^{\infty}\left(\boldsymbol{H}_b,N_0\right)-(N-\nu-1)\sum_{n=1}^{N-\nu}\mathcal{L}^{\infty}\left(\boldsymbol{H}_{b,n},N_0\right). \tag{81}$$

From [33, Lemma 2], when $N=K$, the high SNR power-offset

$$\mathcal{L}^{\infty}\left(\boldsymbol{H}_b,N_0\right)=-\ln\left(\sum_{k=2}^{N-\nu}k^{-1}-\gamma\right)-\frac{1}{N-\nu}\sum_{k=1}^{N-\nu}\ln\tilde{\eta}_k, \tag{82}$$

$$\mathcal{L}^{\infty}\left(\boldsymbol{H}_{b,n},N_0\right)=-\ln\left(\sum_{k=2}^{N-\nu-1}k^{-1}-\gamma\right)-\frac{1}{N-\nu-1}\sum_{k=1,k\neq n}^{N-\nu}\ln\tilde{\eta}_k, \tag{83}$$

where $\gamma\approx 0.5772$ is the Euler-Mascheroni constant. Inserting (82) and (83) back into (81) yields

$$\tilde{\mathcal{L}}\propto-(N-\nu-1)\sum_{k=1}^{N-\nu}\ln\tilde{\eta}_k+\sum_{n=1}^{N-\nu}\sum_{k=1,k\neq n}^{N-\nu}\ln\tilde{\eta}_k=0.$$

Hence, in the high SNR regime, the ergodic AIR is also independent of the selection of $\boldsymbol{x}_a$.

References

[1] 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 Process. Mag.*, vol. 30, no. 1, pp. 40-60, Jan. 2013.

[2] "Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding," v12.4.0, 3GPP TS 36.212, Mar. 2015.

[3] E. G. Larsson, "MIMO detection methods: How they work [Lecture Notes]," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 91-95, May 2009.

[4] P. Silvola, K. Hooli, and M. Juntti, "Suboptimal soft-output MAP detector with lattice reduction," *IEEE Signal Process. Lett.*, vol. 13, no. 6, pp. 321-324, Jun. 2006.

[5] S. Yang, T. Lv, R. G. Maunder, and L. Hanzo, "From nominal to true a posteriori probabilities: An exact Bayesian theorem based probabilistic data association approach for iterative MIMO detection and decoding," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2782-2793, Jul. 2013.

[6] A. Hu, T. Lv, H. Gao, Z. Zhang, and S. Yang,"An ESPRIT-based approach for 2-D localization of incoherently distributed sources in massive MIMO systems," *IEEE J. Sel. Topics Signal Process.,* vol. 8, no. 5, pp. 996-1011, Oct. 2014.

[7] G. D. Forney Jr., "Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 363-378, May 1972.

[8] C. Studer, A. Burg, and H. Bölcskei, "Soft-output sphere decoding: algorithms and VLSI implementation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 290-300, Feb. 2008.

[9] C. Studer and H. Bölcskei, "Soft-input soft-output single tree-search sphere decoding," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4827-4842, Oct. 2010.

[10] B. M. Hochwald and S. Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 389-399, Mar. 2003.

[11] L. G. Barbero and J. S. Thompson , "Fixing the complexity of the sphere decoder for MIMO Detection," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2131-2142, Jun. 2008.

[12] T. M. Aulin, "Breadth-first maximum likelihood sequence detection: Basics," *IEEE Trans. Commun.*, vol. 47, no. 2, pp. 208-216, Feb. 1999.

[13] K. J. Kim, T. Reid, and R. A. Iltis, "Iterative soft-QRD-M for turbo coded MIMO-OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 56, no. 7, pp. 1043-1046, Jul. 2008.

[14] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 491-503, Mar. 2006.

[15] S. Chen, T. Zhang, and Y. Xin, "Relaxed K-best MIMO signal detector design and VLSI implementation," *IEEE Trans. VLSI*, vol. 15, no. 3, pp. 328-337, Mar. 2007.

[16] R. Wang and G. B. Giannakis, "Approaching MIMO channel capacity with reduced-complexity soft sphere decoding," *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Minneapolis, U.S.A., Mar. 2004, vol. 3, pp. 1620-1625.

[17] E. G. Larsson and J. Jaldén, "Fixed-complexity soft MIMO detection via partial marginalization," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3397-3407, Aug. 2008.

[18] D. Persson and E. G. Larsson, "Partial marginalization soft MIMO detection with higher order constellations," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 453-458, Jan. 2011.

[19] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953-1967, Nov. 1994.

[20] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, pp. 2315-2328, Nov. 2000.

[21] I. Abou-Faycal and A. Lapidoth, "On the capacity of reduced complexity receivers for intersymbol interference channels," *Proc.* Conf. Inf. Sci. and Sys. (CISS), Princeton University, Mar. 2000, pp. WA4 32-37.

[22] F. Rusek and A. Prlja, "Optimal channel shortening of MIMO and ISI channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 810-818, Feb. 2012.

[23] S. Hu and F. Rusek, "On the design of reduced state demodulators with interference cancellation for iterative receivers," *IEEE Conf. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Hongkong, Sep. 2015, pp. 981-985.

[24] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "A Low-complexity channel shortening receiver with diversity support for evolved 2G device," *IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1-7.

[25] S. Hu, F. Rusek, and N. Al-Dhahir, "Comparison of two channel shortening approaches for MIMO-ISI channels," *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Doha, Qatar, Apr. 2016, pp. 1-7.

[26] S. Hu and F. Rusek, "On the design of channel shortening demodulators for iterative receivers in MIMO and ISI channels," *submitted to IEEE Trans. Inf. Theory, arXiv preprint: 1506.07331*, May 2015.

[27] M. Sellathurai and S. Haykin, "Turbo-BLAST for wireless communications: theory and experiments," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2538-2546, Oct. 2002.

[28] T. K. Moon, *Error correction coding: mathematical methods and algorithms*, New York, U.S.A., Wiley, 2005.

[29] M. M. Mansour, "A near-ML MIMO subspace detection algorithm," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 408-412, Apr. 2015.

[30] F. Pérez-Cruz, M. R. D. Rodrigues, and S. Verdú, "MIMO Gaussian channels with arbitrary inputs: optimal precoding and power allocation," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1070-1084, Mar. 2010.

[31] M. M. Mansour and L. M. A. Jalloul, "Optimized configurable architectures for scalable soft-input soft-output MIMO detectors with 256-QAM," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4969-4984, Sep. 2015.

[32] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3, Johns Hopkins Studies in the Mathematical Sciences, JHU Press, 2013.

[33] M. R. McKay, I. B. Collings, and A. M. Tulino, "Achievable sum rate of MIMO MMSE receivers: A general analytic framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 396-410, Jan. 2010.

[34] H. Shin and J. H. Lee, "Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2636-2647, Oct. 2003.

[35] C. Xu, D. Liang, S. Sugiura, S. X. Ng and L. Hanzo, "Reduced-Complexity Approx-Log-MAP and Max-Log-MAP Soft PSK/QAM Detection Algorithms," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1415-1425, Apr. 2013.

[36] S. ten Brink, "Convergence of iterative decoding," *Electron. Lett.*, vol. 35, no. 10, pp. 806-808, May 1999.

[37] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727-1737, Oct. 2001.

[38] D. Maiwald and D. Kraus,"Calculation of moments of complex Wishart and complex inverse Wishart distributed matrices," *IEEE Proc. Radar, Sonar and Navigation*, vol. 147, no. 4, Aug. 2008, pp. 162-168.

Paper IV

Modulus Zero-Forcing Detection for MIMO Channels

We propose a modulus based zero-forcing (MZF) detection for multi-input multi-output (MIMO) channels. Traditionally, a ZF detector nulls out all interferences from other layers when detecting a current layer, which can yield suboptimal detection-performance due to the noise-enhancement issue. In many communication systems, finite alphabets such as M quadrature-amplitude-modulation (QAM) are widely used, which comprises $\sqrt{M}$ pulse-amplitude-modulation (PAM) symbols for the real and imaginary parts. With finite alphabets, one feasible way to improve ZF detection is to allow controllable interferences that can be removed away by modulus operations.

Based on: S. Hu and F. Rusek, "Modulus zero-forcing detection for MIMO channels," submitted to *IEEE Access*, Nov. 2017.

1 Introduction

We consider a standard multi-input multi-output (MIMO) channel model with a received signal $\tilde{\boldsymbol{y}}$ expressed as

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{n}}. \tag{1}$$

where $\tilde{\boldsymbol{H}}$, $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{n}}$ are the complex-valued MIMO channel, transmitted symbols and Gaussian noise, respectively.

Given received signal model (1), detecting $\tilde{\boldsymbol{x}}$ is referred to as a MIMO detection problem, which has a history that can be traced back about half a century and a review on it can be found in e.g., [1]. In general, maximum likelihood (ML) detection [2] yields optimal performance but with prohibitive complexity when the MIMO dimension and/or the input alphabet has large cardinality. Effective implementations of ML detection, such as sphere-decoding (SD) [3] can significantly reduce the complexity, but not overcome an exponential complexity in the number of symbol layers [4]. On the other hand, linear detectors [2] such as zero-forcing (ZF) and linear minimum-mean-square-error (LMMSE), have low complexities, but also suboptimal performances. One direction for improving linear detectors is lattice-aided-reduction (LAR) [5] based approaches, which use lattice-reduction (LR) algorithms, e.g., Lenstra-Lenstra-Lovász (LLL), to find a short and nearly orthogonal basis for the lattice induced by the MIMO channel [6].

Other than the existing approaches [1], as the transmitted symbols are drawn from finite alphabets such as quadrature-amplitude-modulation (QAM) and pulse-amplitude-modulation (PAM) symbols, the modulus can also be used in MIMO detection for improving the detection-performance. The modulus operation has been used in Tomlinson-Harashima precoding (THP) [7, 8] as a suboptimal approximation for dirty-paper coding (DPC) [9], and recently it has also be considered in the designs of integer-forcing (IF) receivers for MIMO channels [10–12]. The IF scheme in [10] requires the transmitter to employ the same lattice code [13] for each transmitted layer, which does not apply to most current communication systems. Besides, when higher-order modulations such as p-PAM are used, designing lattice codes over $\mathbb{Z}_p$ is challenging [13]. Simpler IF receivers dealing with linear binary codes such as turbo and LDPC are proposed in [11, 12], and the designs follow the same steps as in [10]. One disadvantage of the IF receivers is that, each transmit-antenna needs a separate encoding/decoding process, which is not the case for practical LTE systems where one codeword is split among transmit-antennas. Moreover, the IF design in [11] needs a separate encoding/decoding process per transmit-antenna and per bit-layer when higher-order modulations are employed. Another disadvantage is that, the receiver has to detect linear combinations of codewords across all transmit-antennas first, followed by a matrix inversion (over a finite-field) to recover the codeword at each layer.

To overcome these disadvantages in IF receiver designs, we consider a new approach to im-

prove linear detection with modulus operation, namely, the proposed modulus ZF (MZF) detector. Note that, MZF is conceptually different from previous IF receivers, although they share some similarities. The fundamental difference is that, with MZF, there is no encoding/decoding over a finite-field, as is the case for IF, to recover the transmit symbols (i.e., the linear combinations of codewords across all transmit-antennas). Alternatively, we design the MZF detector such that the transmit symbols on each transmit-antenna can be recovered directly by modulating away the interferences from the remaining transmit-antennas, and such a process is independent and fully parallel for different transmit-antennas. Such a design principle simplifies the operations, and can be well co-operated into practical systems such as LTE. To achieve this, a matrix which we refer to as the modulus matrix, must be carefully designed and optimized according to the specific modulation-order.

2 Preliminaries

We start with reviewing the standard ZF detection. Before proceeding, without loss of generality, the matrix $\boldsymbol{H}$ is always assumed to be a square matrix, obtained by a QR factorization or padding zero rows to the matrix if necessary. With the following definitions,

$$\boldsymbol{y} = \begin{bmatrix} \mathcal{R}\{\tilde{\boldsymbol{y}}\} \\ \mathcal{I}\{\tilde{\boldsymbol{y}}\} \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} \mathcal{R}\{\tilde{\boldsymbol{x}}\} \\ \mathcal{I}\{\tilde{\boldsymbol{x}}\} \end{bmatrix}, \ \boldsymbol{n} = \begin{bmatrix} \mathcal{R}\{\tilde{\boldsymbol{n}}\} \\ \mathcal{I}\{\tilde{\boldsymbol{n}}\} \end{bmatrix}, \ \boldsymbol{H} = \begin{bmatrix} \mathcal{R}\{\tilde{\boldsymbol{H}}\} & -\mathcal{I}\{\tilde{\boldsymbol{H}}\} \\ \mathcal{I}\{\tilde{\boldsymbol{H}}\} & \mathcal{R}\{\tilde{\boldsymbol{H}}\} \end{bmatrix}, \tag{2}$$

we can rewrite (1) as a real-valued model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{3}$$

where the $K \times K$ channel matrix $\boldsymbol{H}$ is known to the receiver, $\boldsymbol{x} = [x_1 \ \dots \ x_K]^{\mathrm{T}}$ contains PAM symbols from an alphabet $\mathcal{A} = \{\pm 1, \pm 3, \dots, \pm(\sqrt{M} - 1)\}$, and $\boldsymbol{n}$ is random Gaussian noise with a covariance matrix $(N_0/2)\boldsymbol{I}$. As the transmit power depends on M, the signal-to-noise (SNR) is defined as

$$\mathrm{SNR} = 2\mathbb{E}[|x_k|^2]/N_0. \tag{4}$$

The ZF detector is given by

$$\hat{\mathbf{x}} = \mathcal{Q}_{\mathcal{A}}\left(\boldsymbol{H}^{+}\boldsymbol{y}\right), \tag{5}$$

where $\mathcal{Q}_{\mathcal{A}}(\cdot)$ denotes entry-wise quantization to the nearest point in $\mathcal{A}$. This can be slightly rewritten as

$$\hat{x}_k = \mathcal{Q}_{\mathcal{A}}(r_k), \quad 1 \le k \le K$$

with

$$r_k = \boldsymbol{\delta}_k \boldsymbol{H}^+ \boldsymbol{y} \tag{6}$$

and

$$\boldsymbol{\delta}_k = [\underbrace{0 \ \ldots \ 0}_{k-1} \ 1 \ \underbrace{0 \ \ldots \ 0}_{K-k}].$$

For later use, we note that we may just as well replace the "1" in $\boldsymbol{\delta}_k$ with any arbitrary scalar value and equivalently work with

$$r_k = \tau x_k + w_k, \tag{7}$$

where w_k is zero-mean complex Gaussian noise with a variance $N_0 \tau^2 \|\boldsymbol{\delta}_k \boldsymbol{H}^+\|^2/2$. Accordingly, the post-processing SNR, which is independent of τ, becomes

$$\gamma_k = \frac{\mathrm{SNR}}{\|\boldsymbol{\delta}_k \boldsymbol{H}^+\|^2}, \tag{8}$$

with the SNR defined in (4).

3 Description of Proposed Method

A main issue with ZF is that, $\|\boldsymbol{\delta}_k \boldsymbol{H}^+\|^2$ in (8) is typically large and results in noise-enhancement when $\boldsymbol{H}$ is ill-conditioned [14]. To combat that, we make use of the underlying idea of THP but apply it to equalization, *without any involvement of the transmitter.* We propose to replace (6) with

$$r_k = (\tau \boldsymbol{\delta}_k + \boldsymbol{q}_k) \boldsymbol{H}^+ \boldsymbol{y} \tag{9}$$

where $\boldsymbol{q}_k = [q_{k1}, \ q_{k2}, \ \ldots \ q_{kK}]$ and $q_{k\ell} \in 2\mathbb{Z}$, i.e., the even integers. With that,

$$r_k = \tau x_k + \sum_{\ell=1}^{K} q_{k\ell} x_\ell + w_k, \tag{10}$$

and the noise power changes to $N_0 \|(\tau \boldsymbol{\delta}_k + \boldsymbol{q}_k) \boldsymbol{H}^+\|^2/2$.

Note that, (10) is identical to the received signal per user in vector perturbation (VP) [15, 16]. Therefore, further processing of (10) can follow the same steps as those in VP.

To detect x_k from r_k, we first state Property 1.

Property 1. *Let*

$$y = z + \alpha \sum_{m=1}^{M} p_m b_m,$$

where $\alpha \geq 1$, $|z|<2$, *and* $p_m, (b_m-1) \in 2\mathbb{Z}$. *Then,*

$$z = \begin{cases} (y \mod 4\alpha) - 2\alpha, & \text{if } \frac{1}{2}\sum_{m=1}^{M} p_m \text{ is odd,} \\ \big((y+2\alpha) \mod 4\alpha\big) - 2\alpha, & \text{otherwise.} \end{cases} \tag{11}$$

Proof. See Appendix A. □

In view of Property 1, we see that $q_{k\ell}$ and x_ℓ in (10) qualify as p_m and b_m, with setting $\alpha=1$. Further, from Property 1 we have that, τ must be selected such that

$$\tau \max_{a\in\mathcal{A}} |a| = \tau(\sqrt{M}-1) < 2. \tag{12}$$

To finalize the detector, we let

$$z_k = \begin{cases} (r_k \mod 4) - 2, & \text{if } \frac{1}{2}\sum_{\ell=1}^{K} q_{k\ell} \text{ is odd,} \\ \big((r_k+2) \mod 4\big) - 2, & \text{otherwise.} \end{cases} \tag{13}$$

which can be expressed as

$$z_k = \tau x_k + \tilde{w}_k, \tag{14}$$

where $\tilde{w}_k$ has a complicated distribution due to the modulus operation. The detected symbol $\hat{x}_k$ can now be obtained as

$$\hat{x}_k = \mathcal{Q}_{\tau\mathcal{A}}(z_k), \tag{15}$$

where the quantization is implemented on $\tau\mathcal{A}$, i.e., a scaled alphabet from $\mathcal{A}$.

We first remark that, the choice $\tau = 2/(\sqrt{M}-1)$ is not suitable in (12). This is so since if $\tau x_k + w_k = 2+\epsilon$ in (10), for some small $\epsilon>0$, then after modulus operation in (13) it may happen that $z_k = \epsilon - 2$ if $\sum_{\ell=1}^{K} q_{k\ell}$ is even. However, provided that $\tau < 2/(\sqrt{M}-1)$, at high SNR such wrap seldom happens and $\tilde{w}_k = w_k$ with high probability. Further, for constellation points x_k with small magnitude, $\tilde{w}_k = w_k$ holds with much higher probability than for constellation points x_k of large magnitude. To ensure equal error probability for all constellation points, we design τ such that:

The distance from 2 to the largest constellation point in $\tau\mathcal{A}$ *is half the distance between two points in* $\tau\mathcal{A}$.

Following this rule results in

$$\tau = 2^{(1-\log_2\sqrt{M})}, \tag{16}$$

which satisfies (12), and with out loss of generality we assume that $\log_2\sqrt{M}$ is an integer. For instance, if $M=4$, i.e., $\tilde{\boldsymbol{x}}$ is 4-QAM modulated and $\boldsymbol{x}$ comprises 2-PAM symbols, then τ equals 1. Similarly, if $\tilde{\boldsymbol{x}}$ is 16-QAM or 64-QAM modulated and $\boldsymbol{x}$ comprises 4-PAM or 8-PAM symbols, τ is set to 1/2 or 1/4, respectively.

With (16), we have that $\tilde{w}_k$ is "nearly" Gaussian (see [15, 16] for details) at high SNR. To optimize the receiver, we should solve

$$\boldsymbol{q}_k^{\text{opt}} = \arg\min_{\boldsymbol{q_k}} \|(\tau\boldsymbol{\delta}_k + \boldsymbol{q}_k)\boldsymbol{H}^+\|^2 \tag{17}$$

where elements of $\boldsymbol{q}_k$ are even integers. We rewrite (17) as

$$\boldsymbol{q}_k^{\text{opt}} = \arg\min_{\boldsymbol{q}_k} \|\boldsymbol{b}_k - \boldsymbol{q}_k\boldsymbol{B}\|^2, \tag{18}$$

which is an instance of sphere detection with integers [17], where

$$\boldsymbol{b}_k = \tau\boldsymbol{\delta}_k\boldsymbol{H}^+, \tag{19}$$

$$\boldsymbol{B} = -\boldsymbol{H}^+. \tag{20}$$

With the MZF detection introduced and without any extension, the basic MZF algorithm is given in Algorithm 1. We remind the reader that, the inputs $\boldsymbol{H}$ and $\boldsymbol{y}$ to the algorithm are assumed to be real-valued, while M denotes the cardinality of the complex-valued QAM constellation. We give an example of MZF detection in Appendix B to illustrate the detection process.

With the principle of MZF detection introduced, we have a few important remarks as follows.

Remark 1. *The MZF detection generalizes the ZF, and has the latter one as a special case when*

$$\left\|\boldsymbol{q}_k^{\text{opt}} - \boldsymbol{\delta}_k \odot \boldsymbol{q}_k^{\text{opt}}\right\|^2 = 0, \tag{21}$$

where $\odot$ is the Hadamard product. Hence, from a perspective of post-processing SNR, the MZF is always superior than the ZF.

Remark 2. *Following Remark 1, when the condition (21) holds, the modulus operation is unnecessary and ZF estimate shall be used.*

Remark 3. *In general, the minimum value achieved by $\boldsymbol{q}_k^{\text{opt}}$ in (18) increases as τ decreases. That is, for large constellations, the gain of the MZF decreases.*

To resolve the issue in Remark 3 and further improve the detection performance, we develop some useful extensions of the basic MZF detection.

Algorithm 1 MZF Algorithm

$\boldsymbol{H}$ is $K \times K$ real-valued
$\boldsymbol{y}$ is $K \times 1$ real-valued
M is cardinality of QAM constellation

function $\hat{\boldsymbol{x}} =$ModularZF($\boldsymbol{H}, \boldsymbol{y}, M$)
  $\tau = 2^{(1-\log_2\sqrt{M})}$
  $\boldsymbol{B} = -\boldsymbol{H}^+$

Preprocessing for each coherence interval

  for $k = 1$ **to** K
    $\boldsymbol{b}_k = \tau\boldsymbol{\delta}_k\boldsymbol{H}^+$
    Solve : $\boldsymbol{q}_k^{\text{opt}} = \arg\min_{\boldsymbol{q}_k} \|\boldsymbol{b}_k - \boldsymbol{q}_k\boldsymbol{B}\|^2$
  end for

Executed for every channel observation

  for $k = 1$ **to** K
    $r_k = (\tau\boldsymbol{\delta}_k + \boldsymbol{q}_k^{\text{opt}})\boldsymbol{H}^+\boldsymbol{y}$
    if $\left\|\boldsymbol{q}_k^{\text{opt}} - \boldsymbol{\delta}_k \odot \boldsymbol{q}_k^{\text{opt}}\right\|^2 = \boldsymbol{0}$ **then**
      $z_k = \boldsymbol{b}_k\boldsymbol{y}$
    else
      if $\frac{1}{2}\sum_{\ell=1}^{K} q_{k\ell}^{\text{opt}}$ is odd **then**
        $z_k = (r_k \mod 4) - 2$
      else
        $z_k = ((r_k + 2) \mod 4) - 2$
      end if
    end if
    $\hat{x}_k = \mathcal{Q}_{\tau\mathcal{A}}(z_k)$
  end for
end function

4 Extensions

In this section we introduce some extensions to the basic MZF detection. While Extension 1 and 4 are generalizations of the basic algorithm, Extension 2 is to resolve the issue mentioned for larger constellations and improves the detection for weak bit-layers, and Extension 3 is a decision feedback version of Extension 2.

4.1 Extension 1: A Scaled Modulus

This first extension arises from a slight relaxation of $\alpha = 1$ in the MZF detector. From Property 1, we can replace (13) as

$$z_k = \begin{cases} (r_k \mod 4\alpha) - 2\alpha, & \text{if } \frac{1}{2}\sum_{\ell=1}^{K} q_{k\ell} \text{ is odd}, \\ \big((r_k + 2\alpha) \mod 4\alpha\big) - 2\alpha, & \text{otherwise}. \end{cases} \tag{22}$$

This requires us to optimize, instead of (17),

$$(\boldsymbol{q}_k^{\text{opt}}, \alpha^{\text{opt}}) = \arg\min_{\alpha \geq 1,\, \boldsymbol{q}_k} \|(\tau \boldsymbol{\delta}_k + \alpha \boldsymbol{q}_k)\boldsymbol{H}^{+}\|^2. \tag{23}$$

Solving (23) is harder than solving (17) since it can be regarded as an instance of non-coherent sphere detection. Instead, we solve (17) first, and then plug the optimal solution into (23) and solve for the optimal α. That is, $\boldsymbol{q}_k^{\text{opt}}$ is obtained with (17), and

$$\alpha^{\text{opt}} = \arg\min_{\alpha \geq 1} \|(\tau \boldsymbol{\delta}_k + \alpha \boldsymbol{q}_k^{\text{opt}})\boldsymbol{H}^{+}\|^2, \tag{24}$$

where we sligthly abused notation since the pair $(\boldsymbol{q}_k^{\text{opt}}, \alpha^{\text{opt}})$ is in general not *jointly* optimal in the sense of (23). Although Extension 3 is intuitive, the gain seems marginal according to numerical results.

4.2 Extension 2: Bitwise MZF

An underlying assumption of this extension is that the bit-mapping to the symbols in $\mathcal{A}$ follows a natural labeling [11, 12], that is, a PAM symbol x_k should have a following form

$$x_k = \sum_{b=1}^{\log_2 \sqrt{M}} u_{kb} 2^{b-1}, \tag{25}$$

where $u_{kb} \in \{\pm 1\}$ correspond to information bits. Using Algorithm 1, the bits u_{kb} are determined by the output $\hat{x}_k = \mathcal{Q}_{\tau\mathcal{A}}(z_k)$, with setting $\tau = 2^{(1-\log_2 \sqrt{M})}$. As M increases, τ decreases and so are the gains with MZF detection. To resolve this for high-order modulations, we extend the symbol-based MZF detection in Algorithm 1 to a bitwise MZF detection.

Note that we can rewrite (10) as

$$r_k = \tau \sum_{b=1}^{\log_2 \sqrt{M}} u_{kb} 2^{b-1} + \sum_{\ell=1}^{K} q_{k\ell} x_\ell + w_k. \tag{26}$$

Supposing that we are interested in the n-th bit u_{kn}, we let

$$\tilde{x}_k = \sum_{b=1}^{n} u_{kb} 2^{b-1}, \tag{27}$$

which belongs to a 2^n-PAM alphabet. Setting $\tau(n) = 2^{1-n}$ in (26) yields

$$r_k = 2^{1-n}\tilde{x}_k + \sum_{b=n+1}^{\log_2\sqrt{M}} u_{kb} 2^{b-n} + \sum_{\ell=1}^{K} q_{k\ell} x_\ell + w_k. \tag{28}$$

It can be easily seen that $\frac{1}{2}\sum_{b=n+1}^{\log_2\sqrt{M}} u_{kb} 2^{b-n}$ is an odd integer so it qualifies as a valid value of $q_{k\ell}$, and u_{kn} can be detected as

$$\hat{u}_{kn} = \operatorname{sign}(z_k). \tag{29}$$

Therefore, for each bit-layer, a different value of τ is used and only a sign operation is needed for detecting bit u_{kn}. Extension 2 has a complexity increment over Algorithm 1 that, an optimization to find an optimal $\boldsymbol{q}_k$ is needed for each bit-layer.

Note that, according to Remark 2, when detecting the last bit-layer and if (21) holds, the ZF estimate shall be used for detection, while for detecting the other layers, modulus operations are still needed to module away the transmitted bits corresponding to higher bit-layers. With cooperating such a modification, the MZF with Extension 2 is summarized in Algorithm 2.

4.3 Extension 3: A Decision Feedback Version of Extension 2

An obstacle with Extension 2 is that τ decreases as n grows, and as previously mentioned, performance deteriorates. Small values of n correspond to weak bit-layers, and large n correspond to strong bit-layers. Thus, with Extension 2, predominantly the weak bit-layers can gain by the MZF, while the gain could be minuscule for strong bit-layers. A gain for weak bit-layers is important since it is typically these bit-layers that limit ultimate performance. However, we can also harvest a gain for strong bit-layers via a decision feedback mechanism. To prevent error propagation in decision feedback equalization, strong bits are typically detected first and then canceled. That option is not available for MZF, rather we detect the weakest bit-layer first and then move on to stronger ones such as in [12].

The method works as follows. First set $n = 1$ and follow the Extension 2 verbatim to obtain $\hat{\boldsymbol{u}}_1 = [\hat{u}_{11} \ \ldots \ \hat{u}_{K1}]^{\mathrm{T}}$. For notational convenience, define $\boldsymbol{y}_1 = \boldsymbol{y}$. Now construct

$$\boldsymbol{y}_2 = \frac{1}{2}\left(\boldsymbol{y}_1 - \boldsymbol{H}\hat{\boldsymbol{u}}_1\right). \tag{30}$$

Algorithm 2 MZF Algorithm with Extension 2
$\boldsymbol{H}$ is $K \times K$ real-valued
$\boldsymbol{y}$ is $K \times 1$ real-valued
M is cardinality of QAM constellation

function $\hat{\boldsymbol{x}} =$ModularZFExt2$(\boldsymbol{H}, \boldsymbol{y}, M)$
  $N = \log_2 \sqrt{M}$
  $\boldsymbol{B} = -\boldsymbol{H}^{+}$
  for $n = 1$ **to** N
    $\tau(n) = 2^{1-n}$
  end for

Preprocessing for each coherence interval

  for $k = 1$ **to** K
    $\boldsymbol{b}_k = \boldsymbol{\delta}_k \boldsymbol{H}^{+}$
    for $n = 1$ **to** N
      $\boldsymbol{b}_{kn} = \tau(n)\boldsymbol{b}_k$
      Solve : $\boldsymbol{q}_{k,n}^{\text{opt}} = \arg\min_{\boldsymbol{q}_k} \|\boldsymbol{b}_{kn} - \boldsymbol{q}_k \boldsymbol{B}\|^2$
    end for
  end for

Executed for every channel observation

  for $k = 1$ **to** K
    for $n = 1$ **to** N
      if $n = N$ **and** $\left\| \boldsymbol{q}_k^{\text{opt}} - \boldsymbol{\delta}_k \odot \boldsymbol{q}_k^{\text{opt}} \right\|^2 = \boldsymbol{0}$, **then**
        $z_k = \boldsymbol{b}_k \boldsymbol{y}$
      else
        $r_k = (\tau(n)\boldsymbol{\delta}_k + \boldsymbol{q}_k^{\text{opt}})\boldsymbol{H}^{+}\boldsymbol{y}$
        $z_k = (r_k \mod 4) - 2$
      end if
        $\hat{u}_{kn} = \text{sign}(z_k)$
    end for
  end for
end function

Provided that $\hat{\boldsymbol{u}}_1$ is correct, $\boldsymbol{y}_2$ is described with the same MIMO channel as $\boldsymbol{y}_1$, but with $\sqrt{M}/2$-PAM rather than $\sqrt{M}$-PAM inputs. Next, move on to $n = 2$ and keep $\tau = 1$. Since nor the value of τ neither the channel $\boldsymbol{H}$ has changed, the optimal vector $\boldsymbol{q}_k$ for

$n=2$ coincides with that already found for $n=1$. We then have that for $\boldsymbol{y}_2$

$$r_k = u_{k2} + \sum_{b=3}^{\log_2\sqrt{M}} u_{kb}2^{b-2} + \sum_{\ell=1}^{K} q_{k\ell}\frac{1}{2}(x_\ell - \hat{u}_{k1}) + w_k, \tag{31}$$

and $\hat{u}_{k2}$ is obtained by taking the sign of z_k as in (29). We proceed by

$$\boldsymbol{y}_3 = \frac{1}{2}\left(\boldsymbol{y}_2 - \boldsymbol{H}\hat{\boldsymbol{u}}_2\right), \tag{32}$$

and continue the process until all bit-layers have been detected.

Similarly, according to Remark 2, when detecting each bit-layer and when (21) holds, the ZF estimate shall be used. With considering this, the MZF with Extension 3 is summarized in Algorithm 3.

The Extension 3 is similar to Extension 2 in the sense that, the detection for all bit-layers only needs to take the signs of z_k as in (29), but it has less complexity since only one optimization of (17) is needed which is shared for all bit-layers. A drawback with Extension 3 is that, as for all decision-feedback based detections, the process of the bit-layers cannot be parallelized, which is, however, possible with Extension 2. Another drawback is potential error-propagations at low SNRs.

4.4 Extension 4: Replacing ZF by LMMSE

So far we have introduced the modulus detection using ZF, however, $\boldsymbol{H}^+$ can also be replaced by other linear detectors[1] such as LMMSE, which sets

$$\boldsymbol{H}^+ \triangleq \boldsymbol{H}^\dagger\left(\boldsymbol{H}\boldsymbol{H}^\dagger + N_0\boldsymbol{I}\right)^{-1}. \tag{33}$$

Casting in vector form and with a modulus matrix $\boldsymbol{T}$, the received signal after equalization equals

$$\boldsymbol{T}\boldsymbol{H}^+\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{T}\left(\boldsymbol{H}^+\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{x} + \boldsymbol{T}\boldsymbol{H}^+\boldsymbol{n},$$

where $\boldsymbol{T}=\tau\boldsymbol{I}+\boldsymbol{Q}$. The target of optimizing $\boldsymbol{q}_k$ in this case, is to minimize the interference plus noise power (for a given τ) such that

$$\tilde{\boldsymbol{q}}_k = \arg\min_{\boldsymbol{q}_k} \left\|(\tau\boldsymbol{\delta}_k + \boldsymbol{q}_k)\boldsymbol{E}\right\|^2, \tag{34}$$

[1] This is also known as regularized perturbation in VP [15].

Algorithm 3 MZF Algorithm with Extension 3
$\boldsymbol{H}$ is $K \times K$ real-valued
$\boldsymbol{y}$ is $K \times 1$ real-valued
M is cardinality of QAM constellation

function $\hat{\boldsymbol{x}}$ =ModularZFExt3($\boldsymbol{H}, \boldsymbol{y}, M$)
  $\tau = 1$
  $N = \log_2\sqrt{M}$
  $\boldsymbol{B} = -\boldsymbol{H}^+$

Preprocessing for each coherence interval

  for $k = 1$ **to** K
    $\boldsymbol{b}_k = \tau\boldsymbol{\delta}_k\boldsymbol{H}^+$
    Solve : $\boldsymbol{q}_k^{\text{opt}} = \arg\min_{\boldsymbol{q_k}} \|\boldsymbol{b}_k - \boldsymbol{q}_k\boldsymbol{B}\|^2$
  end for

Executed for every channel observation

  $\hat{\boldsymbol{y}} = \boldsymbol{y}$
  for $n = 1$ **to** N
    $\hat{\boldsymbol{u}}_n = 0$
    for $k = 1$ **to** K
      if $\left\|\boldsymbol{q}_k^{\text{opt}} - \boldsymbol{\delta}_k \odot \boldsymbol{q}_k^{\text{opt}}\right\|^2 = \boldsymbol{0}$, **then**
        $z_k = \tau\boldsymbol{\delta}_k\boldsymbol{H}^+\hat{\boldsymbol{y}}$
      else
        $r_k = (\tau\boldsymbol{\delta}_k + \boldsymbol{q}_k^{\text{opt}})\boldsymbol{H}^+\hat{\boldsymbol{y}}$
        $z_k = (r_k \mod 4) - 2$
      end if
      $\hat{u}_{kn} = \text{sign}(z_k)$
    end for
    $\hat{\boldsymbol{u}}_n = [\hat{u}_{1n}\, \hat{u}_{2n}\, \ldots\, \hat{u}_{Kn}]^{\text{T}}$
    $\hat{\boldsymbol{y}} = (\hat{\boldsymbol{y}} - \boldsymbol{H}\hat{\boldsymbol{u}_n})/2$
  end for
end function

where

$$\boldsymbol{E} = \left[\boldsymbol{H}^+\boldsymbol{H} - \boldsymbol{I},\ N_0\boldsymbol{H}^+\right]. \tag{35}$$

Note that, when $\boldsymbol{H}^+$ equals the pseudo-inverse of $\boldsymbol{H}$, $\boldsymbol{E}$ degrades to $\boldsymbol{H}^+$, which shows the generalization of MZF detection. The reason for introducing Extension 4 is that, the

ZF is suboptimal to LMMSE at low SNRs, in which cases it is beneficial to use LMMSE instead of ZF in the MZF detection. Since only $\boldsymbol{H}^{+}$ is replaced by LMMSE equalizer in Extension 4, all Algorithms 1-3 still apply with such a modification in (33) for the MZF detection.

There are also many other possible variations of the MZF detection, but we will not pursue any further. Next, we put an interest on comparing the MZF detector to a traditional LAR detector. The reason is that, solving (17) involves significant complexity, and we put forth an approximated solution based on LR with less computational efforts.

5 A Solution Based on, and a Comparison to, Lattice Reduction

Except for approximately solving (17) with LR, another reason for comparing MZF with LR detection is that, the obtained MZF allows for a direct comparison to LAR detectors. In LAR as well as the MZF, the most burdening task is to execute the LLL algorithm (or other similar algorithms), thus the complexities of LAR and MZF become virtually identical. As we will demonstrate, the detection-performance of MZF is superior in some cases.

5.1 A Quick Review of LAR

Given (3), LAR starts by performing the LLL algorithm on $\boldsymbol{H}$, so that we obtain $\bar{\boldsymbol{H}} = \boldsymbol{H}\boldsymbol{T}$ where $\boldsymbol{T}$ is unimodular and $\bar{\boldsymbol{H}}$ is nearly orthogonal. With $\boldsymbol{z} = \boldsymbol{T}^{-1}\boldsymbol{x}$ we have

$$\boldsymbol{y} = \bar{\boldsymbol{H}}\boldsymbol{z} + \boldsymbol{n}. \tag{36}$$

Performing ZF based on $\bar{\boldsymbol{H}}$ and quantizing to the nearest integers gives

$$\hat{\boldsymbol{z}} = \mathcal{Q}_{\mathbb{Z}}(\bar{\boldsymbol{H}}^{-1}\boldsymbol{y}) \tag{37}$$

from which one can obtain

$$\hat{\boldsymbol{x}} = \mathcal{Q}_{\mathcal{A}}(\boldsymbol{T}\hat{\boldsymbol{z}}).$$

Clearly, once $\bar{\boldsymbol{H}}$ has been established, the remaining steps are of minuscule complexity.

At this point, a reasonable question is, what the relation between LAR and MZF is, and whether they are equivalent? The answers to these questions are that, they are closely related, but not equivalent. Prior to quantization in (37), we can write

$$\begin{aligned} \boldsymbol{r} &= \bar{\boldsymbol{H}}^{-1}\boldsymbol{y} \\ &= \boldsymbol{T}^{-1}\boldsymbol{x} + \boldsymbol{w}. \end{aligned} \tag{38}$$

Since $\boldsymbol{T}$ is unimodular, so is $\boldsymbol{T}^{-1}$. On the other hand, casting in vector form, (9) equals

$$\begin{aligned}\boldsymbol{r} &= (\tau\boldsymbol{I}+\boldsymbol{Q})\,\boldsymbol{H}^{-1}\boldsymbol{y}\\ &= (\tau\boldsymbol{I}+\boldsymbol{Q})\,\boldsymbol{x}+\boldsymbol{w}.\end{aligned} \tag{39}$$

Comparing (38) and (39) with $\tau=1$, we see that in both cases $\boldsymbol{r}$ equals an integer-valued matrix multiplied with the data symbols, plus noise. However, the matrix $\boldsymbol{T}^{-1}$ in (38) has no particular structure (besides being unimodular) so the modulus operation in (13) is not available. This makes LAR, i.e., (38) and MZF, i.e., (39) fundamentally different, as the structure of (38) requires further processing in the form of (37), while (39) allows for further processing via (13).

5.2 An Approximate Solution to (17) based on LLL

In (18) we have the following problem to solve

$$\boldsymbol{q}^{\mathrm{opt}} = \arg\min_{\boldsymbol{q}} \|\boldsymbol{b}-\boldsymbol{q}\boldsymbol{B}\|^2, \tag{40}$$

where we omit the subscript k. Perform the LLL algorithm to $\boldsymbol{B}^{\mathrm{T}}$ so that we have

$$\bar{\boldsymbol{B}} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{T}.$$

Since $\boldsymbol{B}=-\boldsymbol{H}^{+}$ the LLL algorithm needs, similar to LAR, to be executed only once per coherence interval. We can now proceed as in the LAR case,

$$\hat{\boldsymbol{z}} = \mathcal{Q}_{\mathbb{Z}}(\bar{\boldsymbol{B}}^{-1}\boldsymbol{b}^{\mathrm{T}})$$

followed by

$$\boldsymbol{q}^{\mathrm{opt}} = \left[\mathcal{Q}_{2\mathbb{Z}}(\boldsymbol{T}\hat{\boldsymbol{z}})\right]^{\mathrm{T}}. \tag{41}$$

Note that, the optimization (40) itself is also an MIMO detection problem (but only needs to run once for a coherence-interval of the MIMO channel), therefore, there are also other low-complexity suboptimal algorithms to solve (40), such as using ZF or partial marginalization [18]. In the simulations, we will focus on the SD based optimal, and the LLL based suboptimal solutions for (40), respectively.

6 Numerical Results

In this section, we show some numerical results of the proposed MZF detection, as well as its extensions. In all tests, we test with $K\times K$ real-valued MIMO channels (each element is

an independent and identically distributed (i.i.d.) Gaussian variable with a zero-mean and unit-variance) with $\sqrt{M}$-PAM modulated symbols that are transfered from $K/2 \times K/2$ complex-valued MIMO channels and M-QAM modulated symbols. We simulate 50,000 channel realizations for each of the tests.

6.1 SINR Improvements

In Fig. 1 we show the post-processing SNR improvements with the MZF detector using Algorithm 1, and compare to a traditional ZF detector with different PAM modulations (i.e., τ values). As can be seen, the SNRs are greatly improved, especially for low-order modulations (or the weak bit-layers of high-order modulations with Extension 2). When τ decreases, the gains become smaller. We also test the MZF with Extension 1, where we can observe only marginal gains. Therefore, in the remaining tests we set $\alpha = 1$.

6.2 Uncoded Bit-Error-Rate (BER)

Next we show the uncoded BER performance. In Fig. 2 we compare MZF with ZF and ML under 6×6 MIMO with 4-PAM modulation. The MZF uses SD to find optimal $\boldsymbol{q}_k$. As can be seen, the MZF without extensions outperforms the ZF more than 2 dB at 0.1% BER. With Extension 2, the BER of the first bit-layer (weaker layer) is greatly improved by more than 4 dB at 0.1% BER and outperforms the second bit-layer, which justifies the application of Extension 3. With Extension 3, where feedback of the first bit-layer is used, the BER of the second bit-layer is also improved by more than 3 dB at 0.1% BER compared to the MZF with Extension 2. The gaps between ZF and ML are significantly reduced by the MZF, and the slopes of BER with MZF are also much steeper than the ZF, and close to those of ML. However, as also can be observed, the MZF has only marginal gains at low SNRs, and the decision-feedback approach performs even worse due to inaccurate feedbacks. This issue can be relieved by using LMMSE based approaches, i.e., Extension 4.

In Fig. 3 we repeat the tests in Fig. 2 under 4×4 MIMO with 8-PAM modulation, that is, three bit-layers are considered. The MZF with Extension 2 using SD is compared to ZF and ML. As already shown in Fig. 1, setting $\tau = 1/4$ for detecting the third-layer (strongest layer) only has small gains, and the BER performance is also close to ZF and therefore are not shown in Fig. 3. Nevertheless, the BER of the first and second bit-layers are significantly boosted by MZF. As can be seen, the MZF performs around 3 dB better than the ZF at 0.1% BER for the second bit-layer, and 7 dB better for the first bit-layer. Since the weakest bit-layer usually has a stronger impact on the decoding performance, the gains for the first bit-layer is of importance.

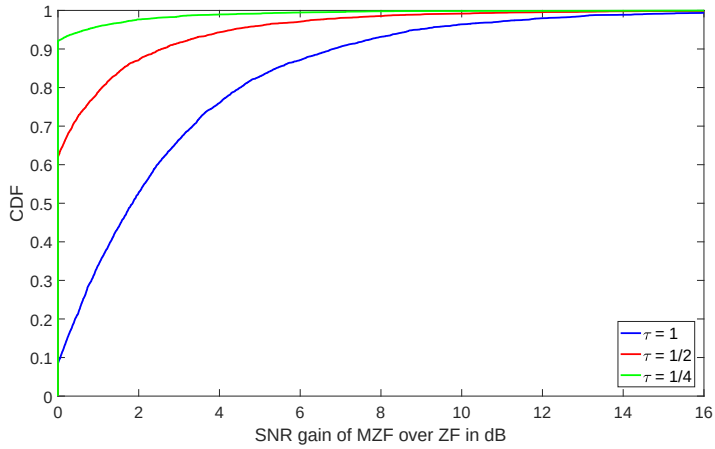


Figure 1: SNR gains under real-valued 12×12 MIMO with 2-PAM, 4-PAM and 8-PAM modulations, respectively.

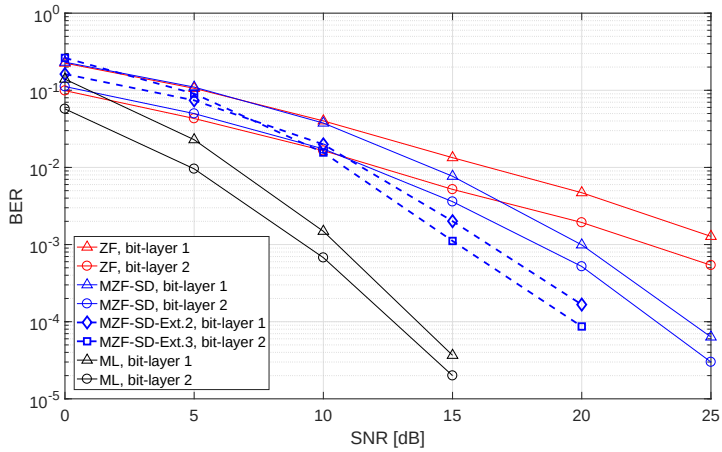


Figure 2: Uncoded BER under real-valued 6×6 MIMO with 4-PAM modulation.

6.3 Comparison with LAR

In Fig. 4 we compare the MZF with the LAR under 8×8 MIMO with 2-PAM modulation. The MZF uses both SD and LLL based approximations to find optimal $\boldsymbol{q}_k$. As can be seen, the MZF outperforms the LAR more than 1.5 dB at 0.1% BER, with a similar complexity for running LLL algorithm for LR.

Moreover, with Extension 4 (the LMMSE based detection), the BER at low SNRs are also improved with the MZF which is inferior to the original ZF based MZF at high SNRs. Nevertheless, with Extension 4, the MZF is 4 dB better than a normal ZF, and more than 2dB better than a normal LMMSE detector at 0.1% BER. Another observation is that, the SD based MZF is more than 2 dB better than the LLL based MZF, which shows that optimal selection of $\boldsymbol{q}_k$ is important.

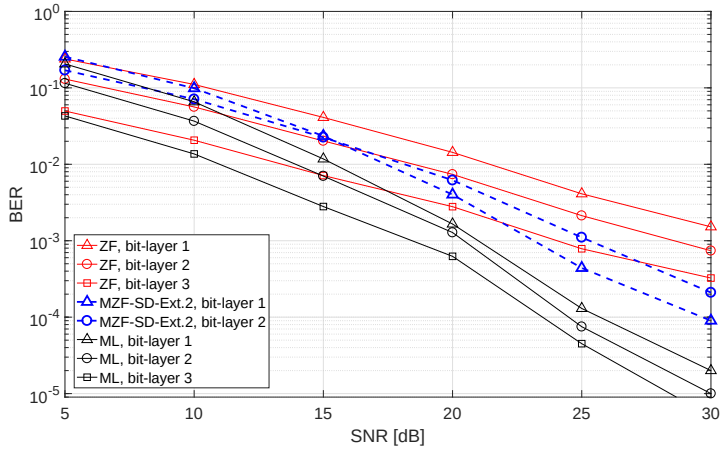


Figure 3: Uncoded BER under real-valued 4×4 MIMO with 8-PAM modulation.

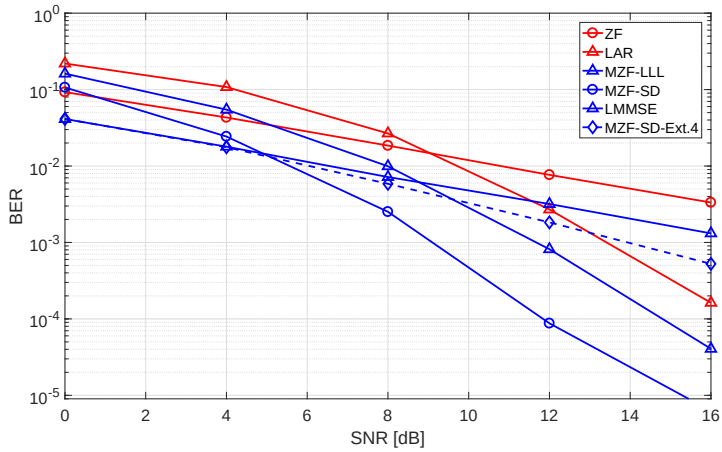


Figure 4: Uncoded BER under real-valued 8×8 MIMO with 2-PAM modulation.

7 Summary

We have proposed a novel modulus base zero-forcing (MZF) detection for multi-input multi-output (MIMO) channels, with possible extensions of the basic algorithm. The MZF detection shows significant gains in terms of post-processing signal-to-noise (SNR) and uncoded bit-error-rate (BER) compared to traditional linear detectors, at medium and high SNR scenarios and in particular for weak bit-layers. At low SNRs and with large modulation-orders, we have provided several possible extensions to improve the detection-performance of the MZF. Finding optimal modulus matrix itself is a complex MIMO detection problem, but it needs to be done only once per a coherence-interval of the MIMO channel using such as sphere-decoding (SD) and other suboptimal algorithms. In particular, with a similar complexity, the MZF with lattice-reduction (LR) based approach

outperforms the traditional lattice-aided-reduction (LAR) detector, which justifies its potential in MIMO detection.

Appendix A: Proof of Property 1

Since $p_m, (b_m-1) \in 2\mathbb{Z}$, we let $p_m = 2\tilde{p}_m$ and $b_m = 2\tilde{b}_m + 1$, where $\tilde{p}_m, \tilde{b}_m \in \mathbb{Z}$. Then,

$$\begin{aligned} y &= z + \alpha \sum_{m=1}^{M} p_m b_m \\ &= z + 4\alpha \sum_{m=1}^{M} \tilde{p}_m \tilde{b}_m + 2\alpha \sum_{m=1}^{M} \tilde{p}_m. \end{aligned} \tag{42}$$

Since $|z| < 2$ and $\alpha \geq 1$, it holds that $z + 2\alpha > 0$. If $\frac{1}{2} \sum_{m=1}^{M} p_m = \sum_{m=1}^{M} \tilde{p}_m$ is odd, we have

$$y \mod 4\alpha = z + 2\alpha; \tag{43}$$

Otherwise, if $\sum_{m=1}^{M} \tilde{p}_m$ is even, it also holds that

$$(y + 2\alpha) \mod 4\alpha = z + 2\alpha. \tag{44}$$

Combing (43) and (44), z can be obtained as in (11).

Appendix B: A 4×4 example for applying the MZF detection

Below we give a 4×4 real-valued MIMO example with 4-PAM modulation to illustrate the process of MZF detection, with assuming the channel, transmitted symbol vector and received signal vector as

$$\boldsymbol{H} = \begin{bmatrix} -6 & 0 & -1 & 5 \\ -3 & -2 & -1 & 1 \\ 1 & -5 & -6 & 0 \\ 1 & -1 & -3 & -2 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} 3 \\ 1 \\ 15 \\ 11 \end{bmatrix},$$

respectively. Then it can be shown that

$$\boldsymbol{H}^{+} = \frac{1}{185} \begin{bmatrix} -5 & -55 & 30 & -40 \\ 35 & -59 & -25 & 58 \\ -30 & 40 & -5 & -55 \\ 25 & -58 & 35 & -59 \end{bmatrix},$$

and the ZF estimate of $\boldsymbol{x}$ equals

$$\tilde{\boldsymbol{x}}_{\mathrm{ZF}} = \boldsymbol{H}^{+}\boldsymbol{y} = \frac{1}{185}\begin{bmatrix} -60 \\ 309 \\ -730 \\ -107 \end{bmatrix},$$

where only the third symbol is correctly detected.

Next we use the basic MZF detection with Algorithm 1. Setting $\tau = 1$ and run SD for optimization (17) yields an optimal $\boldsymbol{Q}$ as

$$\boldsymbol{Q} = \begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & -2 & 0 \\ 2 & 0 & 0 & -2 \end{bmatrix}.$$

We first see that, the MZF shall reuse the ZF estimates for the first and third layers based on (21). Then, we see that with $\boldsymbol{Q}$, the post-processing SNR (assuming the noise power equals 1) for the second bit-layer (which is identical to the fourth bit-layer) is increased from $1/\|\boldsymbol{\delta}_2\boldsymbol{H}^{+}\| = 185/47$ to $1/\|(\boldsymbol{\delta}_2+\boldsymbol{q}_2)\boldsymbol{H}^{+}\|^2 = 185/27$. Next we compute estimates with the MZF for the these two layers.

For the second layer, according to (13) we have

$$r_2 = (\boldsymbol{\delta}_2 + \boldsymbol{q}_2)\tilde{\boldsymbol{x}}_{\mathrm{ZF}} = \frac{1}{185}[0 \;\; 1 \;\; 2 \;\; 0] \begin{bmatrix} -60 \\ 309 \\ -730 \\ -107 \end{bmatrix} = \frac{-1151}{185},$$

and

$$z_2 = (r_2 \mod 4) - 2 = \frac{-7}{38}.$$

Similarly, for the fourth layer we have

$$r_4 = (\boldsymbol{\delta}_4 + \boldsymbol{q}_4)\tilde{\boldsymbol{x}}_{\mathrm{ZF}} = \frac{1}{185}[2 \;\; 0 \;\; 0 \;\; -1] \begin{bmatrix} -60 \\ 309 \\ -730 \\ -107 \end{bmatrix} = \frac{-13}{185},$$

and

$$z_4 = (r_4 \mod 4) - 2 = \frac{357}{185}.$$

As can be seen, the MZF corrects both detections for the second and the fourth layers where the ZF fails.

References

[1] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys & Tutorials,* vol. 17, no. 4, pp. 1941-1988, Sep. 2015.

[2] S. M. Kay, "Fundamentals of statistical signal processing, volume I: Estimation theory," Prentice Hall signal processing series, 1993.

[3] A. Ghasemmehdi and E. Agrell, "Faster recursions in sphere decoding," *IEEE Trans. Inf. Theory,* vol. 57, no. 6, pp. 3530-3536, Jun. 2011.

[4] J. Jaldén, B. Ottersten, "On the complexity of sphere decoding in digital communications", *IEEE Trans. Signal Process.,* vol. 53, no. 4, pp. 1474-1484, Apr. 2005.

[5] D. Wubben, R. Bohnke, V. Kuhn and K. Kammeyer, "Near-maximum-likelihood detection of MIMO systems using MMSE-based lattice reduction," IEEE Int. Conf. Commun. (ICC), Jun. 2004, pp. 798-802.

[6] A. K. Lenstra, H. W. Lenstra and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261 no. 4, pp. 515-534, Jul. 1982.

[7] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electron. Lett.,* vol. 7, no. 5, pp. 138-139, Mar. 1971.

[8] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Trans. Commun.,* vol. 20, no. 4, pp. 774-780, Aug. 1972.

[9] M. H. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 2, pp. 439-441, May 1983.

[10] J. Zhan, B. Nazer, U. Erez, and M. Gastpar, "Integer-forcing linear receivers," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7661-7685, Dec. 2014.

[11] S. H. Chae, M. Jang, and S.-K. Ah, "Multilevel coding scheme for integer-forcing MIMO receivers with binary codes", *IEEE Trans. Wireless. Commun.*, vol. 16, no. 8, pp. 5428-5441, Aug. 2017.

[12] O. Ordentlich and U. Erez, "Achieving the gains promised by Integer-Forcing equalization with binary codes", IEEE Conv. Elect. and Electron. Eng. in Israel (IEEEI), Nov. 2010, pp. 703-707.

[13] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227-5243, Sep. 2013.

[14] R. Böhnke, D. Wübben, V. Kühn, and K. D. Kammeyer, "Reduced Complexity MMSE Detection for BLAST Architectures", *Proc.* IEEE Global Commun. Conf. (GLOBECOM), Orlando, FL, USA, Dec. 2003, pp. 508-512.

[15] B. M. Hochwald, C. B. Peel and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication - Part II: Perturbation," *IEEE Trans. Commun.,* vol. 53, no. 3, pp. 537-544, Mar. 2005.

[16] J. Maurer, J. Jaldén, D. Seethaler and G. Matz, "Vector perturbation precoding revisited," *IEEE Trans. Signal Process.,* vol. 59, no. 1, pp. 315-328, Jan. 2011.

[17] E. Agrell, T. Eriksson, A. Vardy and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory,* vol. 48, no. 8, pp. 2201-2214, Aug. 2002.

[18] S. Hu and F. Rusek, "A soft-output MIMO detector with achievable information rate based partial marginalization," *IEEE Trans. Signal Process.,* vol. 65, no.6, pp. 1622-1637, Mar. 2017.

Paper v

A Generalized Zero-Forcing Precoder with Successive Dirty-Paper Coding in MISO Broadcast Channels

In this paper, we consider precoder designs for multiuser multi-input single-output (MISO) broadcasting channels. Instead of using a traditional linear zero-forcing (ZF) precoder, we propose a generalized ZF (GZF) precoder in conjunction with successive dirty-paper coding (DPC) for data-transmissions, namely, the GZF-DP precoder, where the suffix 'DP' stands for 'dirty-paper'. The GZF-DP precoder is designed to generate a band-shaped and lower-triangular effective channel $\boldsymbol{F}$ such that only the entries along the main diagonal and the ν first lower-diagonals can take non-zero values. Utilizing the successive DPC, the known non-causal inter-user interferences from the other (up to) ν users are canceled through successive encoding. We analyze optimal GZF-DP precoder designs both for sum-rate and minimum user-rate maximizations. Utilizing Lagrange multipliers, the optimal precoders for both cases are solved in closed-forms in relation to optimal power allocations. For the sum-rate maximization, the optimal power allocation can be found through water-filling, but with modified water-levels depending on the parameter ν. While for the minimum user-rate maximization that measures the quality of the service (QoS), the optimal power allocation is directly solved in closed-form which also depends on ν. Moreover, we propose two low-complexity user-ordering algorithms for the GZF-DP precoder designs for both maximizations, respectively. We show through numerical results that, the proposed GZF-DP precoder with a small ν (≤ 3) renders significant rate increments compared to the previous precoder designs such as the linear ZF and user-grouping based DPC (UG-DP) precoders.

S. Hu and F. Rusek,

"A generalized zero-forcing precoder with successive dirty-paper coding in MISO broadcast channels,"

IEEE Trans. Wireless Commun., vol. 16, no. 6, pp. 3632-3645, Jun. 2017.

1 Introduction

In the emerging Internet of things (IoT) [1] and device-to-device (D2D) [2] communication systems, a transmit node equipped with M transmit antennas may broadcast messages simultaneously to N low-cost receive nodes that are equipped with a single antenna. Under the assumption that the number of transmit antennas are much larger than the number of served users, i.e., $M \gg N$, which is known as massive multi-input multi-output (MIMO) systems [3], the multi-input single-output (MISO) broadcasting channels corresponding to different users that link the transmit and receive nodes are approximately orthogonal to each other. Consequently, the zero-forcing (ZF) precoders applied at the transmit nodes can efficiently eliminate the inter-user interference, and the MISO channels can be decomposed into a number of parallel and independent single-input single-output (SISO) channels in such cases.

In small-antenna systems such as small cells [4] with compact base-stations and WiFi systems, however, compared to the number of served users the number of transmit antennas are usually limited. Further, in current 3GPP standard [5], LTE-A systems support only up to 8 transmit antennas. Although future releases may support massive-MIMO or full-dimension MIMO (FD-MIMO)[6] and the number of transmit antennas at the eNode-B may increase to 64 for 2-D antenna array designs, the intended number of served users will also increase due to the vast connections featured in 5G systems. Consider the case where N is comparable to M, in order to fully eliminate the inter-user interference, the linear ZF precoder performs poorly due to the non-orthogonality of the MISO broadcast channel vectors [7]. Therefore, advanced precoder designs are required to improve the transmit power-efficiency and increase the rates of data-transmissions.

Some of the typical precoder designs are to preserve parts of the inter-user interference and mitigate them with the techniques of channel coding with side information (CCSI). CCSI has generated much research interests due to its applications in data hiding [8], precoding for interference channels [9], and transmitter cooperation in Ad-hoc networks[10]. Gelfand and Pinsker in [11] derive the capacity of a single-user memoryless channel with an additive interference signal $\boldsymbol{s}$ known to the transmitter, but not the receiver. Consider a received signal

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{s} + \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ are transmit and receive signals, and $\boldsymbol{z}$ is the unknown Gaussian noise, respectively. The capacity of model (1) is shown to equal

$$\mathcal{C} = \max_{p(\boldsymbol{u},\boldsymbol{x}|\boldsymbol{s})} \left\{ I(\boldsymbol{u};\boldsymbol{y}) - I(\boldsymbol{u};\boldsymbol{s}) \right\}. \tag{2}$$

where $\boldsymbol{u}$ is an auxiliary random variable and the maximum is taken over all joint probability distributions. Based on the result (2), Costa shows in [12] that with dirty-paper coding

(DPC), the channel capacity $\mathcal{C}$ is the same even if the interference $\boldsymbol{s}$ is not present. Utilizing the same principle, the DPC scheme can be extended to multi-user Gaussian vector broadcast channels[13], and DPC capacity regions have been derived via the uplink-downlink duality between broadcast channels and multiple-access channels [14, 15]. Practical DPC designs based on finite-alphabets have been extensively developed such as Tomlinson Harashima precoding [16], lattice precoding [17], and trellis coded quantization and modulation [18, 19].

Caire and Shamai in [9] propose a ZF based DPC (ZF-DP) design for MISO broadcast channels. They show that with successive DPC utilized at transmitter, the sum-rate of the ZF-DP precoder is close to the optimal DPC. In [35], the authors propose a successive ZF-DP (SZF-DP) precoding scheme and show that in the low SNR regime, the SZF-DP has similar performance as a successive ZF (SZF) precoder, where the SZF-DP and SZF precoders are direct extensions of the ZF-DP and linear ZF precoders in [9] for MIMO broadcast channels. In [36, 37] the authors further extend the ZF-DP and SZF-DP precoders subject to per-antenna power constraint (PAPC) instead of a sum-power constraint (SPC). Nevertheless, all the successive DPC based precoder designs in [9, 35–37] assume a full successive DPC scheme. As the number of users N increases, the successive DPC becomes prohibitive as it needs to consider the inter-user interference up to $N-1$ users. Recently, the authors in [20] propose a user-group based DPC precoder (UG-DP), which splits the N users into g disjoint groups with each group containing N_g users[1]. The inter-group interferences are eliminated by the precoder, while the intra-group interferences are canceled with successive DPC that is implemented on each user-group independently. With a small N_g, the DPC has less-complexity and is feasible [16–21]. However, as different user-groups are orthogonalized to each other, the UG-DP also suffers from rate-losses, especially when the channel vectors of different user-groups are spatially correlated.

In this work, we propose a generalized ZF precoder (GZF) design in conjunction with successive DPC, namely, the GZF-DP precoder, which unifies the designs of the UG-DP and the ZF-DP precoders. Instead of considering $N-1$ users in previous designs, we consider inter-user interference up to ν users, where the parameter ν is up to design and provides a trade-off between the rates and implementation complexity of the successive DPC[2]. By setting $\nu=0$, the GZF-DP precoder degrades to the linear ZF precoder, which has low complexity (no DPC is needed) but also low rates. On the other hand, with setting $\nu=N-1$, the GZF-DP precoder is identical to the ZF-DP precoder [9], which performs better than the other settings of ν but also has the highest DPC implementation complexity.

[1]For notational convenience, we assume that N is divisible by g and let $N_g=N/g$. But it can be straightforwardly modified to other cases with minor changes.

[2]Instead of using DPC at transmitter, in cooperative networks [22] the receiver nodes can implement successive interference cancellations (SIC) to achieve the same rates as the DPC. However, that requires a cost of communicating between the receive nodes. In which case, the parameter ν represents a maximal number of communication channels needed for the receive nodes.

Moreover, as the UG-DP precoder can be viewed as a special case of the GZF-DP precoder, it renders lower rates than the GZF-DP precoder with $\nu = N_g - 1$.

With the GZF-DP precoder, we consider two optimal designs: sum-rate maximization and minimum user-rate maximization, that are aiming to maximize the overall throughput and the quality of service (QoS), respectively. Using Lagrange multipliers, the optimal GZF-DP precoder designs for both cases are found in closed-form which depend on optimal power-allocations. For the sum-rate maximization, the optimal power allocation is found through a water-filling scheme in relation to modified water-levels introduced by preserving the inter-user interference up to ν users. While for the minimum user-rate maximization, the optimal power allocation can be solved directly in closed-form which also depends on ν. Moreover, we provide two low-complexity algorithms for optimal user-orderings for both maximizations, respectively. We show through numerical results that, the proposed GZF-DP precoder is superior to the previous ZF and UG-DP precoders, and most interestingly, with a small value of ν (≤ 3) the proposed GZF-DP precoder performs close to the ZF-DP precoder[14], i.e., the GZF-DP precoder with $\nu = N - 1$.

Notice that, as the precoder designs in [35–37] follow similar approaches as those in [9], the proposed GZF-DP precoder can also be extended to MIMO broadcast channels and PAPC constraint, which is a generalization of the SZF-DP precoder by only performing DPC up to ν multiple-receive-antenna users. However, as in [35–37] only the sum-rate maximization with a full DPC is considered, an interesting fact that the sum-rate maximization actually sacrifices the user-rates of some of the last users (corresponding to the last columns of channel matrix $\boldsymbol{H}$) compared to the linear ZF precoder is not shown. With the variable ν increasing from 0 to $N-1$, this property is clear shown in this work, which also motivates us to consider the minimum user-rate maximization for the proposed GZF-DP precoder.

The rest of the paper are organized as follows. In Sec. II, we briefly introduce the MISO system model and the previous precoder designs. In Sec. III, we elaborate the proposed GZF-DP precode designs in detail for sum-rate and minimum user-rate maximizations, respectively. We also analyze the low-complexity ordering algorithms for both maximization problems. Empirical results are provided in Sec. IV, and Sec. V summarizes the paper.

Notations:

Throughout this paper, superscripts $(\cdot)^{-1}$, $(\cdot)^{1/2}$, $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^\dagger$ stand for the inverse, matrix square root, complex conjugate, transpose, and Hermitian transpose, respectively. Boldface letters indicate vectors and boldface uppercase letters designate matrices. We also reserve $a_{m,n}$ to denote the element at the mth row and nth column of matrix $\boldsymbol{A}$, a_m to denote the mth element of vector $\boldsymbol{a}$, and $\boldsymbol{I}$ to represent the identity matrix. The operators $\mathcal{R}\{\cdot\}$ and $\mathrm{Tr}(\cdot)$ take the real part and the trace of the arguments, and $[\cdot]^+$ is the non-

negative protection. In addition, $\mathcal{J}_1 \setminus \mathcal{J}_2$ returns a set that contains all elements in set $\mathcal{J}_1$ that are not in $\mathcal{J}_2$, and the expressions $\boldsymbol{A} \succ \boldsymbol{B}$ and $\boldsymbol{A} \succeq \boldsymbol{B}$ represent that $(\boldsymbol{A}-\boldsymbol{B})$ is positive definite and semi-positive definite, respectively.

2 System Model and Previous Sum-rate Maximization Precoder Designs

Consider an MISO system with an M-antenna transmitter and N single-antenna users with assumption $M \geq N$. The channel vector from the transmitter to the nth user is denoted as $\boldsymbol{h}_n \in \mathbb{C}^{M\times 1}$, and the mth entry h_{mn} of $\boldsymbol{h}_n$ is the channel gain from the mth transmit antenna to the nth user. Denote the $N \times M$ channel

$$\boldsymbol{H} = [\boldsymbol{h}_1\ \boldsymbol{h}_2\ \ldots\ \boldsymbol{h}_N]^{\mathrm{T}}, \tag{3}$$

and let the $N \times 1$ vectors

$$\begin{aligned} \boldsymbol{y} &= [y_1\ y_2\ \ldots\ y_N]^{\mathrm{T}}, \\ \boldsymbol{x} &= [x_1\ x_2\ \ldots\ x_N]^{\mathrm{T}}, \\ \boldsymbol{z} &= [z_1\ z_2\ \ldots\ z_N]^{\mathrm{T}}, \end{aligned} \tag{4}$$

where x_n is the DPC-encoded symbol of the nth user that cancels the non-causal interference from the other users, and y_n, z_n is the received sample and the noise term corresponding to the nth user, respectively. With an $M \times N$ precoding matrix $\boldsymbol{P}$ applied at the transmitter, the received signals at the N autonomous users can be compactly written as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{z}, \tag{5}$$

where the noise term $\boldsymbol{z}$ comprises identical and independently distributed (IID) complex Gaussian variables with zero mean and a covariance matrix $N_0\boldsymbol{I}$. The transmit symbols x_n are uncorrelated due to DPC encoding and have unit-transmit power, that is, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\dagger}] = \boldsymbol{I}$. In addition, the transmit node is subject to a total transmit power constraint P_{T} such that

$$\mathrm{Tr}\left(\boldsymbol{P}\boldsymbol{P}^{\dagger}\right) \leq P_{\mathrm{T}}. \tag{6}$$

2.1 Optimal DPC Precoder

Denote the effective channel $\boldsymbol{F} = \boldsymbol{H}\boldsymbol{P}$, the interference channel corresponding to each of the N users from (5) can be written as

$$y_n = f_{n,n}x_n + \sum_{k=1}^{n-1} f_{n,k}x_k + \sum_{k=n+1}^{N} f_{n,k}x_k + z_n. \tag{7}$$

With a successive DPC [12] encoding scheme, the interference term $\sum_{k=1}^{n-1} f_{n,k}x_k$ is non-causally known and canceled, while the causal interference term $\sum_{k=n+1}^{N} f_{n,k}x_k$ is regarded as additive noise. Therefore, the optimal DPC precoder that maximizes the sum-rate is designed by solving the following problem

$$\begin{aligned} &\underset{\boldsymbol{F}}{\text{maximize}} \quad \sum_{n=1}^{N} \log\left(1+\frac{|f_{n,n}|^2}{N_0+\sum_{k=n+1}^{N}|f_{n,k}|^2}\right) \\ &\text{subject to} \quad (6). \end{aligned} \tag{8}$$

Directly optimizing (8) is computationally complex as it is a non-convex problem. In [23] the authors propose an iterative water-filling scheme to solve (8) based on the uplink-downlink duality. Although the optimal DPC precoder achieves the capacity region [24] of the multi-user MISO broadcast channels, the linear ZF precoder is widely used due to its simple implementation.

2.2 Linear ZF Precoder

The linear ZF precoder is set to

$$\boldsymbol{P} = \boldsymbol{H}^{\dagger}\left(\boldsymbol{H}\boldsymbol{H}^{\dagger}\right)^{-1}\boldsymbol{F}, \tag{9}$$

where $\boldsymbol{F}$ is an $N \times N$ diagonal matrix. With (9), the constraint (6) changes to

$$\mathrm{Tr}\left(\boldsymbol{F}^{\dagger}\left(\boldsymbol{H}\boldsymbol{H}^{\dagger}\right)^{-1}\boldsymbol{F}\right) \leq P_{\mathrm{T}}. \tag{10}$$

Denote $\boldsymbol{G} = \left(\boldsymbol{H}\boldsymbol{H}^{\dagger}\right)^{-1}$, the sum-rate maximization for linear ZF precoder is then formulated as

$$\begin{aligned} &\underset{f_{n,n}}{\text{maximize}} \quad R = \sum_{n=1}^{N} \log\left(1+\frac{|f_{n,n}|^2}{N_0}\right) \\ &\text{subject to} \quad \sum_{n=1}^{N} g_{n,n}|f_{n,n}|^2 \leq P_{\mathrm{T}}. \end{aligned} \tag{11}$$

The optimal power allocation is found through the water-filling scheme,

$$|f_{n,n}|^2 = N_0\left[\frac{1}{\lambda g_{n,n}} - 1\right]^{+}, \tag{12}$$

where $\lambda \geq 0$ is a constant such that power constraint (10) is satisfied. The optimal sum-rate reads

$$R^{\text{sum}} = \sum_{n=1}^{N} \left[-\log\left(\lambda g_{n,n}\right) \right]^{+}, \tag{13}$$

As the linear ZF precoder completely eliminates the inter-user interference, it results in low transmit power-efficiencies (even with regularizations[25]), especially when $\boldsymbol{H}$ is ill-conditioned. In [9], the authors propose a ZF-DP precoder that only nulls out the causal inter-user interference through ZF, and utilize successive DPC to cancel the non-causal interference.

2.3 ZF-DP Precoder

Assuming the channel decomposition $\boldsymbol{H} = \boldsymbol{R}\boldsymbol{U}$, where $\boldsymbol{R}$ is an $N \times N$ lower-triangular matrix and $\boldsymbol{U}$ is an $N \times M$ unitary matrix, the ZF-DP precoder is set to $\boldsymbol{P} = \boldsymbol{U}^{\dagger}\boldsymbol{B}$, and the $N \times N$ diagonal matrix $\boldsymbol{B}$ represents the power allocation whose nth diagonal element is b_n. The effective channel with the ZF-DP precoder equals $\boldsymbol{F} = \boldsymbol{R}\boldsymbol{B}$, and the received sample y_n reads

$$y_n = f_{n,n}x_n + \sum_{k=1}^{n-1} f_{n,k}x_k + z_n. \tag{14}$$

Through successive DPC encoding, the non-casual interference $\sum_{k=1}^{n-1} f_{n,k}x_k$ is nulled out for each of the users, and the sum-rate maximization problem can be formulated as

$$\begin{aligned} \underset{b_n}{\text{maximize}} \quad & \sum_{n=1}^{N} \log\left(1 + \frac{|b_n r_{n,n}|^2}{N_0}\right) \\ \text{subject to} \quad & \sum_{n=1}^{N} b_n^2 \leq P_{\text{T}}. \end{aligned} \tag{15}$$

The optimal power allocation b_n can also be found through standard water-filling. Although the ZF-DP precoder renders promising performance, the implementation of successive DPC becomes over complex when N is large. To reduce the DPC complexity, the authors in [20] propose a low-complexity UG-DP precoder.

2.4 UG-DP Precoder

We next briefly introduce the UG-DP precoder design. Assuming the same channel decomposition as with ZF-DP precoder, but now we constrain $\boldsymbol{R}$ to be block-diagonal, with each

block $\boldsymbol{R}_k$ ($1 \leq k \leq g$) being an $N_g \times N_g$ lower-triangular matrix. Let the $N_g \times M$ sub-matrix $\boldsymbol{H}_k$ comprise the row vectors in $\boldsymbol{H}$ corresponding to the users in the kth group, and the $(N-N_g) \times M$ sub-matrix $\bar{\boldsymbol{H}}_k$ comprise the remaining row vectors. With decomposition $\boldsymbol{U} = [\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_g]^\dagger$, each $M \times N_g$ component $\boldsymbol{U}_k$ can be obtained through

$$\boldsymbol{H}_k \left(\boldsymbol{I} - \bar{\boldsymbol{H}}_k^\dagger \left(\bar{\boldsymbol{H}}_k \bar{\boldsymbol{H}}_k^\dagger \right)^{-1} \bar{\boldsymbol{H}}_k \right) = \boldsymbol{R}_k \boldsymbol{U}_k^\dagger. \tag{16}$$

Then, with the matrix $\boldsymbol{U}_k$ calculated via (16), the optimal $\boldsymbol{P}$ equals $\boldsymbol{P} = \boldsymbol{U}^\dagger \boldsymbol{B}$ and the effective channel becomes $\boldsymbol{F} = \boldsymbol{R}\boldsymbol{B}$, where the diagonal matrix $\boldsymbol{B}$ represents the power allocation to different users. Then, the remaining processes follow the ZF-DP precoder design. Although the UG-DP precoder reduces the complexity of DPC by user-grouping, it also suffers from rate-losses from the orthogonalization of different user-groups. In order to increase the rates of the UG-DP precoder while keeping a similar complexity, we can extend the block-diagonal lower-triangular $\boldsymbol{R}$ and $\boldsymbol{F}$ to be band-shaped matrices. That is, the connections among different user-groups are preserved such that, only the elements along the main diagonal and the first $N_g - 1$ lower-diagonals of $\boldsymbol{R}$ and $\boldsymbol{F}$ can take non-zero values. The proposed GZF-DP precoder design is based on such a principle and is explained in detail next.

3 Optimal Designs of the Proposed GZF-DP Precoder

Instead of assuming $\boldsymbol{F}$ to be diagonal or block-diagonal such as in previous designs, we let $\boldsymbol{F}$ to be a band-shaped and lower-triangular for the GZF-DP precoder design,

$$\boldsymbol{F} = \begin{bmatrix} f_{1,1} & & & & & & \\ f_{2,1} & f_{2,2} & & & & & \\ \vdots & f_{3,2} & \ddots & & & & \\ f_{\nu+1,1} & \vdots & \ddots & \ddots & & & \\ & f_{\nu+2,2} & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & f_{N,N-\nu} & \cdots & f_{N,N-1} & f_{N,N} \end{bmatrix}. \tag{17}$$

The parameter ν denotes the interfering depth of the effective MISO broadcasting channels. For simpler descriptions, we define two operations as

$$\begin{aligned} n \ominus \nu &= \max(n - \nu, 0), \\ n \boxplus \nu &= \min(n + \nu, N). \end{aligned} \tag{18}$$

The GZF-DP precoder generalizes the linear ZF precoder in the sense that ν can be set larger than 0. Under the case $\nu = 0$, the GZF-DP precoder degrades to the linear ZF precoder and no DPC is needed. With $\boldsymbol{F}$ defined in (17), the received sample y_n of the nth user reads

$$y_n = f_{n,n}x_n + \sum_{k=n\ominus\nu}^{n-1} f_{n,k}x_k + z_n. \tag{19}$$

As the interference $\sum_{k=n\ominus\nu}^{n-1} f_{n,k}x_k$ is non-causally known at the transmit node, we can apply the same successive DPC encoding as the ZF-DP precoder[9] to cancel it. That is, we first encode a first user that suffers no interference from the other users after precoding. Then, the second user is encoded utilizing DPC scheme with regarding the encoded symbols from the first user as known interference. The remaining users are successively encoded in the same manner. For each of the N users, as there are at most ν users to be considered in the DPC and $\nu \ll N-1$, the GZF-DP precoder renders much lower-complexity of the successive DPC operations than the ZF-DP precoder and has similar complexity as the UG-DP precoder with $\nu = N_g - 1$.

Before deriving the optimal GZF-DP precoder designs, we make some useful notations. Denote the $\nu\times 1$ vectors that comprise the non-zero entries on each column of $\boldsymbol{F}$ excluding the main diagonal element as

$$\boldsymbol{f}_n^{\nu} = [f_{n+1,n}\ , f_{n+2,n}\ , \cdots\ , f_{n\boxplus\nu,n}]^{\mathrm{T}}. \tag{20}$$

Moreover, define the $(\nu+1)\times(\nu+1)$ principle sub-matrix $\boldsymbol{G}_n^{\nu}$ obtained from $\boldsymbol{G}$ as

$$\boldsymbol{G}_n^{\nu} = \begin{bmatrix} g_{n,n} & g_{n,n+1} & \cdots & g_{n,n\boxplus\nu} \\ g_{n+1,n} & g_{n+2,n+1} & \cdots & g_{n+1,n\boxplus\nu} \\ \vdots & \vdots & \vdots & \vdots \\ g_{n\boxplus\nu,n} & g_{n\boxplus\nu,n+1} & \cdots & g_{n\boxplus\nu,n\boxplus\nu} \end{bmatrix}. \tag{21}$$

and let

$$\boldsymbol{g}_n^{\nu} \quad = \quad [g_{n,n+1}\ , g_{n,n+2}\ , \cdots\ , g_{n,n\boxplus\nu}]^{\dagger}\ . \tag{22}$$

Then, $\boldsymbol{G}_{n+1}^{\nu-1}$ is the $\nu\times\nu$ principle sub-matrix obtained by further removing the first row and column vectors from $\boldsymbol{G}_n^{\nu}$.

3.1 Sum-rate Maximization

We first consider the GZF-DP precoder design for the sum-rate maximization subject to the transmit power constraint (10). The problem can be formulated as

$$\begin{aligned} \underset{\boldsymbol{F}}{\text{maximize}} \ & \sum_{n=1}^{N} \log\left(1+\frac{|f_{n,n}|^2}{N_0}\right) \\ \text{subject to} \ & \operatorname{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) = P_{\mathrm{T}}. \end{aligned} \tag{23}$$

Note that, we have changed the power constraint in (23) from $\operatorname{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) \leq P_{\mathrm{T}}$ to $\operatorname{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) = P_{\mathrm{T}}$. The reason is that, for a solution of (23) the equality of the power constraint always holds. This is so, since if $\operatorname{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) < P_{\mathrm{T}}$ holds, we can scale up $\boldsymbol{F}$ to be some $\tilde{\boldsymbol{F}} = \alpha\boldsymbol{F}$ ($\alpha > 1$) such that $\operatorname{Tr}\left(\tilde{\boldsymbol{F}}^{\dagger}\boldsymbol{G}\tilde{\boldsymbol{F}}\right) = P_{\mathrm{T}}$ holds, and with $\tilde{\boldsymbol{F}}$ the sum-rate in (23) is also increased. By constraining $f_{n,n} \geq 0$, the optimal solution for (23) is stated in Theorem 1.

Theorem 1. *The optimal band-shaped and low-triangular matrix $\boldsymbol{F}$ as defined in (17) for sum-rate maximization (23) satisfies the following conditions*

$$\boldsymbol{f}_n^{\nu} = -f_{n,n}\left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1}\boldsymbol{g}_n^{\nu}, \tag{24}$$

$$f_{n,n} = \sqrt{N_0\left[\frac{1}{\lambda\hat{g}_n^{\nu}} - 1\right]^{+}}, \tag{25}$$

where

$$\hat{g}_n^{\nu} = g_{n,n} - \left(\boldsymbol{g}_n^{\nu}\right)^{\dagger}\left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1}\boldsymbol{g}_n^{\nu}, \tag{26}$$

and $\lambda > 0$ is a constant such that the transmit power constraint is satisfied.

Proof. Consider the Lagrangian function

$$\mathcal{L} = \sum_{n=1}^{N} \log\left(1+\frac{|f_{n,n}|^2}{N_0}\right) - \lambda\left(\operatorname{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) - P_{\mathrm{T}}\right), \tag{27}$$

where λ is the Lagrange multiplier. The necessary conditions [26, 27] for the optimal solution are

$$\left.\begin{aligned} \frac{\partial\mathcal{L}}{\partial f_{n,k}} = 0, \ 1 \leq n,k \leq N \\ \operatorname{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) - P_{\mathrm{T}} = 0 \\ \lambda \geq 0 \end{aligned}\right\}. \tag{28}$$

Note that, with the definitions in (20)-(22), the trace term in (27) can be rewritten as

$$\mathrm{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right)=\sum_{n=1}^{N}\left[f_{n,n}\ \ (\boldsymbol{f}_n^{\nu})^{\dagger}\right]\left[\begin{array}{cc} g_{n,n} & (\boldsymbol{g}_n^{\nu})^{\dagger} \\ \boldsymbol{g}_n^{\nu} & \boldsymbol{G}_{n+1}^{\nu-1}\end{array}\right]\left[\begin{array}{c} f_{n,n} \\ \boldsymbol{f}_n^{\nu}\end{array}\right]. \tag{29}$$

Taking the first-order derivatives of $\mathcal{L}$ with respect to $f_{n,n}$ and $\boldsymbol{f}_n^{\nu}$, and using (29) results in

$$\frac{\partial\mathcal{L}}{\partial f_{n,n}} = \frac{N_0 f_{n,n}}{N_0+|f_{n,n}|^2}-\lambda\left(f_{n,n}g_{n,n}+(\boldsymbol{f}_n^{\nu})^{\dagger}\,\boldsymbol{g}_n^{\nu}\right), \tag{30}$$

$$\nabla_{\boldsymbol{f}_n^{\nu}}\mathcal{L} = -\lambda\left(f_{n,n}\,(\boldsymbol{g}_n^{\nu})^{\dagger}+(\boldsymbol{f}_n^{\nu})^{\dagger}\,\boldsymbol{G}_{n+1}^{\nu-1}\right)^{\mathrm{T}}. \tag{31}$$

Then, by setting $\nabla_{\boldsymbol{f}_n^{\nu}}\mathcal{L}$ in (31) to zero, the vector $\boldsymbol{f}_n^{\nu}$ can be solved for, and the result is given in (24). Inserting (24) back into (30) and setting $\partial\mathcal{L}/\partial f_{n,n}$ to zero, we obtain

$$\frac{N_0}{N_0+|f_{n,n}|^2}=\lambda\left(g_{n,n}-(\boldsymbol{g}_n^{\nu})^{\dagger}\left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1}\boldsymbol{g}_n^{\nu}\right). \tag{32}$$

From (32) it holds that $\lambda>0$ as $N_0>0$, since $\hat{g}_n^{\nu}>0$ which will be shown later in Property 1. Using (26), the optimal $f_{n,n}$ reads

$$|f_{n,n}|^2=N_0\left[\frac{1}{\lambda\hat{g}_n^{\nu}}-1\right]^{+}. \tag{33}$$

As we constrain $f_{n,n}$ to be positive, the solution of $f_{n,n}$ is in (25), which completes the proof. □

With the necessary conditions of $\boldsymbol{f}_n^{\nu}$ and $f_{n,n}$ stated in Theorem 1, the constraint in (23) can be written as

$$\begin{aligned}\frac{1}{N_0}\mathrm{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) &= \frac{1}{N_0}\sum_{n=1}^{N}\hat{g}_n^{\nu}|f_{n,n}|^2 \\ &= \sum_{n=1}^{N}\left[\frac{1}{\lambda}-\hat{g}_n^{\nu}\right]^{+}=\frac{P_{\mathrm{T}}}{N_0}.\end{aligned} \tag{34}$$

and the sum-rate equals

$$R^{\mathrm{sum}}=\sum_{n=1}^{N}R_n^{\mathrm{user}}, \tag{35}$$

where

$$R_n^{\mathrm{user}}=\left[-\log\left(\lambda\hat{g}_n^{\nu}\right)\right]^{+}. \tag{36}$$

Therefore, to find the optimal solution for (23) is equivalent to find an optimal water-level $1/\lambda$ such that (35) is maximized and (34) is satisfied, which can be efficiently solved using water-filling scheme[28]. Comparing (33) with (12), with the GZF-DP precoder a similar water-filling scheme still applies, however, the water-level has changed as $g_{n,n}$ is replaced now by $\hat{g}_n^\nu$, due to the preserved inter-user interference. We state a property below that shows that $\hat{g}_n^\nu$ is positive and non-increasing in ν for all $1 \le n \le N$.

Property 1. *Under the condition that $\boldsymbol{H}$ has full row rank, for $1 \le n \le N$, it holds that*

$$0 < \hat{g}_n^{N-1} \le \hat{g}_n^{N-2} \le \cdots \le \hat{g}_n^1 \le g_{n,n}. \tag{37}$$

Proof. First we show that for $1 \le \nu \le N-1$, $0 < \hat{g}_n^\nu \le g_{n,n}$ holds. Since $\boldsymbol{H}$ has full row rank, $\boldsymbol{G} \succ \boldsymbol{0}$. Consequently, $\boldsymbol{G}_{n+1}^{\nu-1}$ and $\boldsymbol{G}_n^\nu$ are also positive-definite as principle sub-matrices of $\boldsymbol{G}$. Hence, $(\boldsymbol{g}_n^\nu)^\dagger \left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1} \boldsymbol{g}_n^\nu \ge 0$, and $\hat{g}_n^\nu \le g_{n,n}$ follows from (26). On the other hand, from the definition, $\boldsymbol{G}_n^\nu$ equals

$$\boldsymbol{G}_n^\nu = \begin{bmatrix} g_{n,n} & (\boldsymbol{g}_n^\nu)^\dagger \\ \boldsymbol{g}_n^\nu & \boldsymbol{G}_{n+1}^{\nu-1} \end{bmatrix}. \tag{38}$$

Hence, $\hat{g}_n^\nu$ is the Schur-complement[29] of $g_{n,n}$, and by utilizing the matrix-inversion lemma [30], the inverse $(\boldsymbol{G}_n^\nu)^{-1} \succ \boldsymbol{0}$ is in (40), which shows that $\hat{g}_n^\nu > 0$.

Next we show that, $\hat{g}_n^\nu \le \hat{g}_n^{\nu-1}$ holds for $1 \le n \le N$. Firstly, for $n > N-\nu$, by definitions (22) and (21), the equalities $\boldsymbol{g}_n^\nu = \boldsymbol{g}_n^{\nu-1}$ and $\boldsymbol{G}_{n+1}^{\nu-1} = \boldsymbol{G}_{n+1}^{\nu-2}$ hold. Hence, from (26), $\hat{g}_n^\nu = \hat{g}_n^{\nu-1}$ holds. Secondly, for $1 \le n \le N-\nu$, $\boldsymbol{G}_n$ in (38) can also be rewritten as

$$\boldsymbol{G}_n^\nu = \begin{bmatrix} \boldsymbol{G}_n^{\nu-1} & (\tilde{\boldsymbol{g}}_n^\nu)^\dagger \\ \tilde{\boldsymbol{g}}_n^\nu & g_{n+\nu,n+\nu} \end{bmatrix}, \tag{39}$$

where $\tilde{\boldsymbol{g}}_n^\nu = [g_{n+\nu,n},\ g_{n+\nu,n+1},\ \cdots,\ g_{n+\nu,n+\nu-1}]$. By utilizing the matrix-inversion lemma again, the inverse $(\boldsymbol{G}_n^\nu)^{-1}$ can also be written in (41). From (40) we know that, $\left(\hat{g}_n^{\nu-1}\right)^{-1}$ is the first diagonal element of $\left(\boldsymbol{G}_n^{\nu-1}\right)^{-1}$, while $(\hat{g}_n^\nu)^{-1}$ is the first diagonal element of $(\boldsymbol{G}_n^\nu)^{-1}$ and hence, the first diagonal element of $\left(\boldsymbol{G}_n^{\nu-1} - \frac{(\tilde{\boldsymbol{g}}_n^\nu)^\dagger \tilde{\boldsymbol{g}}_n^\nu}{g_{n+\nu,n+\nu}}\right)^{-1}$ from (41). Using the Woodbury matrix identity[30], $\left(\boldsymbol{G}_n^{\nu-1} - \frac{(\tilde{\boldsymbol{g}}_n^\nu)^\dagger \tilde{\boldsymbol{g}}_n^\nu}{g_{n+\nu,n+\nu}}\right)^{-1} \succeq \left(\boldsymbol{G}_n^{\nu-1}\right)^{-1}$ holds. Therefore, $(\hat{g}_n^\nu)^{-1} \ge \left(\hat{g}_n^{\nu-1}\right)^{-1}$ holds, and $\hat{g}_n^\nu \le \hat{g}_n^{\nu-1}$ follows, which completes the proof. □

As $\hat{g}_n^\nu \le g_{n,n}$, from (34) in general the water-level $1/\lambda$ is actually non-increasing when ν increases. Therefore, not all the user-rates are increased with a larger ν. For instance, for the last user, as $\hat{g}_N^\nu = g_{N,N}$ for all ν, the user-rate R_N^{user} is non-increasing as ν increases. In general, we have the following corollary.

$$(\boldsymbol{G}_n^{\nu})^{-1} = \begin{bmatrix} (\hat{g}_n^{\nu})^{-1} & -(\hat{g}_n^{\nu})^{-1} (\boldsymbol{g}_n^{\nu})^{\dagger} \left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1} \\ -(\hat{g}_n^{\nu})^{-1} \left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1} \boldsymbol{g}_n^{\nu} & \left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1} + (\hat{g}_n^{\nu})^{-1} \left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1} \boldsymbol{g}_n^{\nu} (\boldsymbol{g}_n^{\nu})^{\dagger} \left(\boldsymbol{G}_{n+1}^{\nu-1}\right)^{-1} \end{bmatrix}. \tag{40}$$

$$(\boldsymbol{G}_n^{\nu})^{-1} = \begin{bmatrix} \left(\boldsymbol{G}_n^{\nu-1} - \frac{(\tilde{\boldsymbol{g}}_n^{\nu})^{\dagger} \tilde{\boldsymbol{g}}_n^{\nu}}{g_{n+\nu,n+\nu}}\right)^{-1} & -\left(\boldsymbol{G}_n^{\nu-1} - \frac{(\tilde{\boldsymbol{g}}_n^{\nu})^{\dagger} \tilde{\boldsymbol{g}}_n^{\nu}}{g_{n+\nu,n+\nu}}\right)^{-1} \frac{(\tilde{\boldsymbol{g}}_n^{\nu})^{\dagger}}{g_{n+\nu,n+\nu}} \\ -\frac{-\tilde{\boldsymbol{g}}_n^{\nu}}{g_{n+\nu,n+\nu}} \left(\boldsymbol{G}_n^{\nu-1} - \frac{(\tilde{\boldsymbol{g}}_n^{\nu})^{\dagger} \tilde{\boldsymbol{g}}_n^{\nu}}{g_{n+\nu,n+\nu}}\right)^{-1} & \frac{1}{g_{n+\nu,n+\nu}} + \frac{\tilde{\boldsymbol{g}}_n^{\nu}}{g_{n+\nu,n+\nu}^2} \left(\boldsymbol{G}_n^{\nu-1} - \frac{(\tilde{\boldsymbol{g}}_n^{\nu})^{\dagger} \tilde{\boldsymbol{g}}_n^{\nu}}{g_{n+\nu,n+\nu}}\right)^{-1} (\tilde{\boldsymbol{g}}_n^{\nu})^{\dagger} \end{bmatrix}. \tag{41}$$

Corollary 1. *If ν is increased from ν_1 to ν_1+1 for the GZF-DP precoder, as for $n \geq N-\nu_1$, $\hat{g}_n^{\nu} = \hat{g}_n^{\nu-1}$ holds, and as a result of the non-increasing water-level, the user-rates of the last ν_1+1 users are also non-increasing.*

However, the sum-rate never decrease with a larger ν, which is stated in the below property.

Property 2. *If $\nu_2 > \nu_1$, the sum-rate R^{sum} obtained with the GZF-DP precoder with $\nu = \nu_2$ is no less than that obtained with $\nu = \nu_1$. However, under the case that the channel $\boldsymbol{H}$ itself is band-shaped with only the elements along the main diagonal and the first ν_1 lower-diagonals can take non-zero values, increasing ν to be larger than ν_1 will not further increase R^{sum}.*

Proof. The first statement holds from the fact that the effective channel $\boldsymbol{F}$ with $\nu = \nu_1$ is a subset of $\boldsymbol{F}$ with $\nu = \nu_2$. Next we prove the second statement by showing that $\hat{g}_n^{\nu} = \hat{g}_n^{\nu_1}$ for any n and $\nu > \nu_1$, under the condition that $\boldsymbol{H}$ is band-shaped with only the elements along the main diagonal and the first ν_1 lower-diagonals can take non-zero values. Therefore, in such a case, the sum-rate R^{sum} obtained with $\nu > \nu_1$ is equal to R^{sum} with $\nu = \nu_1$.

We first show that, for $n=1$, $\hat{g}_1^{\nu} = \hat{g}_1^{\nu_1}$ holds for $\nu > \nu_1$. We decompose $\boldsymbol{G}$ and $\boldsymbol{H}\boldsymbol{H}^{\dagger}$ into block forms as

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_1^{\nu} & \boldsymbol{G}_2^{\dagger} \\ \boldsymbol{G}_2 & \boldsymbol{G}_3 \end{bmatrix}, \quad \boldsymbol{H}\boldsymbol{H}^{\dagger} = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_2^{\dagger} \\ \boldsymbol{B}_2 & \boldsymbol{B}_3 \end{bmatrix}, \tag{42}$$

where sub-matrix $\boldsymbol{G}_1^{\nu}$ follows the definition in (21) and sub-matrices $\boldsymbol{G}_2$, $\boldsymbol{G}_3$ are deduced from $\boldsymbol{G}_1^{\nu}$. Similarly, sub-matrix $\boldsymbol{B}_1$ has the same size as $\boldsymbol{G}_1^{\nu}$, and sub-matrices $\boldsymbol{B}_2$, $\boldsymbol{B}_3$ are deduced from $\boldsymbol{B}_1$. As $\boldsymbol{G} = \left(\boldsymbol{H}\boldsymbol{H}^{\dagger}\right)^{-1}$, following the matrix inversion lemma we have

$$(\boldsymbol{G}_1^{\nu})^{-1} = \boldsymbol{B}_1 - \boldsymbol{B}_2^{\dagger}\boldsymbol{B}_3^{-1}\boldsymbol{B}_2. \tag{43}$$

As $\boldsymbol{H}$ is band-shaped, when $\nu \geq \nu_1$, the first row vector in $\boldsymbol{B}_2^{\dagger}$ comprises all zero elements. Consequently, from (43) the first diagonal element of $(\boldsymbol{G}_1^{\nu})^{-1}$, which is $(\hat{g}_1^{\nu})^{-1}$, is equal to the first diagonal element of $\boldsymbol{B}_1$. Hence, we have

$$\hat{g}_1^{\nu} = |h_1(1)|^{-2}, \quad \nu \geq \nu_1, \tag{44}$$

where $h_1(1)$ is the first tap of the channel vector corresponding to the first user,

For $n > 1$, we can permute the principle sub-matrix $\boldsymbol{G}_n^{\nu}$ to the upper-left corner with a permutation matrix $\boldsymbol{Q}$ such that,

$$\boldsymbol{Q}\boldsymbol{G}\boldsymbol{Q}^{\dagger} = \begin{bmatrix} \boldsymbol{G}_n^{\nu} & \tilde{\boldsymbol{G}}_2^{\dagger} \\ \tilde{\boldsymbol{G}}_2 & \tilde{\boldsymbol{G}}_3 \end{bmatrix}, \tag{45}$$

where $\tilde{\boldsymbol{G}}_2$, $\tilde{\boldsymbol{G}}_3$ are deduced from $\boldsymbol{G}_n^\nu$. We also permute $\boldsymbol{H}\boldsymbol{H}^\dagger$ accordingly such that

$$\boldsymbol{Q}\boldsymbol{H}\boldsymbol{H}^\dagger\boldsymbol{Q}^\dagger = \begin{bmatrix} \tilde{\boldsymbol{B}}_1 & \tilde{\boldsymbol{B}}_2^\dagger \\ \tilde{\boldsymbol{B}}_2 & \tilde{\boldsymbol{B}}_3 \end{bmatrix}, \tag{46}$$

where sub-matrices $\tilde{\boldsymbol{B}}_1, \tilde{\boldsymbol{B}}_2, \tilde{\boldsymbol{B}}_3$ are defined similarly as before.

As $\boldsymbol{Q}\boldsymbol{G}\boldsymbol{Q}^\dagger = \left(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{H}^\dagger\boldsymbol{Q}^\dagger\right)^{-1}$ holds, following (43) and (44) we have

$$\hat{g}_n^\nu = |h_n(n)|^{-2}, \quad \nu \geq \nu_1, \tag{47}$$

where $h_n(n)$ is the nth tap of the channel vector corresponding to the nth user, which is transfered to be the first user after permutation. Therefore, with $\hat{g}_n^\nu$ given in (47), it holds that, $\hat{g}_n^\nu = \hat{g}_n^{\nu_1}$ for any n and $\nu > \nu_1$, which completes the proof. □

Property 2 reveals that if $\boldsymbol{H}$ is banded, further increasing the band-size of $\boldsymbol{F}$ to be larger than the band-size of $\boldsymbol{H}$ will not increase the sum-rate. Moreover, for a band-shaped $\boldsymbol{H}$, $\hat{g}_n^\nu$ can be easily calculated through (47) for $\nu \geq \nu_1$. Next, we show that the GZF-DP precoder design actually provides a unified framework of the previous ZF based precoder designs.

Corollary 2. *With $\nu = 0$, the GZF-DP precoder becomes the linear ZF precoder without DPC; while with $\nu = N-1$, the GZF-DP precoder is identical to the ZF-DP precoder. In addition, the UG-DP precoder is inferior to the GZF-DP precoder with $\nu = N_g - 1$.*

Proof. When $\nu = 0$, $\hat{g}_n^\nu = g_{n,n}$ for all n, and the GZF-DP precoder is thusly identical to the linear ZF precoder. On the other hand, when $\nu = N-1$, the maximization (23) can be formulated as the same problem in (15), which shows the trade-off between the sum-rate and the complexity of successive DPC. Moreover, as the UG-DP can be reviewed as a special case of the GZF-DP with $\nu = N_g - 1$, the UG-DP precoder is inferior to the GZF-DP precoder in general. □

Although the ZF-DP precoder provides the highest sum-rate, as shown in Corollary 1, it sacrifices user-rates of some of the last users. As a generalization of the ZF-DP precoder, the GZF-DP precoder, however, can provide a trade-off between the sum-rate increment and the user-rate decrement through the parameter ν. Below we illustrate with an example to show different designs of the linear ZF precoder, the UG-DP precoder, and the proposed GZF-DP precoder.

Example 1. *Assume $N_0 = 1$, $P_{\mathrm{T}} = 10$ dB, and consider an MISO channel with 4 transmit antennas and 4 single-antenna users as ($i = \sqrt{-1}$)*

$$\boldsymbol{H} = \begin{bmatrix} 1+4i & 4+3i & 2+3i & 3+3i \\ 4+1i & 1+4i & 1+1i & 2+4i \\ 2+3i & 1+4i & 3+3i & 4+3i \\ 4+4i & 2+3i & 1+4i & 2+2i \end{bmatrix}.$$

The sum-rates (bits/channel use) of the ZF precoder, the UG-DP precoder with $N_{\mathrm{g}} = 2$, and the GZF-DP precoder with $\nu = 1$ are equal to

$$\begin{aligned} R_{\mathrm{ZF}}^{\mathrm{sum}} &= 17.885, \\ R_{\mathrm{UG-DP}}^{\mathrm{sum}} &= 18.206, \\ R_{\mathrm{GZF-DP},\,\nu=1}^{\mathrm{sum}} &= 18.514, \end{aligned}$$

respectively. The optimal effective channels $\boldsymbol{F}$ are listed at the bottom of this page.

With Example 1, the user-rates corresponding to different precoders are equal to

$$\begin{aligned} R_{\mathrm{ZF}}^{\mathrm{user}} &= [4.333,\ 4.830,\ 4.370,\ 4.352], \\ R_{\mathrm{GZF-DP},\,\nu=1}^{\mathrm{user}} &= [4.650,\ 5.106,\ 4.410,\ 4.348], \\ R_{\mathrm{GZF-DP},\,\nu=2}^{\mathrm{user}} &= [5.394,\ 6.047,\ 4.387,\ 4.324]. \end{aligned}$$

As it can been seen, although the sum-rate is increased from $\nu = 0$ to 1, the user-rate of the last user is decreased. Further, from $\nu = 1$ to 2, the user-rates of the last two users are also decreased, which are aligned with Corollary 1. Especially for the last user, the user-rate is continuously decreasing when ν increases from 0 to 2. Therefore, instead of maximizing

---

$$\boldsymbol{F}_{\mathrm{ZF}} = \begin{bmatrix} 4.376 & & & \\ & 5.238 & & \\ & & 4.436 & \\ & & & 4.407 \end{bmatrix},$$

$$\boldsymbol{F}_{\mathrm{UG-DP}} = \begin{bmatrix} 4.899 & 0 & & \\ -1.140+2.340i & -5.217 & & \\ & & 4.490 & 0 \\ & & 0.489+0.607i & 4.389 \end{bmatrix},$$

$$\boldsymbol{F}_{\mathrm{GZF-DP},\,\nu=1} = \begin{bmatrix} 4.910 & 0 & & \\ -1.143+2.345i & 5.784 & & \\ & 2.034+0.416i & 4.501 & 0 \\ & & 0.490+0.609i & 4.400 \end{bmatrix}.$$

the sum-rate, it is also meaningful to consider maximizations of user-rate, which is usually used as a measurement for the fairness of the QoS.

Next we discuss the minimum user-rate maximization with the proposed GZF-DP precoder.

3.2 Minimum User-rate Maximization

For minimum user-rate maximization, the design of the GZF-DP precoder is formulated as

$$\begin{aligned} \underset{\boldsymbol{F},\, R^{\text{user}}}{\text{maximize}} \;\; & R^{\text{user}} \\ \text{subject to} \;\; & R^{\text{user}} \leq \log\left(1+\frac{|f_{n,n}|^2}{N_0}\right), 1 \leq n \leq N \\ & \text{Tr}\left(\boldsymbol{F}^{\dagger}\boldsymbol{G}\boldsymbol{F}\right) \leq P_{\text{T}}, \end{aligned} \tag{48}$$

where the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ are the same as defined for the sum-rate maximization and we constrain $f_{n,n} \geq 0$. Following similar arguments as for the sum-rate maximization, it also holds that the equality in the power constraint always holds for an optimal solution of (48). Furthermore, we have the below lemma.

Lemma 1. *For an optimal solution $\boldsymbol{F}$ of (48), it holds that $R^{\text{user}} = \log\left(1+\frac{|f_{n,n}|^2}{N_0}\right)$ for all n.*

Proof. For an optimal solution $\boldsymbol{F}$ of (48), we denote the maximal and minimal user-rates as $R^{\text{user}}_{n_1}$ and $R^{\text{user}}_{n_2}$, respectively, which equal

$$R^{\text{user}}_{n_i} = \log\left(1+\frac{|f_{n_i,n_i}|^2}{N_0}\right), \quad i=1,2. \tag{49}$$

Then, the minimum user-rate is equal to $R^{\text{user}}_{n_2}$. We further denote the transmit powers of user n_1 and n_2 as P_1 and P_2, respectively. According to (29), it holds that

$$P_i = g_{n_i,n_i}|f_{n_i,n_i}|^2 + 2\mathcal{R}\left\{\left(\boldsymbol{f}^{\nu}_{n_i}\right)^{\dagger}\boldsymbol{g}^{\nu}_{n_i}f_{n_i,n_i}\right\} + \left(\boldsymbol{f}^{\nu}_{n_i}\right)^{\dagger}\boldsymbol{G}^{\nu-1}_{n_i+1}\boldsymbol{f}^{\nu}_{n_i}, \;\; i=1,2. \tag{50}$$

Now let's assume $R_{n_1} > R_{n_2}$, that is, the maximal user-rate is strictly larger than the minimal user-rate. Then, we can scale f_{n_i,n_i} and $\boldsymbol{f}^{\nu}_{n_i}$ to be $\tilde{f}_{n_i,n_i} = \alpha_i f_{n_i,n_i}$ and $\tilde{\boldsymbol{f}}^{\nu}_{n_i} = \alpha_i \boldsymbol{f}^{\nu}_{n_i}$, respectively, where $\alpha_2 > 1 > \alpha_1$ and

$$\alpha_1 = \sqrt{1+\frac{(1-\alpha_2^2)P_2}{P_1}}.$$

Note that, according to (50), with such a scaling operation, the total transmit power of user n_1 and n_2 remains the same, that is, $\alpha_1^2 P_1 + \alpha_2^2 P_2 = P_1 + P_2$. However, according to (49), the user-rate with such a scaling increases $R_{n_2}^{\text{user}}$ and decreased $R_{n_1}^{\text{user}}$. Hence, the minimum user-rate can therefore be increased, which contradicts to the assumption that $\boldsymbol{F}$ is optimal. Therefore, for an optimal $\boldsymbol{F}$, $R_{n_1}^{\text{user}} = R_{n_2}^{\text{user}}$ holds, which shows that all user-rates are equal to each other for an optimal $\boldsymbol{F}$ of (48). □

With the above arguments, we can change (48) to the equivalent problem

$$\begin{aligned} &\underset{\boldsymbol{F},\, R^{\text{user}}}{\text{maximize}} \;\; R^{\text{user}} \\ &\text{subject to} \;\; R^{\text{user}} = \log\left(1 + \frac{|f_{n,n}|^2}{N_0}\right), 1 \le n \le N \\ &\qquad\qquad \text{Tr}\left(\boldsymbol{F}^\dagger \boldsymbol{G} \boldsymbol{F}\right) = P_{\text{T}}. \end{aligned} \tag{51}$$

Then, the necessary conditions for an optimal solution $\boldsymbol{F}$ is stated in Theorem 2.

Theorem 2. *The optimal band-shaped and low-triangular matrix $\boldsymbol{F}$ in (17) for user-rate maximization (51) shall satisfy the the conditions that, the optimal $\boldsymbol{f}_n^\nu$ is in (24) and $f_{n,n}$ equals*

$$f_{n,n} = \sqrt{N_0 \left[\frac{1}{\lambda_n \hat{g}_n^\nu} - 1\right]^+}, \tag{52}$$

where $\lambda_n > 0$ are a set of constants such that the transmit power constraint is satisfied.

Proof. The Lagrangian function for multiple constraints in this case reads

$$\mathcal{L} = R^{\text{user}} - \sum_{n=1}^{N} \mu_n \left(R^{\text{user}} - \log\left(1 + \frac{|f_{n,n}|^2}{N_0}\right)\right) - \lambda\left(\text{Tr}\left(\boldsymbol{F}^\dagger \boldsymbol{G} \boldsymbol{F}\right) - P_{\text{T}}\right), \tag{53}$$

and the necessary conditions are

$$\left.\begin{aligned} \frac{\partial \mathcal{L}}{\partial f_{n,k}} &= 0,\; 1 \le n, k \le N \\ \sum_{n=1}^{N} \mu_n &= 1,\; \mu_n \ge 0,\; 1 \le n \le N \\ R^{\text{user}} &= \log\left(1 + \frac{|f_{n,n}|^2}{N_0}\right),\; 1 \le n \le N \\ \text{Tr}\left(\boldsymbol{F}^\dagger \boldsymbol{G} \boldsymbol{F}\right) - P_{\text{T}} &= 0 \\ \lambda &\ge 0 \end{aligned}\right\}. \tag{54}$$

The first-order derivatives of $\mathcal{L}$ with respect to $f_{n,n}$ is

$$\frac{\partial \mathcal{L}}{\partial f_{n,n}} = \frac{\mu_n N_0 f_{n,n}}{N_0 + |f_{n,n}|^2} - \lambda \left(f_{n,n} g_{n,n} + (\boldsymbol{f}_n^{\nu})^{\dagger} \boldsymbol{g}_n^{\nu} \right), \tag{55}$$

while the gradient of $\mathcal{L}$ with respect to $\boldsymbol{f}_n^{\nu}$ is in (31). Then, from (31) the optimal $\boldsymbol{f}_n^{\nu}$ is solved in (24), and by inserting (24) back into (55) and setting the derivative to zero, we obtain

$$\frac{N_0 \mu_n}{N_0 + |f_{n,n}|^2} = \lambda \left(g_{n,n} - (\boldsymbol{g}_n^{\nu})^{\dagger} \left(\boldsymbol{G}_{n+1}^{\nu-1} \right)^{-1} \boldsymbol{g}_n^{\nu} \right). \tag{56}$$

Hence, as $N_0 > 0$, from (56) it holds that $\lambda > 0$ and $\mu_n > 0$ for all n. Otherwise, if either $\lambda = 0$ or $\mu_n = 0$ for some n, from (56) it holds that $\lambda = \mu_n = 0$ for all n, which contradicts the second necessary condition in (54) (due to $\partial \mathcal{L} / \partial R^{\text{user}} = 0$). By setting $\lambda_n = \lambda / \mu_n > 0$ and from (56) the optimal $f_{n,n}$ equals

$$|f_{n,n}|^2 = N_0 \left[\frac{1}{\lambda_n \hat{g}_n^{\nu}} - 1 \right]^+,$$

where $\hat{g}_{n,n}$ is defined in (26), and the optimal $f_{n,n}$ is then in (52). □

With the necessary conditions of an optimal $\boldsymbol{F}$ in Theorem 2, the user-rate is equal to the minimum user-rate for all users, that is,

$$R^{\text{user}} = \left[-\log \left(\lambda_n \hat{g}_n^{\nu} \right) \right]^+, \quad 1 \le n \le N, \tag{57}$$

and the power constraint can be written as

$$\sum_{n=1}^{N} \left[\frac{1}{\lambda_n} - \hat{g}_n^{\nu} \right]^+ = \frac{P_{\text{T}}}{N_0}. \tag{58}$$

Note that, different from the sum-rate maximization, now the water-level $1/\lambda_n$ varies for different users. From (57) and (58), the minimum user-rate can be solved for in closed-form,

$$R^{\text{user}} = \log \left(1 + \frac{P_{\text{T}}}{N_0 \sum\limits_{n=1}^{N} \hat{g}_n^{\nu}} \right), \tag{59}$$

and the optimal $f_{n,n}$ equals

$$f_{n,n} = \sqrt{N_0 \left(2^{R^{\text{user}}} - 1 \right)}, \quad 1 \le n \le N. \tag{60}$$

Although with the sum-rate maximization some user-rates may be decreased with a larger ν as shown in Corollary 1, for minimum user-rate maximization, R^{user} will not be decreased by a larger ν. Further, as the maximal minimum user-rate R^{user} in (57) is uniquely determined by the values of $\hat{g}_n^{\nu}$, we have the below property.

Property 3. *The conclusions drawn for sum-rate R^{sum} in Property 2 also stand for minimum user-rate R^{user}.*

3.3 Optimal User-Orderings

By permuting the order of the N users with an $N\times N$ permutation matrix $\boldsymbol{Q}$, the received signal model (5) reads

$$\boldsymbol{Qy} = \boldsymbol{QHPx} + \boldsymbol{Qz}. \tag{61}$$

Changing the order of the users may impact[3] the optimizations in (23) and (51), due to that the matrix $\boldsymbol{G}$ is updated with $\tilde{\boldsymbol{G}} = \boldsymbol{QGQ}^{\dagger}$ and the power constraint changes to,

$$\text{Tr}\left(\boldsymbol{F}^{\dagger}\tilde{\boldsymbol{G}}\boldsymbol{F}\right) \leq P_{\text{T}}, \tag{62}$$

Denoting the set that comprises all possible user-orderings as $\mathcal{P}$, and as the size $|\mathcal{P}| = N!$, it is infeasible to find an optimal ordering in a brute-force manner for large values of N. Therefore, we next introduce two efficient suboptimal user-ordering algorithms for the sum-rate and the minimum user-rate maximizations for $0 < \nu < N$ that have complexity orders $\mathcal{O}\left(\binom{N}{\nu+1}\right)$ and $\mathcal{O}(N)$, respectively. We start with the user-ordering for the sum-rate maximization (23). From (35), the optimal user-ordering $\mathcal{U} \in \mathcal{P}$ shall minimize the product[4],

$$\mathcal{U}^{\text{opt}} = \arg\min_{\mathcal{U}\in\mathcal{P}} \lambda^N \prod_{n=1}^{N} \hat{g}_n^{\nu}. \tag{63}$$

Denoting $q = \prod_{n=1}^{N} \hat{g}_n^{\nu}$ and since

$$\lambda = \left(\frac{P_{\text{T}}}{N_0} + \sum_{n=1}^{N} \hat{g}_n^{\nu}\right)^{-1} \leq \left(\frac{P_{\text{T}}}{N_0} + Nq^{\frac{1}{N}}\right)^{-1},$$

[3] This is true for cases $0 < \nu < N$. For $\nu = 0$, i.e., the linear ZF precoder, as the inter-user interferences are completely nulled out, different user-orderings have no impact on both the sum-rate or minimum user-rate maximizations.

[4] Without loss of generality, we assume $\lambda\hat{g}_{n,n} \geq 1$ holds for all users for both sum-rate and minimum user-rate maximizations.

it holds that

$$\lambda^N q \leq \left(\frac{P_{\mathrm{T}}}{N_0 q^{\frac{1}{N}}} + N \right)^{-N}. \tag{64}$$

Instead of directly minimizing (63), from (64) we can minimize the product q instead. On the other hand, from (38) and utilizing the matrix determinant lemma [31], $\hat{g}_n^\nu$ can be rewritten as $\hat{g}_n^\nu = \det \boldsymbol{G}_n^\nu / \det \boldsymbol{G}_{n+1}^{\nu-1}$, and q equals

$$q = \prod_{n=1}^{N} \frac{\det \boldsymbol{G}_n^\nu}{\det \boldsymbol{G}_{n+1}^{\nu-1}}. \tag{65}$$

By noticing that the sub-matrix $\boldsymbol{G}_n^\nu$ comprises $\boldsymbol{G}_{n+1}^{\nu-1}$ and an extra row and column vectors corresponding to the nth user, we can recursively order the users according to (65) as follows.

At a first stage, to minimize $\hat{g}_1^\nu$ we first find the best $\nu+1$ users that minimize $\det \boldsymbol{G}_1^\nu$, which needs to search over in total $\binom{N}{\nu+1}$ possible user combinations[5]. We denote the index set of the obtained $\nu+1$ users as $\mathcal{J}_1$. Then, in a second step, we select one single user from the chosen $\nu+1$ users that maximize $\det \boldsymbol{G}_2^{\nu-1}$, where $\det \boldsymbol{G}_2^{\nu-1}$ is obtained by removing the corresponding row and column vectors of the selected user in $\boldsymbol{G}_1^\nu$. One such user is selected to be the first user and set $\mathcal{U}(1)$ to its user-index.

At a second stage, we continue to order the remaining $N-1$ users, with ν users within the index set $\mathcal{J}_2 = \mathcal{J}_1 \setminus \mathcal{U}(1)$. In order to minimize $\hat{g}_2^\nu$, we first add another user from the remaining $N-\nu-1$ users to the ν users in $\mathcal{J}_2$ and calculate $\det \boldsymbol{G}_2^\nu$ corresponding to the selected $\nu+1$ users. The user from the remaining $N-\nu-1$ users that minimize $\det \boldsymbol{G}_2^\nu$ is selected, which needs $N-\nu-1$ operations. We update $\mathcal{J}_1$ as $\mathcal{J}_2$ plus the selected user-index. Then, we repeat the second step at the first stage to select one user from $\mathcal{J}_2$ (not $\mathcal{J}_1$ in order to keep the value of $\hat{g}_1^\nu$ unchanged) to maximize $\det \boldsymbol{G}_3^{\nu-1}$, and set $\mathcal{U}(2)$ to the index of that user.

Then, we update $\mathcal{J}_2 = \mathcal{J}_1 \setminus \mathcal{U}(2)$, and continue to order the remaining $N-2$ users in the same way until we finish the ordering of all users. Notice that, for the last ν users, we only need to recursively select the best user that maximizes $\det \boldsymbol{G}_{n+1}^{\nu-1}$. Such an algorithm is summarized in Algorithm 1.

Next, we analyze the user-ordering for the minimum user-rate maximization, which renders a simpler user-ordering algorithm. From (58), it holds that

$$\sum_{n=1}^{N} \hat{g}_n^\nu \left(2^{R^{\mathrm{user}}} - 1 \right) \leq \frac{P_{\mathrm{T}}}{N_0}. \tag{66}$$

[5]Note that, the ordering of the $\nu+1$ users inside each combination is independent with $\det \boldsymbol{G}_n^\nu$ since the determinant is invariant under the operation that permutes the row and column vectors in the same manner.

Algorithm 1 User-ordering for sum-rate maximization with the GZF-DP precoder.

1: Initialize $n=1$ and $\mathcal{I}_1=\mathcal{I}_2=[1,2,\cdots,N]$.
2: Search over all $\binom{N}{\nu+1}$ possible combinations to find the best $\nu+1$ users that minimizes the determinant of the principle sub-matrix $\det \boldsymbol{G}_1^{\nu}$ introduced by their indexes, and denote the best user-combination as $\mathcal{J}_1$, then set $\mathcal{J}_2=\mathcal{J}_1$.
3: Select one single user from all users in $\mathcal{J}_2$ to maximize $\det \boldsymbol{G}_2^{\nu-1}$, and denote its user-index as $\mathcal{U}(n)$.
4: Update $\mathcal{I}_1=\mathcal{I}_1\backslash\mathcal{U}(n)$, $\mathcal{J}_2=\mathcal{J}_1\backslash\mathcal{U}(n)$, $\mathcal{I}_2=\mathcal{I}_1\backslash\mathcal{J}_2$, and set $n=n+1$.
5: Replace the index $\mathcal{U}(n-1)$ in $\mathcal{J}_1$ with another user-index from the $N-\nu-n$ users in $\mathcal{I}_2$, such that $\det \boldsymbol{G}_n^{\nu}$ introduced by the updated $\mathcal{J}_1$ is minimized, and keep the updated $\mathcal{J}_1$.
6: Repeat Step 3-5 until $\mathcal{I}_2$ is empty. Then, recursively order the remaining ν users such that $\det \boldsymbol{G}_{n+1}^{\nu-1}$ is maximized at each stage.
7: Output the user-ordering $\mathcal{U}$.

Algorithm 2 User-ordering for minimum user-rate maximization with the GZF-DP precoder.

1: Order the user according to the descending order of the diagonal element $g_{n,n}$.

Therefore, the optimal user-ordering that maximizes R^{user} shall minimize the sum of $\hat{g}_n^{\nu}$,

$$\mathcal{U}^{\text{opt}} = \arg\min_{\mathcal{U}\in\mathcal{P}} \sum_{n=1}^{N} \hat{g}_n^{\nu}. \tag{67}$$

As for the last user, $\hat{g}_N^{\nu}=g_{N,N}$ holds, we can select the user that has the smallest diagonal element $g_{n,n}$ to be the last user $\mathcal{U}(N)$. Then, for the second last user, as

$$\hat{g}_{N-1}^{\nu} = g_{N-1,N-1} - \frac{|g_{N-1,N}|^2}{g_{N,N}}, \tag{68}$$

we can choose the user that has the second smallest diagonal element $g_{n,n}$ to be the second last user $\mathcal{U}(N-1)$. Recursively, based on (26), the users can be ordered in a descending order of $g_{n,n}$, which is summarized in Algorithm 2.

4 Empirical Results

In this section, simulation results are presented to show the promising performance of the proposed GZF-DP precoder for both the sum-rate and minimum user-rate maximizations. The sum-rate of the optimal DPC [14] serve as the upper-bound, while the sum-rate and

minimum user-rate of the linear ZF precoder serve as lower-bounds. For comparisons, we also present the rates of the UG-DP precoder in [20] for sum-rate maximizations, which are inferior to the GZF-DP precoder with $\nu = N_g - 1$ and similar DPC complexity. In all simulations, we set the noise power $N_0 = 1$ and test under Rayleigh fading channels that are based on the Kronecker correlation model

$$\boldsymbol{H} = \boldsymbol{R}_{\mathrm{R}}^{1/2} \boldsymbol{H}_{\mathrm{IID}} \boldsymbol{R}_{\mathrm{T}}^{1/2}, \tag{69}$$

where $N \times M$ matrices $\boldsymbol{H}_{\mathrm{IID}}$ denote IID complex Gaussian channels with zero mean and a covariance matrix being an identity matrix. The $M \times M$ matrix $\boldsymbol{R}_{\mathrm{T}}$ and $N \times N$ matrix $\boldsymbol{R}_{\mathrm{R}}$ denote the correlations at the transmit and receive sides, respectively. We use an exponential correlation model [32] for both $\boldsymbol{R}_{\mathrm{T}}$ and $\boldsymbol{R}_{\mathrm{R}}$, which is defined as

$$\boldsymbol{R} = \begin{bmatrix} 1 & \beta & \cdots & \cdots & \beta^{K-1} \\ \beta & 1 & \beta & \cdots & \beta^{K-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \beta \\ \beta^{K-1} & \beta^{K-2} & \cdots & \beta & 1 \end{bmatrix}, \tag{70}$$

where $K = M$, $\beta = \beta_{\mathrm{T}}$ and $K = N$, $\beta = \beta_{\mathrm{R}}$ for transmit and receive correlation, respectively.

4.1 Optimal Orderings

In Fig. 1, we evaluate the sum-rate with the channel given in Example 1 for all possible $4! = 24$ user-ordering schemes in $\mathcal{P}$. As it can be seen that, different user-orderings provide different sum-rate for $1 \leq \nu \leq 3$.

In Fig. 2, we evaluate the performance of Algorithm 1 for user-ordering for the sum-rate maximization with $M = N = 5$ and under IID complex Gaussian channels, that is, $\beta_{\mathrm{T}} = \beta_{\mathrm{R}} = 0$. The optimal ordering utilizes the brute-force method to select one best user-ordering over all $5! = 120$ possible combinations under each channel realization. The average sum-rate averages the sum-rate over all 120 user-orderings in $\mathcal{P}$. As can be seen, the proposed user-ordering performs 0.5 to 1 dB better than the averaged sum-rate in terms of transmit power P_{T}.

In Fig. 3, we evaluate the performance of Algorithm 2 for user-ordering for the minimum user-rate maximization with $M = N = 6$ and under IID complex Gaussian channels. As can be seen, the proposed Algorithm 2 performs around 1 dB better than the averaged sum-rate in terms of transmit power P_{T}, and quite close to the optimal user-ordering that is selected over $6! = 720$ possible schemes in $\mathcal{P}$ with brute-force method for each channel realization.

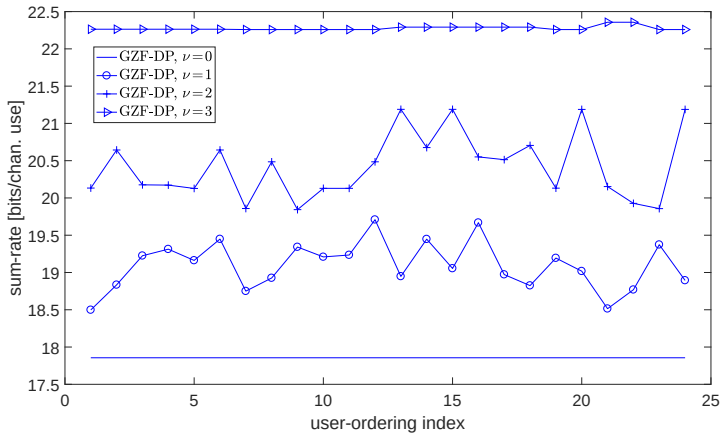


Figure 1: The sum-rate of the GZF-DP precoder with different ν evaluated with $N_0 = 1$ and $P_T = 10$ dB.

4.2 Sum-rate Maximization

Next we evaluate the sum-rate maximizations with $M = N = 8$. In Fig. 4 we simulate under IID complex Gaussian channels. As can be seen, the GZF-DP precoder with $\nu = 1$ renders around 1.5 dB and 4 dB gains compared to the UG-DP precoder and the linear ZF precoder in terms of transmit power P_T, respectively. With $\nu = 3$, which means that in the effective channel we preserve at most interference from 3 other users for each of the users, the GZF-DP precoder is only less than 1.5 dB away from the optimal DPC, and performs quite close to the ZF-DP precoder [9], i.e., the GZF-DP precoder with $\nu = 7$.

In Fig. 5, we repeat the tests in Fig. 4 under Rayleigh fading channels with correlation factors $\beta_T = 0.2$ and $\beta_R = 0.8$. As can be seen, the GZF-DP precoder with $\nu = 1$ renders around 2 dB and 5 dB gains compared to the UG-DP and ZF precoders in this case, respectively. The P_T gains of the GZF-DP precoder are larger than those gains as in Fig. 4, due to the fact that the MISO broadcast channels are correlated in this case. Moreover, we also evaluate the GZF-DP precoder with user-ordering based on Algorithm 1. For the UG-DP precoder, we use the brute-force method to select the optimal user-ordering under each channel realization. As it can be seen, with user-orderings both the GZF-DP and UG-DP precoders renders higher sum-rates. But still, even with the optimal user-ordering, the UG-DP precoder is 1.5 dB away from the proposed GZF-DP precoder without user-ordering.

4.3 Minimum User-rate Maximization

Next we evaluate the minimum user-rate maximizations with $M = N = 8$ and repeat the tests in Fig. 4 and Fig. 5, respectively. As can be seen, in Fig. 6 the proposed GZF-DP precoder with $\nu = 1$ is around 2 dB better than the linear ZF precoder, while in Fig. 7 the gain is more than 4 dB due to spatial correlated channels. In addition, in both cases, the

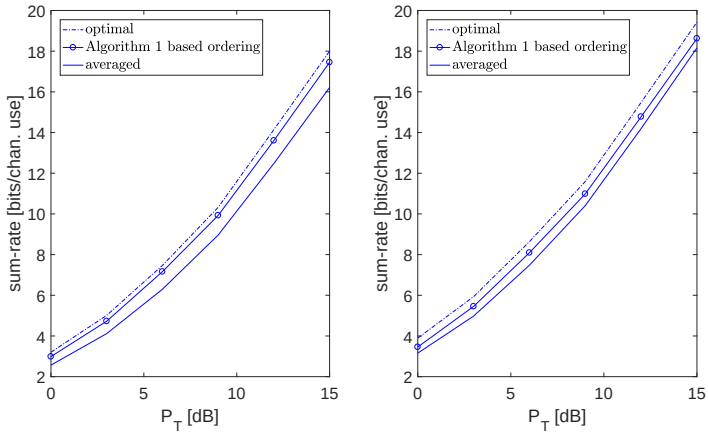


Figure 2: The sum-rate of the proposed Algorithm 1 for user-ordering with the GZF-DP precoder with $\nu = 1$ (the left figure) and $\nu = 2$ (the right figure) for $M = N = 5$.

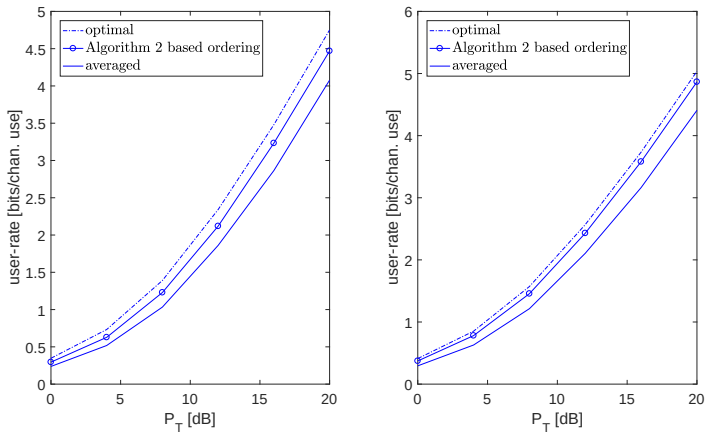


Figure 3: The user-rates of the proposed Algorithm 2 for user-ordering with the GZF-DP precoder with $\nu = 1$ (the left figure) and $\nu = 2$ (the right figure) for $M = N = 6$.

GZF-DP precoder with $\nu = 3$ performs close to the GZF-DP precoder with $\nu = 7$, i.e., the ZF-DP precoder.

4.4 Impact of the Number of Users and Correlation Factors

Next we evaluate the impacts of increasing the number of users and the spatial correlation factors. In all simulations, we set the total transmit power $P_{\mathrm{T}} = 10$ dB. In Fig. 8, we set the number of transmit antennas $M = 24$ and increase the user number N from 4 to 24. As can be seen, as the number of users increases, the sum-rate first increases and then decreases both for the linear ZF precoder and the GZF-DP precoder with $\nu < N-1$. This is so, since as N increases the degrees of freedom (DoF) for the precoder designs also increase and consequently the sum-rate is getting higher. However, the inter-user interference increased

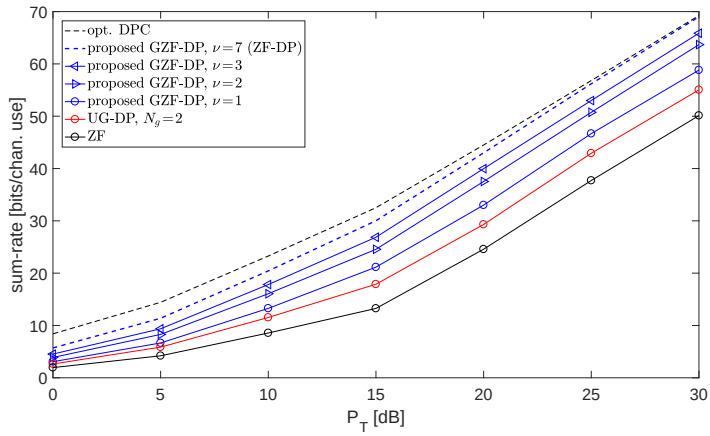


Figure 4: The sum-rate maximization with $M = N = 8$ under IID complex Gaussian channels.

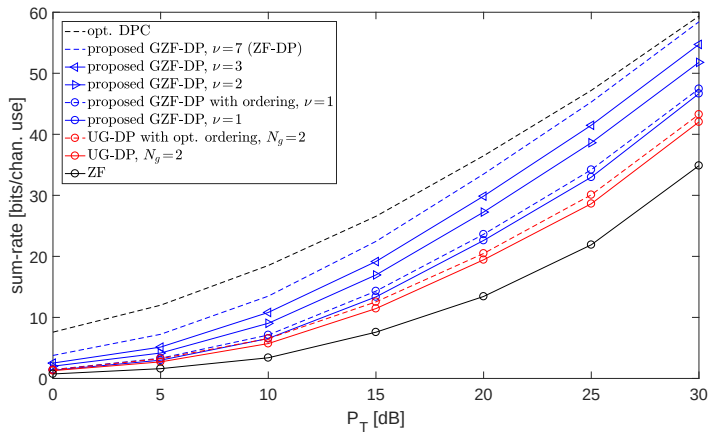


Figure 5: The sum-rate maximization with $M = N = 8$ under Rayleigh-fading channels and $\beta_{\mathrm{T}} = 0.2$ and $\beta_{\mathrm{R}} = 0.8$

with a larger N causes sum-rate degradation for small values of ν. Nevertheless, it can be seen that the GZF-DP precoder with $\nu = 1$ renders the same sum-rate as the ZF precoder with one user less.

In Fig. 9, we set $M = N = 8$ and $\beta_{\mathrm{T}} = \beta_{\mathrm{R}} = \beta$. We increase β from 0.1 to 0.9. As can be seen, as β gets higher, the sum-rate decreases for all precoders. At low and medium correlations, the GZF-DP precoder shows significant gains over the linear ZF precoder. For instance, the GZF-DP precoder with $\beta = 0.5$ renders the same sum-rate as the linear ZF precoder with $\beta = 0$. Therefore, the GZF-DP precoder is more robust against the transmit and receive correlations compared to the linear ZF precoder. In addition, with the user-ordering proposed in Algorithm 1 the correlation gain is even larger.

In Fig. 10, we repeat the tests in Fig. 8 for minimum user-rate maximizations. As can be seen, unlike the cases of the sum-rate maximizations, as the number of users increases, the

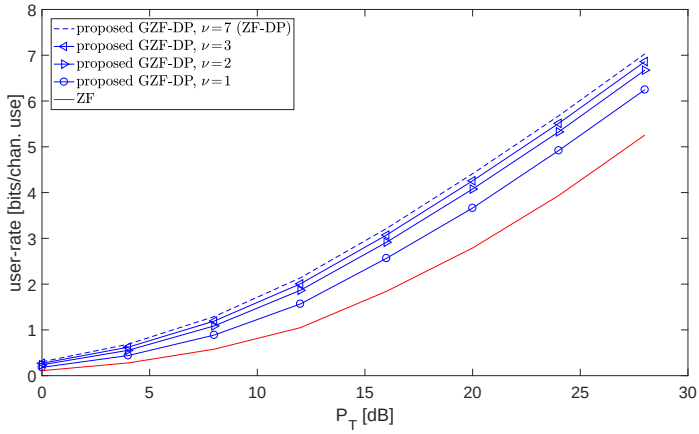


Figure 6: Repeat the test in Fig. 4 for minimum user-rate maximization under IID complex Gaussian channels.

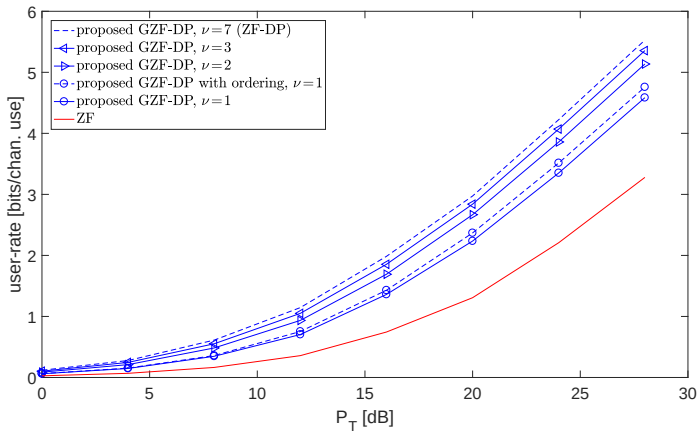


Figure 7: Repeat the test in Fig. 5 for minimum user-rate maximization under Rayleigh-fading channels.

user-rates of all precoder designs decrease. We also present a contour line of the sum-rate, which shows that the sum-rate also decreases when N is close to M. For large N, we can see that the GZF-DP precoder with $\nu = 1$ renders the same user-rate as the linear ZF precoder with one user less.

In Fig. 11, we repeat the tests in Fig. 9 for minimum user-rate maximizations. As can be seen, as the correlation factor β gets higher, the user-rates also decrease for all precoders. The GZF-DP precoder again shows superior performance compared to the linear ZF precoder, and is more robust against transmit and receive correlations.

4.5 Practical FD-MIMO Scenario

At last, we evaluate the proposed GZF-DP precoder in an FD-MIMO downlink scenario considering a 3D channel model [38]. The test scenario is depicted in Fig. 12, where we have

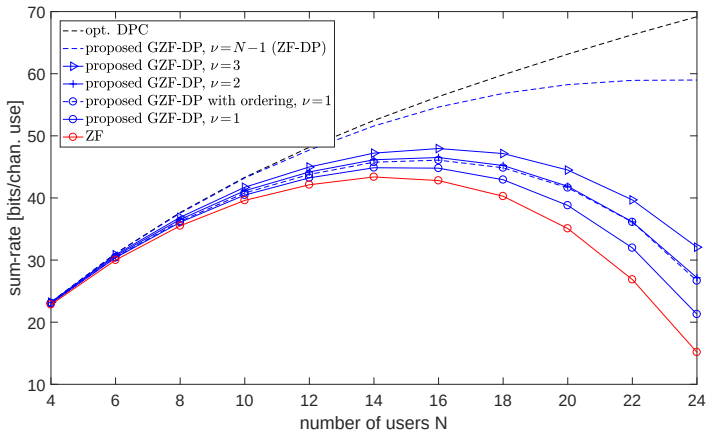


Figure 8: The sum-rate maximization with $M = 24$ and different number of users N under IID complex Gaussian channels. Note that the transmit power is constant no matter the number of users N.

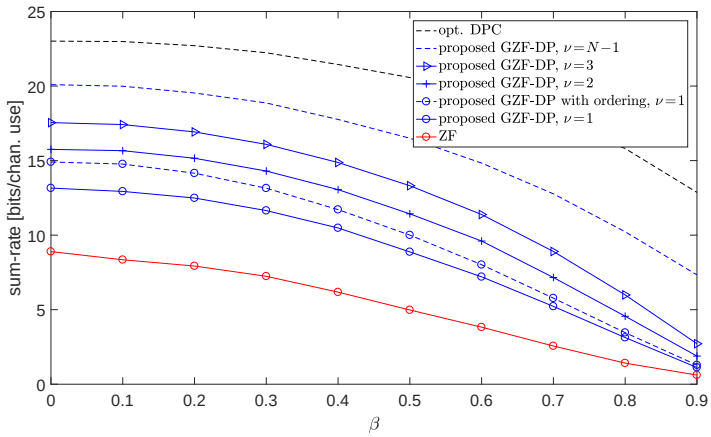


Figure 9: The sum-rate maximization with $M = N = 8$ under Rayleigh-fading channels. The correlation factors $\beta_{\mathrm{T}} = \beta_{\mathrm{R}}$, and change from 0.1 to 0.9.

an 8×8 2D antenna-array deployed at an e-NodeB that is 20 meters above the ground. The spacing between to adjacent antenna elements (both in horizontal and vertical dimensions) is 1/2 wave-length. The e-NodeB broadcasts at 2.4 GHz to 8 single-antenna users that are placed along a line which is perpendicular to the 2D antenna-plane. The distance between two adjacent users is 10 meters and the first user is 20 meters away from the e-NodeB. For simplicity, we consider an ideal line-of-sight (LOS) situation with channels constructed from the free-space path loss.

As shown in Fig. 13, the sum-rate of the proposed GZF-DP precoder with $\nu = 1$ is much higher than that of the linear ZF precoder. And with $\nu = 3$ the GZF-DP precoder significantly outperforms the UG-DP precoder with $N_g = 4$. Moreover, the GZF-DP precoder with $\nu = 3$ also performs close to the ZF-DP precoder which requires a full successive DPC scheme.

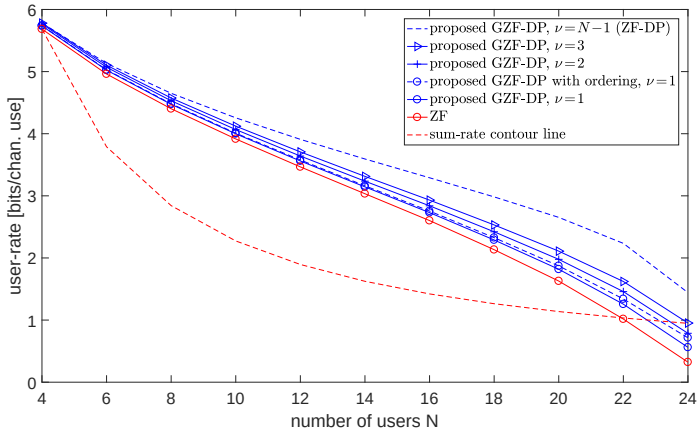


Figure 10: Repeat the tests in Fig. 8 for user-rate maximization.

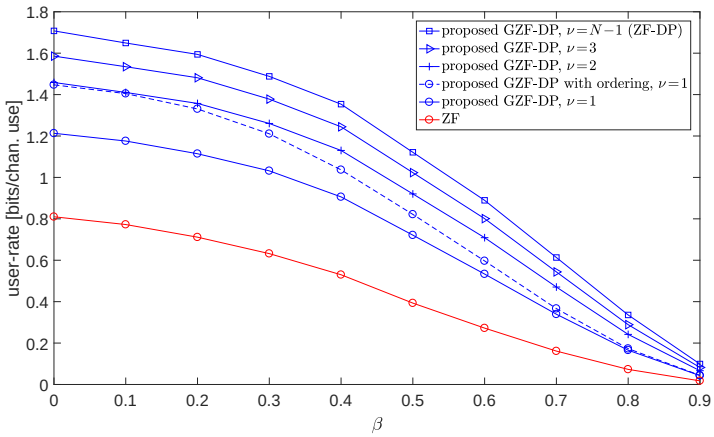


Figure 11: Repeat the tests in Fig. 9 for user-rate maximization.

5 Summary

We have proposed a generalized zero-forcing precoder (GZF) in conjunction with successive dirty-paper coding (DPC), namely, the GZF-DP precoder, for multi-input-single-output (MISO) broadcast channels. Utilizing the successive DPC encoding scheme at the transmitter to cancel the known non-causal interference, the GZF-DP precoder preserves up to ν interferers for each of the users and results in significant rate-increments. We analyze optimal designs of the proposed GZF-DP precoder both for sum-rate and minimum user-rate maximizations. The optimal GZF-DP precoder designs are solved in closed-forms in relation to optimal power allocations. For the sum-rate maximization, the optimal power allocation can be efficiently found with modified water-filling schemes introduced by inter-user interference, while for the minimum user-rate maximization, the optimal power allocation is solved in closed-from. We have also derived two efficient and low-complexity user-ordering algorithms for the GZF-DP precoder for the sum-rate and minimum user-

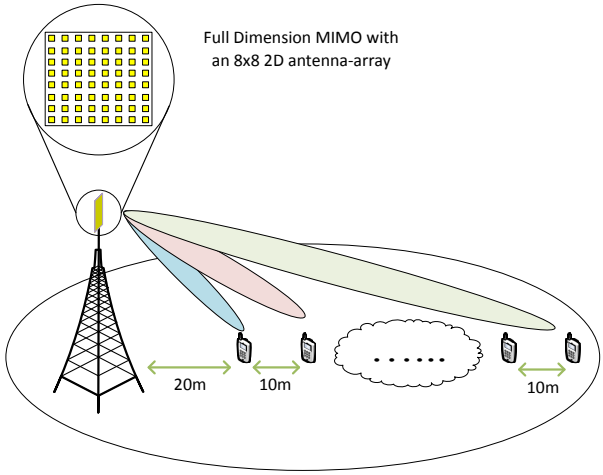


Figure 12: An FD-MIMO scenario where an e-NodeB equipped with an 8×8 2D antenna-array is broadcasting to 8 lined-up single-antenna users.

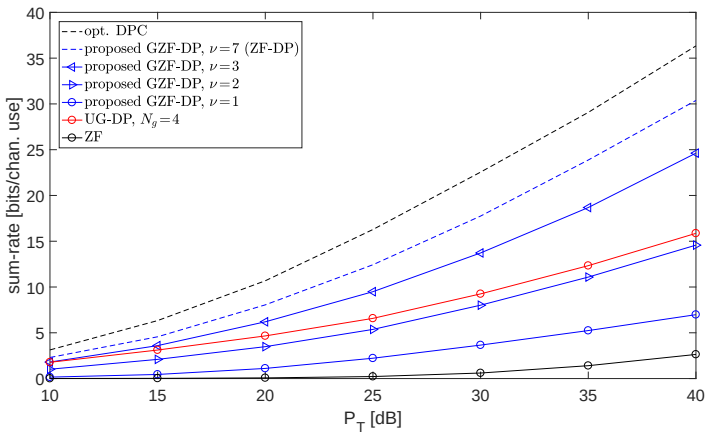


Figure 13: The sum-rate maximization for the FD-MIMO scenario considered in Fig. 12.

rate maximizations, respectively. We show through numerical results that, the proposed GZF-DP precoder yields both much higher sum-rate and minimum user-rate compared to the traditional linear ZF precoder and the previous user-grouping based DPC (UG-DP) precoder, and is close to the ZF with full complexity DPC (ZF-DP) precoder.

Acknowledgment

The authors would like to thank the Associate Editor and anonymous reviewers for their helpful and constructive comments which significantly improved the quality of this paper.

References

[1] Ericsson, White Paper, "Cellular networks for massive IoT," Jan. 2016.

[2] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "A Low-complexity channel shortening receiver with diversity support for evolved 2G device," IEEE Int. Conf. Commun. (ICC), Kuala Lumpur, Malaysia, May, 2016, pp. 1-7.

[3] 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 Process. Mag.*, vol. 30, no. 1, pp. 40-60, Jan. 2013.

[4] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 41-51, May 2015.

[5] 3GPP, TS 36.201, *Evolved universal terrestrial radio access (E-UTRA); LTE physical layer*, Release 13, Jun. 2016.

[6] Y. Kim, H. Ji, J. Lee, Y. H. Nam, B. L. Ng, I. Tzanidis, Y. Li, and J. Z. Zhang, "Full dimension MIMO (FD-MIMO): The next evolution of MIMO in LTE systems," *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, pp. 92-100, Jun. 2014.

[7] S. Hu, X. Gao, and F. Rusek, "Linear precoder design for MIMO-ISI broadcasting channels under channel shortening detection," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1207-1211, Sep. 2016.

[8] F. Petitcolas, R. Anderson, and M. Kuhn, "Information hiding-A survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062-1078, Jul. 1999.

[9] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691-1706, Jul. 2003.

[10] A. Hørest-Madsen, "Capacity bounds for cooperative diversity," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1522-1544, Apr. 2006.

[11] S. Gelfand and M. Pinsker, "Coding for channel with random parameters," *Problems Cont. and Inf. Theory*, vol. 9, no. 1 pp. 19-31, 1980.

[12] M. H. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 2, pp. 439-441, May 1983.

[13] A. E. Gamal and Y. H. Kim, *Network information theory*, Cambridge University Press, 2011.

[14] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 768-783, May 2004.

[15] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink–downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912-1921, Aug. 2003.

[16] Y. Sun, Y. Yang, A. D. Liveris, V. Stankovic, and Z. Xiong, "Near-capacity dirty-paper code design: A source-channel coding approach," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3013-3031, Jul, 2009.

[17] U. Erez, S. Shamai (Shitz), and R. Zamir, "Capacity and lattice-strategies for cancelling known interferences," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 3820-3833, Nov. 2005.

[18] U. Erez and S. ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3417-3412, Oct. 2005.

[19] W. Yu, D. P. Varodayan, and J. M. Cioffi, "Trellis and convolutional precoding for transmitter based interference presubtraction," *IEEE Trans. Comm.,* vol. 53, no.7, pp. 3013-3031, Jul. 2005.

[20] S. K. Mohammed, and E. G. Larsson, "Improving the performance of the zero-forcing multiuser MISO downlink precoder through user grouping," *IEEE Trans. Wireless Commun.,* vol. 15, no. 2, pp. 811-826, Feb. 2016.

[21] S. C. Lin and H. J. Su, "Practical vector dirty paper coding for MIMO Gaussian broadcast channels," *IEEE J. Sel. Areas in Commun.,*, vol. 25, no. 7, pp. 1345-1357, Sep. 2007.

[22] T. H. A. Nosratinia and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol 42. no. 10, pp. 68-73, Oct. 2004.

[23] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570-1580, Apr. 2005.

[24] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory,* vol. 52, no. 9, pp. 3936-3964, Sep. 2006.

[25] Z. Wang and W. Chen,"Regularized zero-forcing for multiantenna broadcast channels with user selection," *IEEE Wireless. Commun. Lett.,* vol. 1, no. 2, pp. 129-132, Apr. 2012.

[26] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Mar. 2004.

[27] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*, the first edition, Academic Press, Nov. 1982.

[28] W. Yu, "Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition" *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 754-759, Feb. 2006.

[29] F, Zhang, *The Schur complement and its applications,* Numerical Methods and Algorithms, vol. 4, Springer, Jan. 2005.

[30] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3, Johns Hopkins Studies in the Mathematical Sciences, JHU Press, 2013.

[31] D. A. Harville, *Matrix algebra from a statistician's perspective,* Springer, 2008.

[32] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369-371, Sep. 2001.

[33] J. Choi and D. J. Love, "Bounds on eigenvalues of a spatial correlation matrix," *IEEE Commun. Lett.*, vol. 18, no. 8, pp. 1391-1394, Aug. 2014

[34] J. N. Pierce and S. Stein, "Multiple diversity with nonindependent fading," *Proc. IRE,* vol. 48, no. 1, pp. 89-104, Jan. 1960.

[35] A. D. Dabbagh and D. J. Love, "Precoding for multiple antenna Gaussian broadcast channels with successive zero-forcing," *IEEE Trans. Signal Process.,* vol. 55, no. 7, pp. 3837-3850, Jul. 2007.

[36] L.-N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, "Weighted sum rate maximization for MIMO broadcast channels using dirty paper coding and zero-forcing methods," *IEEE Trans. Commun.,* vol. 61, no. 6, pp. 2362-2373, Jun. 2013.

[37] L.-N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, "Beamformer designs for MISO broadcast channels with zero-forcing dirty paper coding," *IEEE Trans. Wireless Commun.,* vol. 12, no. 3, pp. 1173-1185, Mar. 2013.

[38] Y. H. 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 Commun. Mag.*, vol. 51, no. 6, pp. 172-179, Jun. 2013.

Paper VI

Linear Precoder Design for MIMO-ISI Broadcasting Channels under Channel Shortening Detection

We consider optimal precoder design for multi-user multi-input multi-output (MIMO) broadcasting channels in single-carrier (SC) systems. Instead of linear detection, we assume that advanced non-linear channel shortening (CS) detectors are utilized at the receivers. Such a scenario is challenging for precoder design as the uplink-downlink duality is inapplicable. The target of our linear precoder design is to maximize the sum of the achievable information rate (sum-AIR), with AIR of each user being explicitly derived. We analyze such a precoder design in general, and provide an efficient per-user based optimization algorithm for the design of block-diagonalization precoder.

S. Hu, X. Gao, and F. Rusek

"Linear precoder design for MIMO-ISI broadcasting channels under channel shortening detection,"

IEEE Trans. Signal Process. Lett., vol. 23, no. 9, pp. 1207-1211, Sep. 2016.

1 Introduction

Although multi-carrier (MC) modulation nowadays is the main stream, single-carrier (SC) modulation has its potential in the emerging Internet-of-Things (IoT) and machine-to-machine (M2M) type of communications [1]. Compared to MC modulation, SC modulation has advantages of better peak average power ratio (PAPR), stringent adjacent channel emissions, lower radio frequency integrated circuit cost, and being compatible to vast legacy 2G devices [2]. To combat intersymbol interference (ISI) in SC system introduced by delay dispersion in propagation channels, advanced non-linear channel shortening (CS) detectors are widely deployed [3]. This is due to the fact that in most practical scenarios, the optimal maximum likelihood (ML) detection results in unacceptable computational cost, while linear detectors, such as zero-forcing (ZF) and linear minimum-mean-squared-error (LMMSE), render unsatisfying performance.

The concept of CS detection is to filter the received signal with a prefilter so that the target response has a much shorter duration than the delay dispersion of the ISI channel, and then applying the Viterbi algorithm [4] to the shorter channel response. CS detectors can be designed via several criteria such as minimum-phase filtering [3], minimum-mean-squared-error (MMSE) [5, 6], minimum-mean-output-energy (MMOE) [7–9], and achievable information rate (AIR) maximization. CS based on AIR-maximization was first introduced in [10], where the AIR has a closed-form expression, under the assumption that transmit symbols are Gaussian distributed. This enables analytical study of the CS detector, a feature generally unavailable for other CS detectors. Moreover, in [11] the authors showed that it is beneficial to take the detection into account when designing a transmit filter for single-user multi-input multi-output (MIMO) ISI channels. As an extension to [11], this letter considers the design of AIR maximized precoder for multi-user broadcast SC systems with non-linear CS detection at the autonomous receivers.

Linear precoder design is a crucial part to increase data rate and improve transmit energy-efficiency for MIMO-ISI channels. Conventionally, precoder design at the base-station (BS) assumes that either linear or ML detection is utilized at the receiver. In both cases, the problem of precoder design can be solved via the uplink-downlink duality [12, 13] between the multiple access channel and broadcasting channel. However, when designing precoder for CS detector, the duality does not carry over. The is so since the closed-form expression of AIR is related to the principle submatrix of the mean squared error (MSE) matrix, not only its diagonal elements as considered in [13] or the full matrix as for ML detection. As a consequence, the precoder design under CS detection is challenging. Moreover, optimizing the precoder for MMSE or ML detection, but then applying CS at the receivers render unsatisfying performance as the precoder is not taking the receiver into account.

In this letter we consider linear precoder design, explicitly constructed for CS detection, for

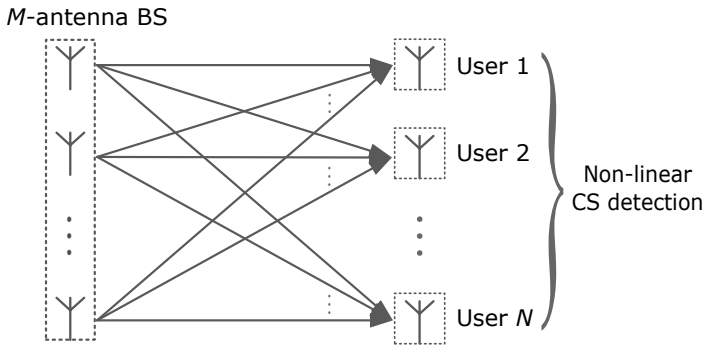


Figure 1: Broadcasting channel in SC system with an M-antenna BS and N single-antenna users that utilize CS detection.

MIMO-ISI broadcasting channels. We consider the general form precoder that maximizes sum-AIR, and solve the design problem for the block-diagonalization precoder. In the latter case multi-user crosstalk is eliminated and only ISI remains, which can be dealt with by CS detector at receivers. Hence, the proposed precoder design is more efficient, with the freedom of preserving some ISI, than the traditional linear precoder design that assumes linear detection. As we are dealing with ISI channels, Fourier analysis for Toeplitz matrix is extensively used, including Szegő's theorem and discrete-time Fourier transform (DTFT), see e.g., [10, 14].

Notations

Operators '$\star$', '$\otimes$', '$[\cdot]^+$', and '$\lfloor\cdot\rfloor$' denote linear convolution, Kronecker product, non-negative protection, and floor operation, respectively. Superscripts '-1', '$*$', 'T', and '$\dagger$' denote the inverse, complex conjugate, transpose, and Hermitian transpose, respectively. In addition, $\mathrm{Tr}\{\cdot\}$ takes the trace, and $\mathcal{R}\{\cdot\}$ fetches the real part of the arguments.

2 Multi-user MIMO-ISI Signal Model

Consider a single-carrier MIMO system with M-antennas BS and N single-antenna users. We denote K as the transmit data block size[1], L as the longest duration of all ISI channels, and S as the longest tap length of all precoders. The received signal vector $\boldsymbol{y}_n$ at the nth user can be written as

$$\boldsymbol{y}_n = \sum_{m=1}^{M} \boldsymbol{h}_{n,m} \star \left(\sum_{k=1}^{N} \boldsymbol{p}_{k,m} \star \boldsymbol{d}_k \right) + \boldsymbol{w}_n, \tag{1}$$

[1]Since we are dealing with ISI channels, later we implicitly let K go to infinity in order to apply Szegő's theorem.

where $\boldsymbol{h}_{n,m}$ is the $L\times 1$ channel vector linking the mth BS antenna and the nth user, $\boldsymbol{p}_{n,m}$ is the $S\times 1$ precoder, and $\boldsymbol{d}_n$ is the $K\times 1$ transmit signal vector. The noise term $\boldsymbol{w}_n$ contains independent and identically distributed (IID) complex Gaussian variables with zero mean and covariance matrix $N_0\boldsymbol{I}$.

We let $\boldsymbol{H}_{n,m}$ and $\boldsymbol{P}_{n,m}$ represent $K\times K$ circular convolutions generated from $\boldsymbol{h}_{n,m}$ and $\boldsymbol{p}_{n,m}$, respectively. As K grows large, circular convolutions can represent normal convolutions to any given precision, see [15] for a rigorous information theoretical treatment. Denote $\boldsymbol{Y}=\left[\boldsymbol{y}_1^{\mathrm{T}}\ \boldsymbol{y}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{y}_N^{\mathrm{T}}\right]^{\mathrm{T}}$, $\boldsymbol{D}=\left[\boldsymbol{d}_1^{\mathrm{T}}\ \boldsymbol{d}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{d}_N^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\boldsymbol{W}=\left[\boldsymbol{w}_1^{\mathrm{T}}\ \boldsymbol{w}_2^{\mathrm{T}}\ \cdots\ \boldsymbol{w}_N^{\mathrm{T}}\right]^{\mathrm{T}}$ as the vectors of received signal, transmit signal and noise vectors that comprise all users, respectively. Further, denote $\boldsymbol{H}_n=[\boldsymbol{H}_{n,1}\,\boldsymbol{H}_{n,2}\cdots\boldsymbol{H}_{n,M}]^{\mathrm{T}}$, $\boldsymbol{P}_n=[\boldsymbol{P}_{n,1}\,\boldsymbol{P}_{n,2}\cdots\boldsymbol{P}_{n,M}]^{\mathrm{T}}$, and let $\boldsymbol{H}=[\boldsymbol{H}_1\,\boldsymbol{H}_2\ \cdots\ \boldsymbol{H}_N]^{\mathrm{T}}$, $\boldsymbol{P}=[\boldsymbol{P}_1\,\boldsymbol{P}_2\ \cdots\ \boldsymbol{P}_N]$. Then, we can rewrite (1) in the matrix form as

$$\boldsymbol{Y}=\boldsymbol{H}\boldsymbol{P}\boldsymbol{D}+\boldsymbol{W}. \tag{2}$$

The design of precoder $\boldsymbol{P}$ is subject to the power constraint

$$c=\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{s=1}^{S}\left|p_{n,m}(s)\right|^2\leq 1. \tag{3}$$

Conventionally, precoders are optimized according to ZF, MMSE, or the ML criteria [13]. These methods assume either ML detection, or linear detection at receivers. In this work, we extend the state-of-the-art by assuming CS detectors, which operate in between ML and linear detection, in terms of both performance and complexity. With CS detection considered, we elaborate our precoder design in the next section.

3 Linear Precoder Design with CS Detection

Denoting the $NK\times NK$ effective channel $\boldsymbol{F}=\boldsymbol{H}\boldsymbol{P}$ in (2), the purpose is to design $\boldsymbol{F}$ with CS detectors at the receivers. With $\boldsymbol{F}$ found, the precoder $\boldsymbol{P}$ can be obtained through

$$\boldsymbol{P}=\boldsymbol{H}^{\dagger}\left(\boldsymbol{H}\boldsymbol{H}^{\dagger}\right)^{-1}\boldsymbol{F}. \tag{4}$$

If $\boldsymbol{F}$ is diagonal, $\boldsymbol{P}$ is ZF precoder that eliminates both multi-user crosstalk and ISI. We formulate the precoder design under CS detection for a general form of $\boldsymbol{F}$, and focus on the case that $\boldsymbol{F}$ is block-diagonal, with ZF precoder being a trivial case.

3.1 Problem Formulation

Consider the effective channel

$$\boldsymbol{F}=\begin{bmatrix} \boldsymbol{F}_{1,1} & \boldsymbol{F}_{1,2} & \cdots & \boldsymbol{F}_{1,N} \\ \boldsymbol{F}_{2,1} & \boldsymbol{F}_{2,2} & \cdots & \boldsymbol{F}_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{F}_{N,1} & \boldsymbol{F}_{N,2} & \cdots & \boldsymbol{F}_{N,N} \end{bmatrix}, \tag{5}$$

where each block $\boldsymbol{F}_{n,k}$ is a $K\times K$ circular convolution matrix generated from an effective ISI channel after precoding. Based on (2) and (4), the received signal vector of the nth user reads

$$\boldsymbol{y}_n = \boldsymbol{F}_{n,n}\boldsymbol{d}_n + \sum_{k=1,k\neq n}^{N} \boldsymbol{F}_{n,k}\boldsymbol{d}_k + \boldsymbol{w}_n. \tag{6}$$

The CS detection on model (6) is to maximize the AIR based on the detection model

$$T(\boldsymbol{y}_n|\boldsymbol{d}_n) = \exp\Big(-2\mathcal{R}\left\{\boldsymbol{d}_n^{\dagger}\boldsymbol{V}_n\boldsymbol{y}_n\right\} + \boldsymbol{d}_n^{\dagger}\boldsymbol{G}_n\boldsymbol{d}_n\Big), \tag{7}$$

which can be carried out over a trellis with $|\mathcal{X}|^{\nu}$ states, where $|\mathcal{X}|$ is the cardinality of the symbol constellation and ν is the memory length of the CS detector. The $K\times K$ circular convolution matrix $\boldsymbol{V}_n$ represents a prefiltering on the received samples, and $\boldsymbol{G}_n$ is a Hermitian Toeplitz matrix with only the middle $(2\nu+1)$ diagonals taking non-zero values.

In order to optimize $(\boldsymbol{V}_n, \boldsymbol{G}_n)$, the AIR corresponding to detection model (7) is adopted as the objective function. With optimal $(\boldsymbol{V}_n, \boldsymbol{G}_n)$ in [10, Proposition 2], the AIR equals

$$I_n = -\log\det(\boldsymbol{B}_n^{\nu}), \tag{8}$$

where the $(\nu+1)\times(\nu+1)$ Hermitian matrix $\boldsymbol{B}_n^{\nu}$ is the principal submatrix formed by any contiguous $(\nu+1)$ rows and the corresponding columns of the MSE matrix

$$\boldsymbol{B}_n = \boldsymbol{I} - \boldsymbol{F}_{n,n}^{\dagger}\left(\sum_{k=1}^{N}\boldsymbol{F}_{n,k}\boldsymbol{F}_{n,k}^{\dagger} + N_0\boldsymbol{I}\right)^{-1}\boldsymbol{F}_{n,n}. \tag{9}$$

Denoting $B_n(\omega)$ and $F_{n,k}(\omega)$ as the DTFTs of $\boldsymbol{B}_n$ and $\boldsymbol{F}_{n,k}$, respectively [14], and from (9), it holds that

$$B_n(\omega) = 1 - \frac{|F_{n,n}(\omega)|^2}{N_0 + \sum_{k=1}^{N}|F_{n,k}(\omega)|^2}. \tag{10}$$

Further, as the (t_1, t_2)th element of $\boldsymbol{B}_n^{\nu}$ equals

$$B_n^{\nu}(t_1, t_2) = \frac{1}{2\pi} \int_{-\pi}^{\pi} B_n(\omega) \exp\left(j\omega(t_1 - t_2)\right) \mathrm{d}\omega, \tag{11}$$

the AIR (8) is a function of $F_{n,k}(\omega)$. Moreover, the power c in (3) can be rewritten as

$$\begin{aligned} c &= \frac{1}{K} \mathrm{Tr}\left\{\boldsymbol{P}^{\dagger}\boldsymbol{P}\right\} \\ &= \frac{1}{K} \mathrm{Tr}\left\{\boldsymbol{F}^{\dagger}\left(\boldsymbol{H}\boldsymbol{H}^{\dagger}\right)^{-1}\boldsymbol{F}\right\} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{n=1}^{N} \sum_{k=1}^{N} \left(\sum_{i=1}^{N} F_{n,i}(\omega) F_{k,i}^{*}(\omega) \right) \Lambda_{k,n}(\omega) \mathrm{d}\omega, \end{aligned} \tag{12}$$

where $\Lambda_{n,k}(\omega)$ is the DTFT of the (n,k)th block entry in $(\boldsymbol{H}\boldsymbol{H}^{\dagger})^{-1}$, which can be obtained by the DTFT of $\boldsymbol{H}$.

Next we state our main theorem, which shows the stationary condition for the optimal $\boldsymbol{F}$.

Theorem 1. *For the sum-AIR optimization problem*

$$\begin{aligned} &\max_{\{F_{n,k}(\omega)\}} \sum_{n=1}^{N} I_n \\ &\text{subject to } c \le 1, \end{aligned} \tag{13}$$

the optimal $F_{n,k}(\omega)$ satisfies, with a set of constants $\{A_t^n\}$,

$$\mathcal{R}\left\{\sum_{t=0}^{\nu} A_t^n \exp\left(j\omega t\right)\right\} \frac{\zeta_{n,k}(\omega) F_{n,k}^{*}(\omega)}{\left(N_0 + \sum_{i=1}^{N} \left|F_{n,i}(\omega)\right|^2\right)^2} = \sum_{i=1}^{N} F_{i,k}^{*}(\omega) \Lambda_{n,i}(\omega), \tag{14}$$

where

$$\zeta_{n,k}(\omega) = \begin{cases} -|F_{n,n}(\omega)|^2 & k \neq n, \\ N_0 + \sum_{i=1, i\neq n}^{N} |F_{n,i}(\omega)|^2 & k = n. \end{cases} \tag{15}$$

Proof. From [11, Theorem 1], the functional derivative of I_n in (8) with respect to $B_n(\omega)$ equals

$$\frac{\delta I_n}{\delta B_n(\omega)} = \mathcal{R}\left\{\sum_{t=0}^{\nu} A_t^n \exp\left(j\omega t\right)\right\}, \tag{16}$$

where $\{A_t^n\}$ are a set of constants independent of ω. Further, with $\zeta_{n,k}(\omega)$ defined in (15), the derivatives of $B_n(\omega)$ in (10) and c in (12) with respect to $F_{n,k}(\omega)$ are

$$\frac{\delta B_n(\omega)}{\delta F_{n,k}(\omega)} = -\frac{\zeta_{n,k}(\omega)F_{n,k}^*(\omega)}{\left(N_0 + \sum_{i=1}^N |F_{n,i}(\omega)|^2\right)^2}, \tag{17}$$

$$\frac{\delta c}{\delta F_{n,k}(\omega)} = \frac{1}{2\pi}\sum_{i=1}^N F_{i,k}^*(\omega)\Lambda_{n,i}(\omega). \tag{18}$$

Utilizing chain rule and the Euler-Lagrange [16] equation

$$\frac{\delta I_n}{\delta B_n(\omega)}\frac{\delta B_n(\omega)}{\delta F_{n,k}(\omega)} = -\lambda\frac{\delta c}{\delta F_{n,k}(\omega)},$$

and combining (16)-(18), we can reach (14). □

As (14) is quintic with respect to $F_{n,k}(\omega)$ at each frequency ω, finding optimal $F_{n,k}(\omega)$ is challenging. We next turn to the tractable case that $\boldsymbol{F}$ is block diagonal. Later we show through simulation results that such a block-diagonalization precoder has neglectable rate loss compared to the full $\boldsymbol{F}$ when the spatial correlation is small, which is a favorable outcome since optimizing the block diagonal precoder is much simpler.

3.2 Block-Diagonalization Precoder

In this case we assume that $\boldsymbol{F}$ is block diagonal,

$$\boldsymbol{F} = \mathrm{diag}\big[\,\boldsymbol{F}_{1,1} \quad \boldsymbol{F}_{2,2} \quad \cdots \quad \boldsymbol{F}_{N,N}\,\big], \tag{19}$$

where inter-user interference is completely eliminated and ISI remains for each user. The received signal model (6) now reads

$$\boldsymbol{y}_n = \boldsymbol{F}_{n,n}\boldsymbol{d}_n + \boldsymbol{w}_n.$$

Based on the effective channel $\boldsymbol{F}_{n,n}$, the nth user applies CS detection as described in Sec. III-A. The power c in (12) can be rewritten as $c = \sum_{n=1}^N c_n$, where c_n is the power allocated to the nth user,

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} |F_{n,n}(\omega)|^2\,\Lambda_{n,n}(\omega)\mathrm{d}\omega.$$

From Theorem 1, the optimal $F_{n,n}(\omega)$ satisfies

$$\mathcal{R}\left\{\sum_{t=0}^{\nu} A_t^n \exp\left(j\omega t\right)\right\}\frac{N_0}{(N_0 + |F_{n,n}(\omega)|^2)^2} = \Lambda_{n,n}(\omega),$$

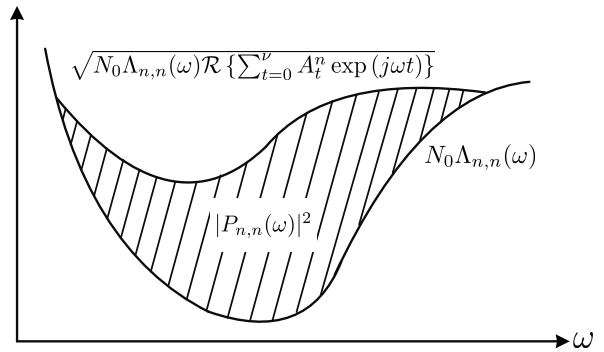


Figure 2: Generalized water-filling scheme with the precoder $P_{n,n}(\omega)$ satisfying $|P_{n,n}(\omega)|^2 = \Lambda_{n,n}(\omega)\,|F_{n,n}(\omega)|^2$ from (4).

which can be solved as

$$|F_{n,n}(\omega)|^2 = \left[\sqrt{\frac{N_0\mathcal{R}\{\sum_{t=0}^{\nu} A_t^n \exp(j\omega t)\}}{\Lambda_{n,n}(\omega)}} - N_0\right]^+. \tag{20}$$

As illustrated in Fig. 2, (20) can be viewed as a generalized water-filling scheme under CS detection. Notice that, the AIR calculated via (8) and (11) only depends on the magnitude $|F_{n,n}(\omega)|^2$ and is irrelevant of the phase of $F_{n,n}(\omega)$. For $\nu = 0$, the CS detection degrades to LMMSE and (20) becomes a standard water-filling as shown in the corollary below.

Corollary 1. *When $\nu = 0$, the optimal $F_{n,n}(\omega)$ satisfies*

$$|F_{n,n}(\omega)|^2 = \left[\sqrt{\frac{N_0 A_0^n}{\Lambda_{n,n}(\omega)}} - N_0\right]^+,$$

and A_0^n is the power allocation that is optimized via (13).

With Theorem 1, the optimization in (13) reduces to the optimization of a set of coefficients $\{A_t^n\}_{0\le t\le\nu}$ for all N users. For a large N, joint optimization of the coefficients over users can be very complex, and it is more efficient to separate the joint optimization into N individual optimizations. Such a strategy is summarized in Algorithm 1, which comprises two major steps. The first step is to optimize $\{A_t^n\}_{0\le t\le\nu}$ for each user separately with a allocated power β_n,

$$\begin{aligned} &\max_{\{A_t^n\}_{0\le t\le\nu}} I_n, \\ &\text{subject to (20) and } c_n \le \beta_n. \end{aligned} \tag{21}$$

Then the second step is to adjust the power allocation over iterations to maximize the sum-AIR.

Algorithm 1 Block-Diagonalization Precoder Optimization.

1: Initialize $\beta_n = 1/N$ for all users, set ϵ, and let $\ell = 1$.
2: Solve optimization (21) for all N users in parallel..
3: Repeat Step 2 with $\tilde{\beta}_n = \beta_n + \epsilon_\ell$, where $\epsilon_\ell = \min_n(\epsilon/\ell, \beta_n)$.
4: Find the $\lfloor N/2 \rfloor$ users that have the largest AIR increments and retain $\tilde{\beta}_n$.
5: Find the $\lfloor N/2 \rfloor$ users that have the smallest AIR increments and update $\tilde{\beta}_n = \beta_n - \epsilon_\ell$. Remove the users with $\tilde{\beta} = 0$ and update N accordingly. Repeat Step 2 for the rest users.
6: If N is odd, the remaining user retain $\tilde{\beta}_n = \beta_n$.
7: After Step 3-6, the total power c is still 1. Update $\beta_n = \tilde{\beta}_n$ and $\ell = \ell + 1$. Go to Step 3 until sum-AIR converges.

4 Numerical Results

In this section we provide simulation results of the proposed precoder design. In all settings, the ISI duration $L = 5$, and the initial stepsize $\epsilon = 0.1$ is used in Algorithm 1. Unless explicitly pointed out, we simulate with the block-diagonalization precoder (19). The convergence of Algorithm 1 and sum-AIRs in different types of ISI channels are evaluated. The rates of dirty paper coding (DPC) [17] and orthogonal frequency division multiplexing with ZF precoder (OFDM-ZF) are also shown. The rate of OFDM-ZF without CP serves as an upper-bound for SC systems. However, with CP [18], the rate of OFDM-ZF decreases. Nevertheless, the comparison between OFDM-ZF and SC shows that the proposed precoder design is efficient since it renders significant signal-to-noise (SNR) gains over linear detection and approaches the rate of OFDM-ZF.

4.1 Convergence Speed

In Fig. 3, the sum-AIRs obtained from Algorithm 1 are normalized by those obtained with the full-complexity optimization over (13) through an interior-point algorithm[19]. The ISI channels have uniform power delay profiles (PDP), and all taps are IID complex Gaussian variables with zero mean. The convergence of Algorithm 1 is evaluated at an SNR of 10 dB. Note that the complexity of Algorithm 1 grows only linearly in N, and as can be seen in Fig. 3, Algorithm 1 converges fast in around 5 iterations (each iteration comprises Step 3-7).

4.2 IID Complex Gaussian Channel

Next, we simulate the sum-AIR for the same ISI channels as in Fig. 3. We evaluate the sum-AIR with different memory length ν for CS detectors. The case $\nu = 0$ is equivalent to

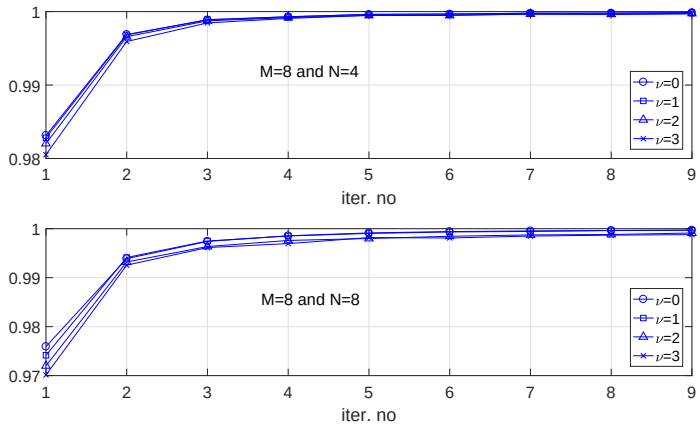


Figure 3: Convergence speed of Algorithm 1.

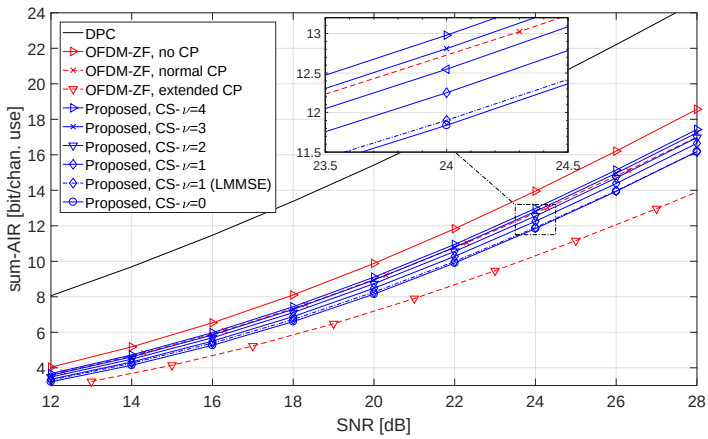


Figure 4: 4-antenna BS and 4 users in IID complex Gaussian ISI channel with a uniform PDP.

precoders optimized for LMMSE detection at receivers, which serves as the lower bound. As shown in Fig. 4, with an 4-antenna BS and 4 users, the sum-AIRs are improved in terms of SNR gain 0.5 dB with $\nu = 1$, and 1 dB with $\nu = 4$, both compared to the linear detection $\nu = 0$. We also show the sum-AIR for a mismatched design, where the precoder is designed for LMMSE detection, but the receivers apply CS detectors with $\nu = 1$. As can be seen in Fig. 4, the sum-AIR of such a mismatched design (blue-dashed curve) is inferior to that of the proposed precoder aimed for $\nu = 1$. In addition, DPC and OFDM-ZF without CP, perform better than the proposed precoders as expected. With CP, however, the rates of OFDM-ZF are inferior to SC systems with CS detection.

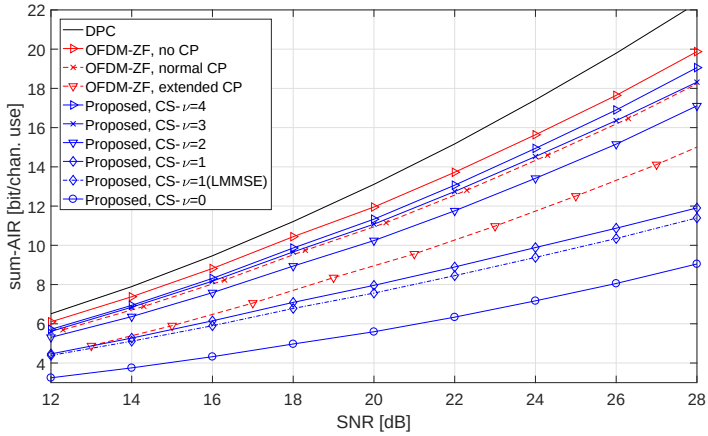


Figure 5: 8-antenna BS and 6 users in Proakis-C channel with a random phase rotation for each ISI link.

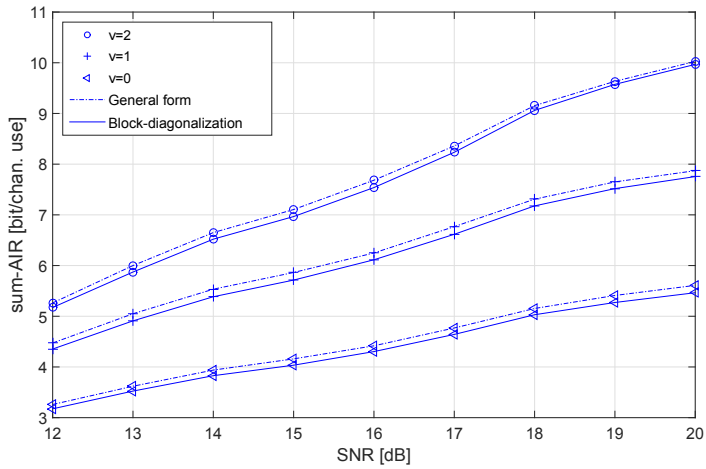


Figure 6: Repeat the test in Fig.5, and compare the sum-AIR for $\boldsymbol{F}$ with block-diagonal form (19) and the general form (5).

4.3 Proakis-C Channel

In Fig. 5, we test the case where each ISI link is Proakis-C channel [20], and is rotated independently with a random rotation $e^{j\theta}$, with θ being uniformly distributed over $[0, 2\pi)$. As shown in Fig. 5, with 8-antenna BS and 6 users, more than 4 dB SNR gain is obtained by increasing ν from 0 to 1, and with the proposed precoder design. For $\nu = 4$, the sum-AIR is close to the rates of OFDM-ZF without CP. Also note that, CS detectors with $\nu = 2$ have major gains over $\nu = 0$. This is so, since most of the channel power in Proakis-C channel is concentrated on three taps, and CS detection with $\nu = 2$ almost captures the entire power.

In Fig. 6, we show the sum-AIR losses with the block-diagonalization precoder compared to the fully-optimized general form (5). As can be seen, the rate losses are neglectable compared to the rate increments by considering CS detection with $\nu = 1$ other than the linear case $\nu = 0$.

5 Summary

We proposed precoder design for single-carrier systems with multi-user MIMO-ISI broadcasting channels under channel shortening detection at receivers. Compared to linear detection, significant gains can be observed when utilizing non-linear CS detectors, especially when the data transmission suffers from severe intersymbol interference. Furthermore, we provided an efficient per-user based optimization algorithm to solve the design problem of the block-diagonalization precoder, with a complexity that only grows linearly in the number of users.

References

[1] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "A Low-complexity channel shortening receiver with diversity support for evolved 2G device," IEEE Int. Conf. Commun. (ICC), May 2016.

[2] Y. Chen and L. M. Davis, "Single carrier filtering system architecture for flexible frequency domain multiplexing uplink," Int. Conf. Comm. Workshop (ICCW), pp. 1048-1053, Jun. 2015.

[3] W. H. Gerstacker, F. Obernosterer, R. Meyer, and J. B. Huber, "On prefilter computation for reduced-state equalization," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 793-800, Oct. 2002.

[4] W. Koch amd A. Baier, "Optimum and sub-optimum detection of intersymbol interference," IEEE Global Telecommun. Conf. (GLOBECOM), pp. 1679-1984, Dec. 1990.

[5] N. Al-Dhahir, "FIR channel-shortening equalizers for MIMO ISI channels," *IEEE Trans. Comm.*, vol. 49, pp. 213-218, Feb. 2001.

[6] C. Toker, S. Lambotharan, and J. A. Chambers, "Joint transceiver design for MIMO channel shortening," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3851-3866, Jul. 2007.

[7] D. Darsena, and F. Verde, "Minimum-mean-output-energy blind adaptive channel shortening for multicarrier SIMO transceivers," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5755-5771, Dec. 2007.

[8] D. Darsena, G. Gelli, L. Paura, and F. Verde, "Blind channel shortening for space-time-frequency block coded MIMO-OFDM systems," *IEEE Trans. Wireless Comm.*, vol. 11, no. 3, pp. 1022-1033, Mar. 2012.

[9] D. Darsena, G. Gelli, L. Paura, and F. Verde, "Blind channel shortening for asynchronous SC-IFDMA systems with CFOs," *IEEE Trans. Wireless Comm.*, vol. 12, no. 11, pp. 5529-5543, Nov. 2013.

[10] F. Rusek and A. Prlja, "Optimal channel shortening of MIMO and ISI channels," *IEEE Trans. Wireless Comm.*, vol. 11, no. 2, pp. 810-818, Feb. 2012.

[11] A. Modenini, F. Rusek, and G. Colavolpe, "Optimal transmit filters for ISI channels under channel shortening detection," *IEEE Trans. Comm.*, vol. 61, no. 12, pp. 4997-5005, Dec. 2013.

[12] W. Yu, "Uplink-downlink duality via minimax duality," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361-374, Feb. 2006.

[13] R. Hunger, M. Joham, and W. Utschick, "On the MSE-duality of the broadcast channel and the multiple access channel," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 698-713, Feb. 2009.

[14] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Commun. and Inform. Theory*, vol. 2, no. 3, pp. 155-239, 2006.

[15] W. Hirt, "Capacity and information rates of discrete-time channels with memory," Ph.D thesis, no. ETH 8671, Inst. Signal and Inf. Process., Swiss Federal Inst. Technol., Zürich, 1988.

[16] F. Charles, *An introduction to the calculus of variations*, Dover, 2010.

[17] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570-1580, Apr. 2005.

[18] 3GPP TS 36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," Release 13, Sep. 2015.

[19] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Programming*, vol. 107, no. 3, pp. 391-408, Jul. 2006.

[20] J. G. Proakis and M. Salehi, *Digital communications*, the fifth edition, McGraw-Hill, 2008.

Paper VII

Beyond Massive-MIMO: The Potential of Data-Transmission with Large Intelligent Surfaces

In this paper, we consider the potential of data-transmission in a system with a massive number of radiating and sensing elements, thought of as a contiguous surface of electromagnetically active material. We refer to this as a Large Intelligent Surface (LIS). The "LIS" is a newly proposed concept, which conceptually goes beyond contemporary massive MIMO technology, that arises from our vision of a future where man-made structures are electronically active with integrated electronics and wireless communication making the entire environment "intelligent".

Firstly, we consider capacities of single-antenna autonomous terminals communicating to the LIS where the entire surface is used as a receiving antenna-array in a perfect line-of-sight (LOS) propagation environment. Under the condition that the surface-area is sufficiently large, the received signal after a matched-filtering (MF) operation can be closely approximated by a sinc-function-like intersymbol interference (ISI) channel. Secondly, we analyze a normalized capacity measured per unit-surface, for a fixed transmit power per volume-unit with different terminal-deployments. As terminal-density increases, the limit of the normalized capacity [nats/s/Hz/volume-unit] achieved when wavelength λ approaches zero is equal to half of the transmit power per volume-unit divided by noise spatial power spectral density (PSD). Thirdly, we show that the number of independent signal dimensions that can be harvested per meter deployed surface is $2/\lambda$ for one-dimensional terminal-deployment, and π/λ^2 per square meter for two and three dimensional terminal-deployments. Lastly, we consider implementations of the LIS in the form of a grid of conventional antenna-elements, and show that the sampling lattice that minimizes the surface-area and simultaneously obtains one independent signal dimension for every spent antenna is the hexagonal lattice.

Based on: S. Hu, F. Rusek, and O. Edfors, "Beyond massive-MIMO: The potential of data-transmission with Large Intelligent Surfaces," submitted to *IEEE Trans. Signal Process.*, Nov. 2017.

1 Introduction

A Large Intelligent Surface (LIS) is an entirely new concept in wireless communication [1, 2], where we envision a future where man-made structures become more and more electronically active, with integrated electronics and wireless communication making the entire environment intelligent. The LIS concept can be seen as an extension of traditional massive multi-input multi-output (MIMO)[3–7], scaled up beyond the traditional large antenna-array concept. As an extension of traditional massive MIMO systems, LIS retains all the advantages such as allowing for an unprecedented focusing of energy in three-dimensional space which enables wireless charging, remote sensing with extreme precision and unprecedented data-transmissions. This makes it possible to fulfill the most grand visions for the next generation of communication systems and the concept of Internet-of-Things (IoT) [8, 9], where billions of devices are expected to be connected to the Internet. In Fig. 1, we show an example of three terminals communicating to a LIS in indoor and outdoor scenarios, respectively.

On the other hand, as a new concept beyond massive MIMO, there are also some substantial differences between the envisioned LIS and traditional massive MIMO systems. Firstly, LIS in its fundamental form uses the whole contiguous surface for transmitting and receiving. A more practical and implementation-friendly version of LIS approximates transmission and reception across a contiguous surface with an antenna-array spread across the same surface structure. We will investigate both these in our attempts to establish the fundamental limits of the LIS concept. Secondly, a traditional massive MIMO system is essentially thought of as a base-station so that users are in the far-field (the distance to the antenna-array is well beyond the Fraunhofer distance [10]). With LIS we consider a system that is so large that users that are reasonably close to it. As an example, one could think of an airport where the walls in the departure hall are covered with radiating antennas. In such a situation users are in the near field of the radiating structure, and for analysis purposes it is convenient to model the antennas as one contiguous surface. However, it is not only convenient, it is also: 1) a very close approximation when radiating antennas are packed tightly enough; 2) powerful since mathematical treatment is tractable; 3) General, since this represents the limit to what can be achieved with transmission for a given surface-area. Lastly, in contrast to traditional aperture antennas where the actual physical structure determines the electromagnetic radiation pattern of transmitted and received signals, with a LIS we can control the electromagnetic field on the entire surface and adapt signal transmission and receiving such as implementing a match-filtering (MF) procedure across the surface.

The LIS is also different from a traditional lens antenna arrays design aimed to reduce signal processing complexity and radio-frequency (RF) chain cost in millimeter wave (mm-Wave) communications with large MIMO systems [42–45]. The fundamental principle of lens antenna arrays is to provide variable phase shifting for electromagnetic rays at different

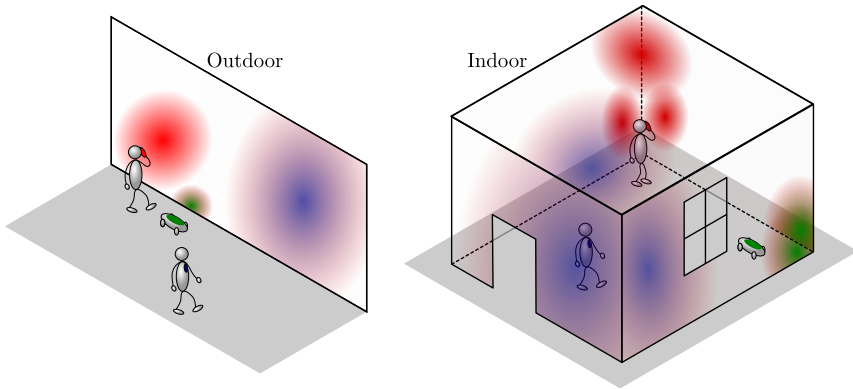


Figure 1: An example of three terminals communicating to a LIS in indoor and outdoor scenarios, respectively.

points on the lens aperture and steer the incident signals with sufficiently separated angle-of-arrivals (AoAs) to or angle-of-departures (AoDs) from different antenna subsets. The substantial difference between LIS and lenses is that, the contiguous surface of the lens is only used to steer signal, and the receiving device is still a traditional discrete MIMO system cascade connected with the lens.

The LIS introduces new properties, substantial gains and implementation challenges compared to massive MIMO systems. In this paper, we take a first look at the information-transfer capabilities of the LIS in the uplink (UL). For analytical tractability we assume an ideal situation where no scatterers or reflections are present, yielding a perfect line-of-sight (LOS) propagation environment, and each autonomous terminal is assumed to propagate an isotropic signal. Due to the reciprocity of the effective channel, the analysis can be straightforwardly applied to downlink (DL) data-transmission as well. Note that, these information-theoretical analysis do not coincide with those for traditional massive MIMO systems since with a LIS, users are inherently in the near-field. For scenarios where users are beyond the Fraunhofer distance of the LIS, our analysis does not extend much beyond those made for massive MIMO systems due to the fact that, the received signals at the LIS after MF process is band-limited and thusly the LIS can be sampled and approximated by a discrete large antenna-array system. Nevertheless, in both cases, the analysis with the LIS provides ultimate limits for data-transmission with a traditional large antenna-array and for a given surface-area.

Under the LOS assumption, we first analyze the normalized capacity $\hat{\mathcal{C}}$ per volume-unit in space, and show that its limit achieved when wavelength λ approaches zero is $\hat{P}/(2N_0)$ [nats/s/Hz/volume-unit], where $\hat{P}$ is the transmit power per volume-unit for different terminal-deployments, and N_0 is the spatial power spectral density (PSD) of additive white Gaussian noise (AWGN). Then, we also analyze the number of independent signal dimensions, i.e., the spatial degrees of freedom (DoF) [32, 35, 37–39], that can be harvested with a LIS, which is measured as the pre-log factor of $\hat{\mathcal{C}}$ when the deployed surface-area is infinitely large. Specially, we show that for an infinitely large LIS, $2/\lambda$ terminals can be spatially mul-

tiplexed per meter (m) deployed surface[1] for one-dimensional terminal-deployment, while π/λ^2 terminals can be spatially multiplexed per m^2 deployed surface-area for two and three dimensional terminal-deployments, respectively. We also demonstrate through numerical simulations that, with a fairly small LIS deployed in a medium sized room, around 100 terminals can be accommodated in the UL with only a minuscule per-terminal capacity loss compared to a case where only one terminal is present, due to effective interference suppressing of the LIS. Lastly, we also analyze optimal implementation of the LIS based on sampling theory [11, 12], and show that, the hexagonal lattice [13] minimizes the surface-area of the LIS while simultaneously obtaining one independent signal dimension for every spent antenna-element on the LIS. With the same number of independent signal dimensions achieved, the hexagonal lattice yields 23% surface-area saving over a rectangular lattice.

The rest of the paper is organized as follows. In Sec. II we describe the received signal model for LIS and introduce a sinc-function based approximation of the ISI channel after a match-filter (MF) procedure for analytical tractability. In Sec. III we analyze the capacities for both the optimal and MF terminals for one, two and three dimensional terminal-deployments, and put a special interest on the independent signal dimensions that can be harvest per unit surface. In Sec. IV we derive the optimal lattice that minimizes the surface-area of a LIS while achieving one independent signal dimension for every spent antenna. Numerical results are presented in Sec. V, and Sec. VI summarizes the paper.

Notation

Throughout this paper, superscripts $(\cdot)^{-1}$, $(\cdot)^{\frac{1}{2}}$, $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$, and $(\cdot)^\dagger$ stand for the inverse, matrix square root, complex conjugate, transpose, and Hermitian transpose, respectively. Boldface letters indicate vectors and boldface uppercase letters designate matrices. We also reserve $a_{m,n}$ to denote the element at the mth row and nth column of matrix $\boldsymbol{A}$, a_m to denote the mth element of vector $\boldsymbol{a}$, and $\boldsymbol{I}$ to represent the identity matrix. The operators '$\mathcal{R}\{\cdot\}$' and '$\mathrm{Tr}(\cdot)$' take the real part and the trace of the arguments, '$\star$' denotes linear convolution, and '$\mathbb{E}[\cdot]$' is the expectation operation.

2 Received Signal Model at LIS for Multiple Terminals

We consider the transmission from K autonomous single-antenna terminals located in a three-dimensional space to a two-dimensional LIS deployed on the xy-plane as shown in Fig. 2. Expressed in Cartesian coordinates, the LIS center is located at $x = y = z = 0$,

[1] We here assume a rectangular LIS and measure its size only by its length, while its height is assumed to extend infinitely.

while terminals are located at $z > 0$ and arbitrary x, y coordinates. For analytical tractability, we assume a perfect LOS propagation. The reason for focusing on the LOS case is four-fold: Firstly, it simplifies calculations to a level where pursuing analytic bounds on performance is tractable. Secondly, with a dense LIS deployment in an environment, most terminals will be close to at least one of the sub-surfaces constituting the LIS and LOS conditions will dominate. Thirdly, if the propagation environment is of the NLOS type, the additional scattering processes in the environment will enhance the ultimate performance, as compared to the LOS case, due to a higher dimensionality provided by the multipath propagation. Lastly, if all multi-path components (MPCs) are independent from each other in an NLOS environment, then each of them can be assumed as an independent LOS path and the capacity involves a summation over all MPCs [36]. As shown later, after MF process the effective channel can be modeled with a sinc-like function, and as long as the distance between two reflections are larger than $\lambda/2$ in the direction parallel to the LIS, the MPCs origin from these two reflections can be assumed orthogonal to each other. For these reasons, our analysis is analytically tractable, valid for LOS environments and will not over-estimate achievable performance in NLOS environments.

2.1 Narrow-band Received Signal Model at the LIS

Assuming that the kth terminal located at (x_k, y_k, z_k) transmits data symbols $u_k[m]$ with power P_k (per Hz), and $u_k[m]$ are independent Gaussian variables with zero-means and unit-variances. Denote λ as the wavelength and T as the symbol period, and consider a narrow-band system where the transmit times from terminals to the LIS are negligible compared to T which results in no temporal interference. The received baseband signal at the LIS location $(x, y, 0)$ corresponding to the kth terminal at time t cam be modeled as[2]

$$\tilde{s}_{x_k,y_k,z_k}(x,y,t) = s_{x_k,y_k,z_k}(x,y)\sqrt{P_k}\sum_{m=-\infty}^{\infty} u_k[m]\mathrm{sinc}_T\left(t-mT\right)+ n(x,y,t), \tag{1}$$

where 'sinc_T(t)' is a unit-energy sinc pulse with two-sided bandwidth $W = 1/T$ that equals

$$\mathrm{sinc}_T(t) = \frac{T}{\pi t}\sin(\pi t/T), \tag{2}$$

and the noise-term $n(x, y, t)$ is independent over locations $(x, y, 0)$ on the LIS, and modeled as wide-sense stationary (WSS) Gaussian process with zero-mean and a spatial PSD N_0 at each position on the LIS.

According to Fig. 2, the effective channel $s_{x_k,y_k,z_k}(x, y)$ can be modeled as

$$s_{x_k,y_k,z_k}(x,y) = \sqrt{\varepsilon_L\cos\phi(x,y)}\exp(-2\pi j f_c\Delta_k(x,y)), \tag{3}$$

[2]We omit the z-coordinate for both the received signals at the LIS and the noise terms in (1), since $z = 0$ across the whole LIS as shown in Fig. 2.

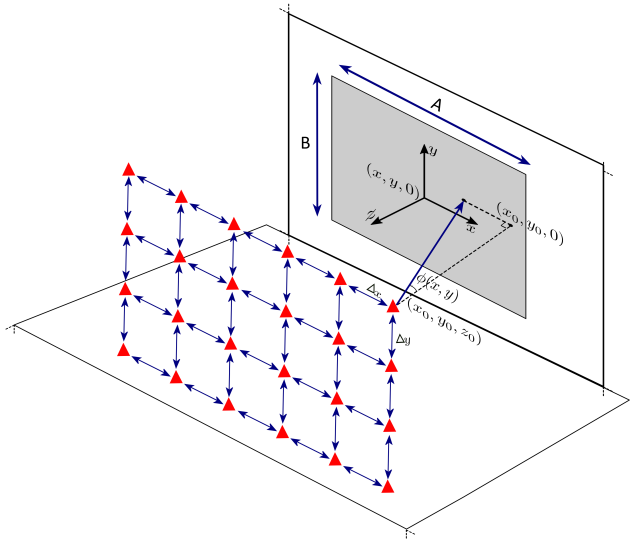


Figure 2: The radiating model of transmitting signal to the LIS with terminals deployed in front of the LIS. We integrate the received signal across the entire surface-area spanned by the LIS. Therefore, the received signal model (5) holds for both near-field and far-field scenarios with respect to the LIS.

where the parameters are defined as follows:

- η_k is the square of distance from the kth terminal to the location $(x, y, 0)$ at the LIS which equals

$$\eta_k = z_k^2 + (y - y_k)^2 + (x - x_k)^2. \tag{4}$$

- $\varepsilon_L = 1/(4\pi\eta_k)$ denotes the free-space attenuation [3].
- $\phi(x, y)$ is the AoA with $\cos\phi(x, y) = z_k/\sqrt{\eta_k}$.
- f_c is the carrier-frequency and c is the speed-of-light.
- $\Delta_k(x, y) = \sqrt{\eta_k}/c$ is the transmit time from the kth terminal to the location $(x, y, 0)$ at the LIS.

Inserting these parameters back into (3) yields an effective channel of the LIS corresponding to the kth terminal as

$$s_{x_k, y_k, z_k}(x, y) = \frac{\sqrt{z_k}}{2\sqrt{\pi}\eta_k^{3/4}} \exp\left(-\frac{2\pi j\sqrt{\eta_k}}{\lambda}\right). \tag{5}$$

Signal model (5), which has also been discussed in [2, Proposition 1], is more accurate than what is usually considered in traditional large antenna-array systems [21], where in the latter case terminals are assumed to be in the far-field and a planar-wave approximation is used in (3) such that the term $\cos\phi(x, y)$ is approximated by 1 and removed.

2.2 Received Signal Model for Multiple Terminals with MF Procedure

Based on (1), the received signal at the LIS location $(x, y, 0)$ comprising signals from all K terminals equals

$$r(x,y,t) = \sum_{k=0}^{K-1} \sum_{m=-\infty}^{\infty} s_{x_k,y_k,z_k}(x,y)\sqrt{P_k}u_k[m]\mathrm{sinc}_T\left(t-mT\right) + n(x,y,t). \quad (6)$$

Given received signal (6) across the LIS, optimum processing includes applying both a spatial and a temporal correlator to each transmit signal, a procedure we call "MF". The discrete received signal at sampling time mT after the MF process[3] is

$$\begin{aligned} r_k[m] &= \sqrt{P_k} \iint\limits_{(x,y)\in\mathcal{S}} s^*_{x_k,y_k,z_k}(x,y)\left(r(x,y,t) \star \mathrm{sinc}_T(t)\Big|_{t=mT}\right)\mathrm{d}x\mathrm{d}y \\ &= \sum_{\ell=0}^{K-1} \sqrt{P_kP_\ell}u_\ell[m] \iint\limits_{(x,y)\in\mathcal{S}} s_{x_\ell,y_\ell,z_\ell}(x,y)s^*_{x_k,y_k,z_k}(x,y)\mathrm{d}x\mathrm{d}y + w_k[m] \\ &= \sum_{\ell=0}^{K-1} \sqrt{P_kP_\ell}\phi_{k,\ell}u_\ell[m] + w_k[m], \end{aligned} \quad (7)$$

where $w_k[m]$ is the effective discrete noise after MF, and the coefficient

$$\phi_{k,\ell} = \iint\limits_{(x,y)\in\mathcal{S}} s_{x_\ell,y_\ell,z_\ell}(x,y)s^*_{x_k,y_k,z_k}(x,y)\mathrm{d}x\mathrm{d}y, \quad (8)$$

where $\mathcal{S}$ is the surface-area spanned by the two-dimensional LIS. As the received signal after MF is identical for all samples and there is no temporal interference, we omit the index m and assemble the notation in (7) into a matrix formulation as

$$\boldsymbol{r} = \boldsymbol{G}\boldsymbol{u} + \boldsymbol{w}, \quad (9)$$

where the (ℓ, k)th element of matrix $\boldsymbol{G}$ equals

$$g_{k,\ell} = \sqrt{P_\ell P_k}\phi_{k,\ell}, \quad (10)$$

which represents the received signal power when $k = \ell$ and the inter-user interference when $k \neq \ell$, respectively. Note that, with the MF process the noise variables are still zero-mean but colored with a covariance matrix

$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}] = N_0\boldsymbol{G}. \quad (11)$$

[3]The MF process can take place either just before the conversion from analog to digital depending on the specific front-end design, which is out of the scope of this paper. In general, the analog circuit implementation discussed is more power-efficient for shorter and faster MF process, and conversely, the digital circuit is more power efficient where the match filters are longer and slower[40].

In the rest of this paper, we assume equal terminal transmit powers (per Hz) $P_k = P$ and study the capability of the terminals to communicate with the LIS.

2.3 Independent Signal Dimensions Harvested with the LIS

As the channel capacity grows linearly with the number of spatial DoF for MIMO systems, it is of particular interest to evaluate the independent signal dimensions, i.e, the spatial DoF of the LIS systems that can be harvested. With statistical MIMO models, the authors in [35, 37] show that the number of independent signal dimensions is the minimum of the number of transmit and receive antennas for a traditional antenna-array. With LIS, however, for a certain deployed surface-area, the number of transmit antennas at the LIS can be reviewed as infinitely many. Let's consider packing single-antenna transmitters in the same size of deployed surface-area of the LIS, which can be seen as a large 2D transmit antenna-array that is of the same size. Obviously, the number of independent signal dimensions cannot go to indefinitely with packing more and more transmitters in a given surface-area, as the interferences among transmitters also become stronger as the number of transmitters increases. Therefore, under the condition that a total transmit power per m or m^2 is fixed, we next analyze the limit of the independent signal dimensions that can be obtained by the LIS as its surface-area goes to ∞.

With the received ISI signal model (9), the channel capacity $\mathcal{C}$ averaged by the number of terminals, in nats per channel use, equals[35]

$$\mathcal{C} = \frac{1}{K} \log \det \left(\boldsymbol{I} + \frac{\boldsymbol{G}}{N_0} \right). \tag{12}$$

The capacity $\mathcal{C}$ is also identical to the capacity in [nats/s/Hz] by noting that $\mathcal{C}/(TW) = \mathcal{C}$. Hence, in the rest of the paper, we always assume that $\mathcal{C}$ has the unit [nats/s/Hz], and pay no attention to the properties of $\mathcal{C}$ on time or frequency domains. Bearing this in mind, we put an emphasis on the number of independent signal dimensions per deployed area-unit of the LIS that is possible to harvest, which is calculated based on the capacity normalized with the total deployed surface-area; a quantity we refer to as $\hat{\mathcal{C}}$. Therefore, the capacity $\hat{\mathcal{C}}$ has the unit [nats/s/Hz/area-unit]. This can be interpreted as the number of available signal space dimensions per area-unit, in perfect analogy with Shannon's original ideas [20]. Reaching this space-normalized capacity $\hat{\mathcal{C}}$ in practice requires, of course, the K terminals to be (i) sufficiently many, and (ii) favorably located in space. In the next we define $\hat{\mathcal{C}}$ in detail.

For a one-dimensional terminal-deployment such as in Fig. 2, K terminals are uniformly deployed along a line that is in parallel to the LIS and with a spacing Δ_x between two adjacent terminals. As $K \to \infty$, the length of the terminal-deployment $K\Delta_x \to \infty$. We

then consider a rectangular shaped[4] LIS whose length grows at the same rate[5]. The space-normalized capacity $\hat{\mathcal{C}}$ [nats/s/Hz/m] is calculated as

$$\hat{\mathcal{C}} = \lim_{K\to\infty} \frac{K\mathcal{C}}{K\Delta_x} = \lim_{K\to\infty} \frac{\mathcal{C}}{\Delta_x}. \tag{13}$$

For two or three dimensional terminal-deployments where the spacings between two adjacent terminals are Δx and Δy for x and y dimensions[6] such as in Fig. 2, respectively, we also consider a rectangular LIS whose surface-area grows at the same rate when $K \to \infty$. Denote $\Delta_s = \Delta_x \Delta_y$, the space-normalized capacity $\hat{\mathcal{C}}$ [nats/s/Hz/m^2] is calculated as

$$\hat{\mathcal{C}} = \lim_{K\to\infty} \frac{K\mathcal{C}}{K\Delta_x\Delta_y} = \lim_{K\to\infty} \frac{\mathcal{C}}{\Delta_s}. \tag{14}$$

With $\hat{\mathcal{C}}$ defined in (13) and (14), respectively, the number of independent signal dimensions ρ is calculated as the pre-log factor, i.e., the high signal-to-noise ratio (SNR) slope of $\hat{\mathcal{C}}$,

$$\rho = \lim_{\hat{P}/N_0\to\infty} \frac{\hat{\mathcal{C}}}{\log\left(\hat{P}/N_0\right)}, \tag{15}$$

where $\hat{P}$ is the transmit power (per Hz) per volume-unit of the terminal-deployments, i.e.,

$$\hat{P} = \frac{P}{\Lambda}, \tag{16}$$

where $\Lambda = \Delta_x$ and Δ_s for one and two dimensional deployments, respectively. We point out that, $\hat{P}$ instead of P shall be used in (15) to calculate ρ. Otherwise, the normalized capacity $\hat{\mathcal{C}}$ becomes infinitely large when Λ is small for a given P.

We first point out that, although we use uniformly located terminals to derive the normalized capacity $\hat{\mathcal{C}}$ and the number of independent signal dimension ρ, these results are substantial and independent of the uniform-distribution assumption, due to the following two facts: Firstly, $\hat{\mathcal{C}}$ and ρ are obtained when the terminal-spacing goes to 0, that is, we pack as many terminals as possible in a unit-area, in which case the deployments of the terminals become irrelevant. Secondly, $\hat{\mathcal{C}}$ is maximized when the terminal-spacing is $\lambda/2$

[4]The shape of the LIS becomes irrelevant when the surface-area is infinitely large.

[5]In one-dimensional case, we consider normalizing the capacity $\mathcal{C}$ by the length of the LIS, i.e., $K\Delta_x$, for analyzing the number of independent signal dimensions ρ. Otherwise, if we normalized $\mathcal{C}$ by the surface-area, the number of independent signal dimensions ρ approaches zero when the width of the LIS also goes to infinity.

[6]For three-dimensional case we omit the spacings between terminals in z-dimension as we can ideally project all the K terminals to a xy-plane in front of the LIS for the purpose of analyzing the number of independent signal dimensions ρ, as what will become clear later in Sec. III-C.

and λ^2/π for 1D and 2D terminal deployment, respectively. Further decreasing the spacings will no longer increase $\hat{\mathcal{C}}$, due to the limits of independent signal dimensions. Hence, the uniform-distributions are actually optimal in the sense of reaching maximal normalized capacities.

Note that, the spatial DoF that can be harvested for a given surface-area limitation with deterministic channel has also been consider in [32] by assuming a discrete spherical antenna-array. However, there are several differences between our analysis and those in [32]. Firstly, the authors in [32] derive the spatial DoF by considering the solid angles subtended by the scattering clusters at both the transmit and receive arrays in the angular domain, whereas we assume a perfect LOS propagation environment without scattering clusters. Secondly, the analysis in [32] is carried out based on decomposition of the integral kernel function of the channel response on angular domain, whereas we derive the normalized capacity $\hat{\mathcal{C}}$ directly with the MF process and calculate the number of independent signal dimensions by taking the surface-area into ∞. Lastly, the analysis in [32] uses the far-field approximation, while with LIS the channel model holds both for near and far field, where we are in particular more interested in near filed properties.

2.4 Array Gain Considerations

Let us first consider the received signal power (per Hz) at the LIS from an omni-directional antenna with power P that is located at coordinates $x = y = 0$ and $z = z_0$, that is, z_0 meters from the LIS and perpendicular to its center. The received power (per Hz) at the LIS, according to (8) and (10), equals

$$g_{k,k} = P \iint\limits_{(x,y)\in\mathcal{S}} |s_{0,0,z_0}(x,y)|^2 \mathrm{d}x\mathrm{d}y = \zeta P, \tag{17}$$

where

$$\zeta = \frac{1}{4\pi} \iint\limits_{(x,y)\in\mathcal{S}} \frac{z_0}{\left(z_0^2 + x^2 + y^2\right)^{\frac{3}{2}}} \mathrm{d}x\mathrm{d}y. \tag{18}$$

Assuming a rectangular LIS with $-A \le x \le A$ and $-B \le y \le B$, then ζ equals

$$\zeta = \frac{1}{\pi} \tan^{-1}\left(\frac{AB}{z_0\sqrt{A^2 + B^2 + z_0^2}}\right). \tag{19}$$

Moreover, If one dimension of the LIS is much larger than the other dimension, e.g., $A \gg B$, the received power at the LIS reads

$$g_{k,k} = \frac{P}{\pi} \tan^{-1}\left(\frac{B}{z_0}\right). \tag{20}$$

Furthermore, if both dimensions of the LIS are asymptotically large, i.e., $A=B=\infty$, then it holds that $g_{k,k}=P/2$, which makes intuitive sense, since half of the isotropically transmitted power from the terminal reaches the LIS, and the other half propagates away from. This number should now be compared to the free-space attenuation ε_L that would result from a single receive antenna at distance z_0, which is typically many orders of magnitudes smaller than $P/2$. Thus we obtain, in addition to a possibly large value of independent signal dimensions, an impressive array gain.

3 Space-Normalized Capacities and Independent Signal dimensions

In this section, we take an information-theoretical analysis on signal model (9) for one, two and three dimensional deployments of the K terminals, and derive the number of independent signal dimensions that can be harvested with a LIS for a given transmit power per volume-unit. As from (8), working with LIS results in solving an integral to calculate $\phi_{k,\ell}$. However, for the cases $\ell\neq k$, closed-form solutions seem out of reach and we seek for close approximations. We first state the following property that can be used to approximate $\phi_{\ell k}$.

Property 1. *For sufficiently small λ, the integral*

$$g(\Delta)=\int_{-\infty}^{\infty}\left(1+x^2\right)^{-\frac{3}{4}}\left(1+(x+\Delta)^2\right)^{-\frac{3}{4}}\exp\left(-\frac{2\pi\jmath}{\lambda}\left[\sqrt{1+x^2}-\sqrt{1+(x+\Delta)^2}\right]\right)\mathrm{d}x, \tag{21}$$

can be well approximated by a sinc function[7]

$$g(\Delta)\approx 2\mathrm{sinc}\left(\frac{2\Delta}{\lambda}\right). \tag{22}$$

Argumentations leading to Property 1 are in Appendix A. To answer what is meant by "sufficiently small λ", we need also to take the distance z_k from the terminal to the LIS into account. From the argumentation in Appendix A, it can be observed that $\lambda/z_k\lesssim 1$. For wavelengths encountered in radio transmission, and reasonable distances from the surface, this condition is usually well satisfied. In Fig. 3 we show an example of calculating $g(\Delta)$ with the exact integral (21) and the approximation (22) for $\lambda=0.4$ m. As can be seen, the two curves are almost aligned with each other and the approximation errors are relatively small. With (22), we can then analyze the information-theoretical properties of the LIS in the next section.

[7]The sinc-function without any subscript denotes a standard sinc-function with unit-energy and a double bandwidth 1.

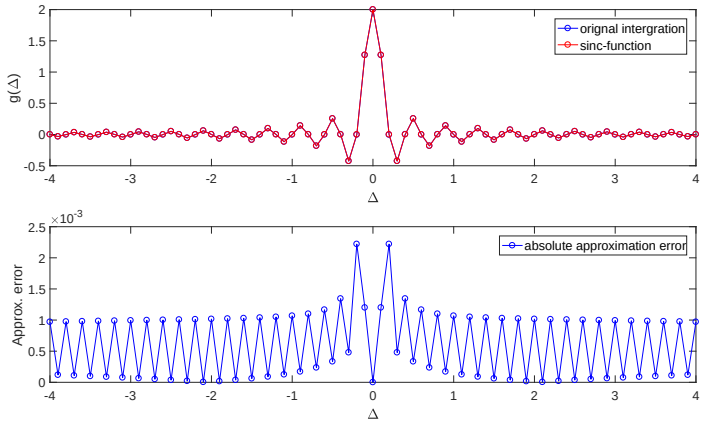


Figure 3: The approximation of integration (21) and the approximation errors for $g(\Delta)$ with $\lambda = 0.4$ m. As can be seen, the errors are relatively small compared to the maximum value $g(0)$, i.e., the received power of the considered terminal.

3.1 Capacity for One-Dimensional Case: Terminals on a Line

We start with the one-dimensional terminal-deployment and consider an infinitely long LIS with a rectangular shape with $-\infty \leq x \leq \infty$ and $-B \leq y \leq B$, where terminals are uniformly located along the line with coordinates $y = 0$ and $z = z_0$, with a spacing-distance Δ_x between two adjacent terminals[8] as shown in Fig. 2. For notational convenience, we first define the ratio between the half wavelength and the terminal-spacing in one-dimensional deployment as

$$\theta = \frac{\lambda}{2\Delta_x}. \tag{23}$$

As will be seen later, θ plays a key role in the following analysis.

From (9), the received signal at the LIS for the kth terminal can be expressed as

$$r_k = \sum_{\ell=0}^{K-1} g_{k,\ell} u_\ell + w_k, \tag{24}$$

where the noise variables w_k are zero-mean Gaussian variables with variances $\mathbb{E}[w_k^* w_\ell] = N_0 g_{k,\ell}$.

Property 2. *Using Property 1, the effective ISI channel tap $g_{k,\ell}$ is real and equals*

$$g_{k,\ell} = \frac{P}{\pi} \tan^{-1}\left(\frac{B}{z_0}\right) \operatorname{sinc}\left(\frac{k-\ell}{\theta}\right). \tag{25}$$

[8]Although an infinitely long wall and equi-distant terminal locations are unreasonable in practice, these assumptions are made for analytical tractability. General capacity results will be obtained, from which the insights of general capacity behavior can be concluded. Numerical results on LIS with finite sizes and random terminal positions will be given in Sec. VI, which are shown to be well predicted by the theoretical analysis.

Proof. See Appendix B. □

We point out that, when $k = \ell$, (25) is well aligned with (20) and the approximation in Property 1 is in fact exact.

Now considering applying an optimal receiver[9] on signal model (24), and the capacity $\mathcal{C}$ [nats/s/Hz] of each terminal can be calculated as [16, 19]

$$\mathcal{C} = \frac{1}{\theta} \int_{-\theta/2}^{\theta/2} \log \left(1 + \frac{G(f)}{N_0} \right) \mathrm{d}f, \tag{26}$$

where $G(f)$ is the frequency response of ISI channel $g_{k,\ell}$ in (24). Since $g_{k,\ell}$ are discrete samples of the sinc-function at a sampling rate θ, and by the Poisson summation formula [22], $G(f)$ can be expressed as

$$G(f) = \zeta P \theta \sum_{k=-\infty}^{\infty} G_0(f - k\theta), \tag{27}$$

and $G_0(f)$ is the standard rectangular function i.e., the Fourier transform of $\mathrm{sinc}(x)$.

Defining two useful auxiliary variables

$$\alpha = \frac{1}{\theta} - \beta \ \text{ and } \ \beta = \left\lfloor \frac{1}{\theta} \right\rfloor, \tag{28}$$

and with the definition in (16), the capacity (26) for the one-dimensional terminal-deployment is stated in Property 3.

Property 3. *With an infinitely long LIS in the direction where the terminals are deployed along a line with equal spacing, the capacity, with an optimal receiver, for each terminal is*

$$\mathcal{C} = \alpha \log \left(1 + \frac{(\beta+1)\lambda\zeta\hat{P}}{2N_0} \right) + (1-\alpha) \log \left(1 + \frac{\beta\lambda\zeta\hat{P}}{2N_0} \right). \tag{29}$$

Proof. See Appendix C. □

Whenever $\alpha = 0$, i.e., $1/\theta$ is an integer, from (29) the capacity equals

$$\mathcal{C} = \log \left(1 + \frac{\zeta P}{N_0} \right) \tag{30}$$

[9]We assume an optimal receiver for analyzing the capacity which implements a maximum a *posteriori* based detection, and the complexity is usually prohibitive in practical systems [14, 15]. For data demodulation, a particularly well suited low-complexity method is channel shortening (CS) [16–18].

which is the resulting capacity of a terminal if no other terminals are present and with an SNR equal to $\zeta P/N_0$. This is so since under such cases, $g_{k,\ell} = 0$ for $\ell \neq k$. We remark that, the analysis and discrete-time model of the one-dimensional case is identical to that of a faster-than-Nyquist (FTN) signaling system using a sinc-pulse [23, 24].

With the capacity $\mathcal{C}$ given in Property 3, we can obtain the space-normalized capacity $\hat{\mathcal{C}}$ defined in (13). By directly evaluating the limit as $\lambda \to 0$, we have the below corollary.

Corollary 1. *As $\lambda \to 0$, for any θ (i.e., Δ_x) the space-normalized capacity $\hat{\mathcal{C}}$ converges to $\zeta \hat{P}/N_0$ [nats/s/Hz/m].*

Instead of using an optimal receiver, we also consider capacity with MF receiver corresponding to model (24), and we summarize our findings in Property 4.

Property 4. *Under the same assumptions in Property 3, the capacity [nats/s/Hz] per-terminal with only the MF process applied in front is*

$$\mathcal{C} = \log\left(1 + \frac{\zeta P}{N_0 + I}\right), \tag{31}$$

where the interference power I equals

$$I = \zeta P\left(\theta^2\left(\beta^2 + 2\alpha\beta + \alpha\right) - 1\right). \tag{32}$$

Proof. See Appendix D. □

Note that from (32), under the cases that $1/\theta$ is an integer, the interference power $I = 0$ and the MF capacity (31) is the same as the capacity for the interference-free case. As seen from Property 4, the capacity depends on ζ and interference power I. The coefficient ζ is the power attenuation factor depending on the LIS surface-area and the distance z_0 from the terminals to the LIS, while the interference power I depends on variable θ. This is a natural result, since the interference power I between two terminals is determined by the spacing Δx normalized by the wavelength λ, which is $1/\theta$.

With the capacity $\mathcal{C}$ obtained with the optimal receiver in (29), from (15) it can be shown that

$$\rho = \begin{cases} 2/\lambda & \theta \geq 1 \\ 2\theta/\lambda & \text{otherwise.} \end{cases}$$

Therefore, the maximal number of independent signal dimensions per m is $2/\lambda$ for one-dimensional terminal deployments. When $1/\theta$ is an integer (or when λ is sufficiently small), the MF can also achieve the same same asymptotic slope of the normalized capacity curve $\hat{\mathcal{C}}$ from (31).

3.2 The Two-Dimensional Case: Terminals on a Plane

We next move on to the case that, terminals are located on a two-dimensional plane at $z = z_0$ which is in parallel to the LIS plane such as in Fig. 2. We are interested in the number of independent signal dimensions per m^2 deployed surface-area, and we therefore let $A, B \to \infty$ to avoid edge effects. In this case, $\zeta = 1/2$ for all z_0 and capacity does not depend on distance as a direct result from Property 2.

The first step is to study the spatial PSD of received signal $r(x, y, t)$ in the absence of noise. Technically, we look at the received signal after sinc-based matched-filtering only (i.e., the convolution in (7)), but not the spatial correlator (i.e. the integrals in (7)). The PSD is given by the two-dimensional Fourier transform [25] of the autocorrelation

$$g(\Delta_x, \Delta_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_{0,0,z_0}(x, y) s^*_{\Delta_x, \Delta_y, z_0}(x, y) \mathrm{d}x \mathrm{d}y. \tag{33}$$

Note that, as the LIS is infinitely long in both dimensions, only the distance

$$\tau = \sqrt{\Delta_x^2 + \Delta_y^2} \tag{34}$$

between two adjacent terminals matters for calculating $g(\Delta_x, \Delta_y)$. Under the approximation of Property 2, we have

$$g(\tau) \approx \frac{1}{2} \mathrm{sinc}\left(\frac{2\tau}{\lambda}\right). \tag{35}$$

As this function has radial symmetry, it follows that its Fourier transform is given by the Hankel transform [26] of degree zero, i.e.,

$$\begin{aligned} G(s) &= 2\pi \mathcal{H}_0 \left\{ g(\tau) \right\} \\ &= \pi \int_0^{\infty} \tau \mathrm{sinc}\left(\frac{2\tau}{\lambda}\right) J_0(2\pi s \tau) \mathrm{d}\tau \\ &= \begin{cases} \frac{\lambda}{4\pi} \left(\sqrt{\frac{1}{\lambda^2} - s^2} \right)^{-1}, & 0 \le s < \frac{1}{\lambda} \\ 0, & s > \frac{1}{\lambda} \end{cases} \end{aligned} \tag{36}$$

where $J_0(x)$ is the zeroth-order Bessel function of the first kind [27]. With the transmit power $\hat{P}$ per m^2 defined in (16), the space-normalized capacity $\hat{\mathcal{C}}$ [nats/s/Hz/m^2] equals

$$\begin{aligned} \hat{\mathcal{C}} &= \int_0^{2\pi} \int_0^{1/\lambda} s \log\left(1 + \frac{\hat{P}}{N_0} G(s)\right) \mathrm{d}s \mathrm{d}\theta \\ &= \pi \left(\frac{\log(1 + \lambda N)}{\lambda^2} + N^2 \log\left(\frac{N\lambda}{1 + N\lambda}\right) + \frac{N}{\lambda} \right), \end{aligned} \tag{37}$$

where

$$N = \frac{\lambda \hat{P}}{4\pi N_0}.$$

By directly evaluating the limit of (37) as $\lambda \to 0$, we obtain the below corollary.

Corollary 2. *As $\lambda \to 0$, the limit of space-normalized capacity $\hat{C}$ equals $\hat{P}/2N_0$ [nats/s/Hz/m^2], which is the same as the one-dimensional case with $B = \infty$ in Corollary 1.*

Then, the number of independent signal dimensions that can be harvested for the two-dimensional terminal-deployment can be computed by directly evaluating (15) with $\hat{C}$ in (37), which is stated in the below property.

Property 5. *The number of independent signal dimensions for the two-dimensional terminal-deployment is*

$$\rho = \frac{\pi}{\lambda^2}. \tag{38}$$

Thus, for every λ^2 deployed surface-area of the LIS, we obtain π independent signal dimensions.

3.3 The Three-Dimensional Case: Terminals in a Sphere

From the derivations for two-dimensional case in Sec. 3.2, we have already furnished for a solution of the dimensionality for the three-dimensional terminal-deployment. Consider the Fourier transform $S_{x_0,y_0,z_0}(f_1, f_2)$ of a signal $s_{x_0,y_0,z_0}(x, y)$. From the convolutional property of Hankel transforms [26], it follows that $G(s)$ in (36) is given by

$$G\left(\sqrt{f_1^2 + f_2^2}\right) = \left|S_{x_0,y_0,z_0}(f_1, f_2)\right|^2. \tag{39}$$

Since $G(s)$ in (36) does not depend on z_0, (39) implies that, the domain of $S_{x_0,y_0,z_0}(f_1, f_2)$ is independent of the distance z_0 from the wall. Since the number of independent signal dimensions that can be accommodated is proportional to the area of the domain of $S_{x_0,y_0,z_0}(f_1, f_2)$, it follows that the same number of dimensions is obtained in the three-dimensional case as in the two-dimensional case.

An alternative way to realize this result is to consider a hyper plane $\mathcal{P} = \{x, y, z : z = z_0\}$ for some small z_0. All signals transmitted from terminals at $z_k > z_0$ has to pass the plane $\mathcal{P}$. From the Huygens-Fresnel principle [28] it, however, follows that the signal that reaches the LIS can be expressed as point sources at the plane $\mathcal{P}$ that radiate the signals arrived $\mathcal{P}$ from the terminals. However, the number of signal space dimensions at the plane $\mathcal{P}$ is π/λ^2 per m^2, which means that the number of dimensions in the three-dimensional volume is unaltered compared to the two-dimensional case.

4 Implementing the LIS based on Sampling Theory

We have seen in Sec. 3 that, the received signal at the LIS has a two-dimensional Fourier transform that is band-limited to a disc of radius $1/\lambda$. A direct consequence is that, there is no loss if the received signal $\tilde{s}_{x_0,y_0,z_0}(x,y)$ is sampled sufficiently dense so that no aliasing occurs. Thus, a LIS can be implemented as a grid of discrete antenna-elements. In this section we take a look at optimal sampling of the LIS. As we deal with LISs with unbounded physical dimensions, we make use of lattice theory [11, 12, 29].

Before proceeding to the discussions on the sampling of LIS, it is of importance to emphasis the difference between LIS and traditional large antenna-array systems. LIS, in its fundamental form, uses the whole contiguous surface for transmitting and receiving signals. Therefore, LIS provides ultimate limits (both physical and theoretical) for a traditional large antenna-array system that packs as many antenna-elements as possible within a given surface-area. Moreover, with traditional massive MIMO systems, the users are usually in the far-field and the received power is assumed to be the same for all antenna-elements (in perfect LOS case), whereas with LIS, our signal model holds for both near-field and far-field cases. However, the LIS can also be approximated by a traditional large antenna-array system from two different aspects. A first and important aspect is that, the received signal at the LIS is band-limited after MF process and a discrete-sampling can be implemented according to such a nice property. The other aspect is that, for a massive MIMO systems equipped with millimeter (mm) or Terahertz-band wave communications [33, 34, 38, 41, 45], the spacing between two adjacent antenna-element is rather small and the deployment of antenna-elements are getting denser. The latter one is a natural outcome, and we put an interest on exploiting the first aspect, that is, analyzing the optimal sampling of the LIS.

With LIS, there are two possible objectives for designing the sampling: (i) One view of optimal sampling is that the antenna units are costly, while area is available in excess. With this view, we should constrain ourselves to obtain one signal space dimension for every spent antenna. Once this constraint is met, we should then find the sampling lattice that minimizes the area. (ii) An alternative view is that, as established in Sec. 3 a LIS can at most offer π/λ^2 dimensions per m^2 deployed surface-area, one can constrain the sampled LIS to offer the same number of dimensions per m^2, and then to ask for the least number of antenna elements per m^2 that meets the constraint. With this view, the physical resource to be minimized is area, while antenna units are considered cheap. Views (i) and (ii) are similar, and in this paper we treat (i) in depth. We point out that for (ii), the resulting lattice problem to be faced is to find the densest lattice whose fundamental cell circumscribes a circle of given radius [30].

Suppose that sampling of the LIS is made on the basis of the sampling matrix $\boldsymbol{S}$. With

that, the asymptotic number of placed antennas per m^2 is

$$A_{\mathrm{d}} = \frac{1}{|V(\boldsymbol{S})|}, \tag{40}$$

where $V(\boldsymbol{S})$ is the fundamental cell of a lattice generated from $\boldsymbol{S}$, and $|V(\boldsymbol{S})|$ is its fundamental volume. Let $\tilde{s}_{m,n}$ denote the samples of $\tilde{s}_{x_0,\,y_0,\,z_0}(x,y)$ generated from sampling at

$$\begin{bmatrix} x \\ y \end{bmatrix} = \boldsymbol{S} \begin{bmatrix} m \\ n \end{bmatrix}.$$

The samples $\{\tilde{s}_{m,n}\}$ have a Fourier transform $\tilde{S}(f_1,f_2)$ that is defined on the fundamental volume $V(\boldsymbol{S}^{-T})$ which is the reciprocal lattice corresponding to $V(\boldsymbol{S})$. The capacity per antenna for the sampled LIS is then given by[31, Theorem II.2]

$$C_{\mathrm{ant}} = \frac{1}{|V(\boldsymbol{S}^{-T})|} \int_{V(\boldsymbol{S}^{-T})} \log\left(1 + \frac{\hat{P}}{N_0|V(\boldsymbol{S})|}|\tilde{S}(f_1,f_2)|^2\right) \mathrm{d}f_1 \mathrm{d}f_2. \tag{41}$$

Recalling the band-limited structure of $S_{x_0,\,y_0,\,z_0}(f_1,f_2)$, the support of $\tilde{S}(f_1,f_2)$ is limited to

$$\tilde{\boldsymbol{V}}(\lambda,\boldsymbol{S}) = D(\lambda^{-1}) \cap V(\boldsymbol{S}^{-T}), \tag{42}$$

where $D(\lambda^{-1})$ denotes a disc with radius $1/\lambda$. Thus, from (41) we have that

$$C_{\mathrm{ant}} = \frac{1}{|V(\boldsymbol{S}^{-T})|} \int_{\tilde{\boldsymbol{V}}(\lambda,\boldsymbol{S})} \log\left(1 + \frac{\hat{P}}{N_0|V(\boldsymbol{S})|}|\tilde{S}(f_1,f_2)|^2\right) \mathrm{d}f_1 \mathrm{d}f_2. \tag{43}$$

Similar to (15), the number of independent signal dimensions that we can harvest per spent antenna is defined as

$$\rho_{\mathrm{ant}} = \lim_{\hat{P}/N_0 \to \infty} \frac{C_{\mathrm{ant}}}{\log(\hat{P}/N_0)}. \tag{44}$$

It can be readily verified (see Appendix E) that this limit is given by

$$\rho_{\mathrm{ant}} = \frac{|V(\boldsymbol{S}^{-T})|}{|\tilde{\boldsymbol{V}}(\lambda,\boldsymbol{S})|}, \tag{45}$$

and attains a maximum value $\rho_{\mathrm{ant}} = 1$ whenever $\tilde{\boldsymbol{V}}(\lambda,\boldsymbol{S}) = V(\boldsymbol{S}^{-T})$, that is, $V(\boldsymbol{S}^{-T}) \subseteq D(\lambda^{-1})$.

Let us now return to path (i). To satisfy the constraint of a maximum number of harvested dimensions per spent antenna, we know that we must sample with a lattice generator $\boldsymbol{S}$

satisfying $V(\boldsymbol{S}^{-T}) \subseteq D(\lambda^{-1})$. To then minimize area, we should choose the lattice generator $\boldsymbol{S}$ such that its fundamental volume is minimized, i.e., the problem to be addressed is

$$\begin{aligned} &\min_{\boldsymbol{S}} |V(\boldsymbol{S})| \\ &\text{such that } V(\boldsymbol{S}^{-T}) \subseteq D(\lambda^{-1}). \end{aligned} \tag{46}$$

We summarize the solution to this problem in the following property.

Property 6. *The lattice generator that solves (46) is the scaled hexagonal lattice generator*

$$\boldsymbol{S} = \begin{bmatrix} \frac{2\lambda}{3} & \frac{\lambda}{3} \\ 0 & \frac{\lambda}{\sqrt{3}} \end{bmatrix}. \tag{47}$$

Proof. See Appendix E. □

For this generator we have $\rho_{\text{ant}} = 1$ and an antenna density per m^2 that equals

$$A_{\text{d}} = \frac{1}{\det(\boldsymbol{S})} = \frac{3\sqrt{3}}{2\lambda^2}. \tag{48}$$

Let us now compare the hexagonal lattice in (47) with a LIS sampling according to a rectangular lattice generator. The most natural one to choose would be the generator

$$\boldsymbol{S} = \begin{bmatrix} \frac{\lambda}{2} & 0 \\ 0 & \frac{\lambda}{2} \end{bmatrix}, \tag{49}$$

which fails to meet the constraint $V(\boldsymbol{S}^{-T}) \subseteq D(\lambda^{-1})$. In fact, the generator for the densest rectangular lattice that meets the constraint, and thus results in $\rho_{\text{ant}} = 1$, is

$$\boldsymbol{S} = \begin{bmatrix} \frac{\lambda}{\sqrt{2}} & 0 \\ 0 & \frac{\lambda}{\sqrt{2}} \end{bmatrix}. \tag{50}$$

However, for this lattice generator we have $A_{\text{d}} = 2/\lambda^2$. Compared to (48), we see that a hexagonal sampling of the LIS requires only a fraction $4/(3\sqrt{3}) \approx 0.77$ of the surface-area required with a rectangular sampling, that is, 23% of the surface-area can be saved. Note that, the efficiency gain of the hexagonal sampling exceeds the normal packing gain $\pi/(2\sqrt{3}) \approx 0.91$ of the hexagonal lattice, this being a result of our constraint that for each spent antenna, we should be able to obtain one signal space dimension.

5 Numerical Results

In this section, we present simulation results to illustrate the information-theoretical properties for data-transmission with the LIS discussed in previous sections. In what follows, the mentioned noise PSD and transmit powers are linear values.

5.1 Capacities with LIS for One-Dimensional Terminal-deployments

In Fig. 4, we evaluate the space-normalized capacity $\hat{\mathcal{C}}$ for one-dimensional terminal-deployment. We plot $\hat{\mathcal{C}}$ [nats/s/Hz/m] obtained with the optimal receiver with $N_0 = 1, \zeta = 0.1$, and $\hat{P} = 10$, and for different values of Δ_x and λ. As can be seen, as $\lambda \to 0$, $\hat{\mathcal{C}}$ converges to a limit 1, which is aligned with Corollary 1.

In Fig. 5, we evaluate the differences of space-normalized capacities $\hat{\mathcal{C}}$ between the optimal and the MF receivers for $N_0 = 0.05, \zeta = 0.5$, $\tilde{P} = 40$, and with different values of Δ_x and λ. As can be seen, whenever $1/\theta$ is an integer, terminals do not interfere with each other and the normalized capacities $\hat{\mathcal{C}}$ of the optimal and the MF receivers are identical. Otherwise, the MF receiver is inferior to the optimal receiver as expected.

In Fig. 6 we also depict $\bar{\mathcal{C}}$ obtained with the MF as a function of terminal-distance Δx, with peaks attained when $1/\theta$ is an integer.

In Fig. 7, we evaluate the space-normalized capacity $\hat{\mathcal{C}}$ for randomly allocated terminals draw from a uniform distribution along a 10 m long line with different values of Δ_x, where $1/\Delta_x$ representing the density of random allocations, i.e., in L meters, we have L/Δ_x users randomly located. As can be seen, as Δ_x decreases to 0, the space-normalized capacity $\hat{\mathcal{C}}$ reaches the capacity limit with the optimal receiver and starts to saturate at $\Delta_x = \lambda/2 = 0.1$ m. With the MF receiver, the capacity also converges but is suboptimal.

5.2 Capacities with LIS for Two and Three Dimensional Terminal-deployments

Next, we evaluate the capacities for two and three dimensional terminal-deployments. In Fig. 8, we evaluate the space-normalized capacity $\hat{\mathcal{C}}$ for uniformly random located terminals in a two-dimensional plane with length and width both equal to 20 m. The locations of terminals are also drawn from a uniform distribution for a given terminal-density $1/\Delta_s$. As can be seen, when Δ_s decreases to 0, $\hat{\mathcal{C}}$ reaches a limit and starts to saturate at $\Delta_s = \lambda^2/\pi$ with the optimal receiver. With the MF receiver, the capacity also converges but is inferior to the optimal receiver similar to the conclusions drawn from the one-dimensional case.

In Fig. 9, we evaluate the three-dimensional case, where we consider a room with length,

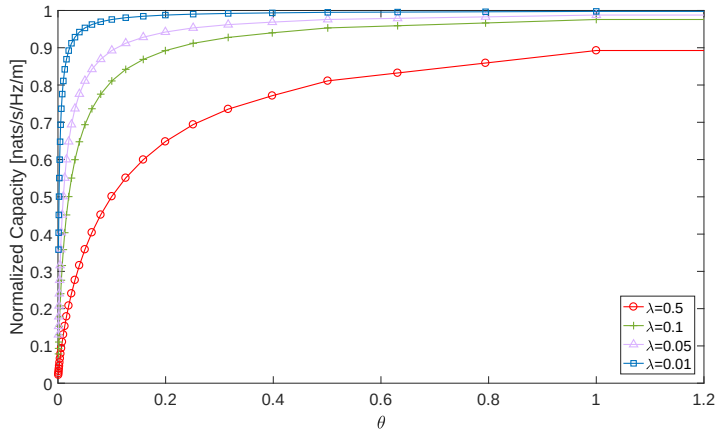


Figure 4: The space-normalized capacity $\hat{\mathcal{C}}$ in relation to θ for optimal receiver with $N_0 = 1, \eta = 0.1$, and $\hat{P} = 10$.

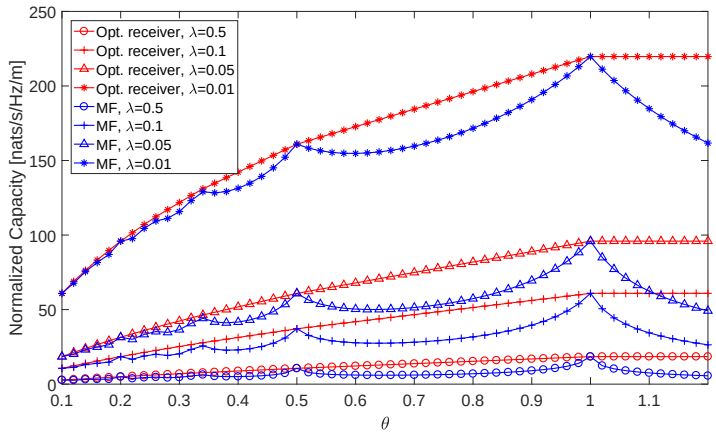


Figure 5: The space-normalized capacity $\hat{\mathcal{C}}$ for the optimal and the MF terminals with $N_0 = 0.05, \eta = 0.5$, and $\hat{P} = 40$.

width and height all equal to 4 m. For simplicity, we do not account for any reflections. On the front wall of the room, we assume a rectangular LIS, with length 2 m and width 1 m, deployed in the middle. For instance, we can use a white-board in a room as the LIS. Since we have a LIS with finite size, we use numerical computations to calculate the elements $\phi_{k,\ell}$ instead of using the sinc-function approximation. We evaluate the space-normalized capacity $\hat{\mathcal{C}}$ for randomly located terminals drawn from a uniform distribution in the room with different terminal-density $1/\Delta_v$, and consider two different cases. The first case is that, we fix the transmit power of each terminal to be $P = 10$ and then measure the capacity $\mathcal{C}$ per terminal. The other case is that, we fix the transmit power per m^3 to $\hat{P} = P/\Delta_v = 10$ (similar as the definition in (16)) and estimate the space-normalized capacity $\hat{\mathcal{C}}$ per m^3. As can be seen, when Δ_v decreases to 0, $\hat{\mathcal{C}}$ increases both for the optimal and MF terminals, like in the one and two dimensional cases. The capacity $\mathcal{C}$, however, is fairly flat when the number of terminals increases from 32 to 320, while the latter one

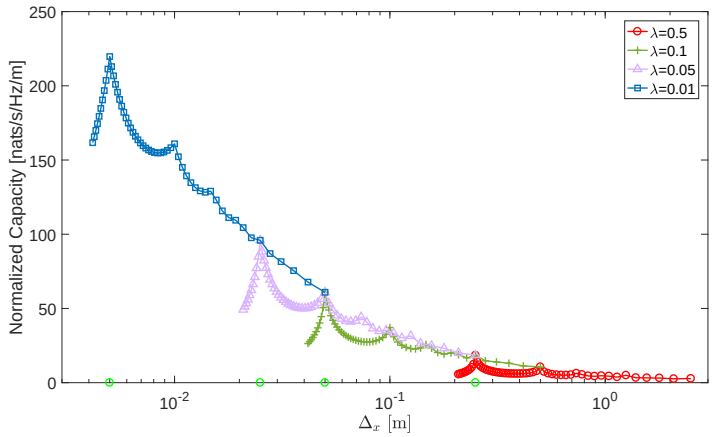


Figure 6: The same test as in Fig. 5. The normalized capacity in relation to Δx. The green-circles correspond to $\theta = 1$ for different values of λ.

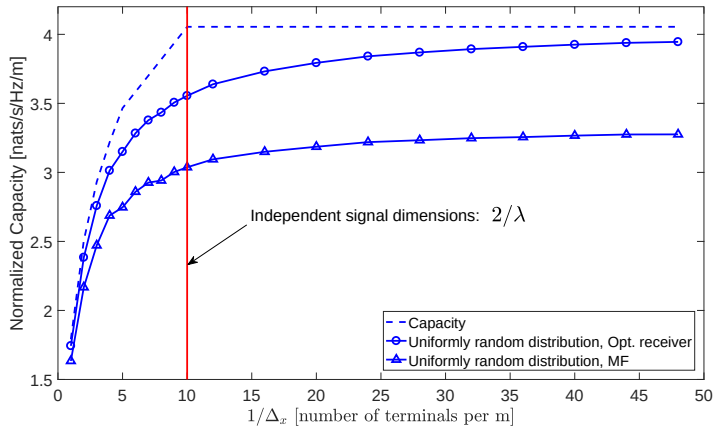


Figure 7: The space-normalized capacity $\hat{\mathcal{C}}$ with randomly located terminals along a line with length 10 m. We assume that $A = B = \infty$, $N_0 = 1$, $\hat{P} = 10$ and $\lambda = 0.2$ m.

results in more interferences among terminals. This clearly shows the potential of the LIS for interference suppression in data-transmission.

6 Summary

In this paper, we have considered using large intelligent surfaces (LIS) as large antenna-array systems for data-transmissions with multiple single-antenna autonomous terminals. We have shown that under the constraint that the transmit power per volume-unit $\hat{P}$ is fixed, the limit of a space-normalized capacity $\hat{\mathcal{C}}$ per volume-unit is $\hat{P}/(2N_0)$ when the wavelength λ approaches zero. We have also derived that the numbers of independent signal dimensions can be harvested for different terminal-deployments, which are shown

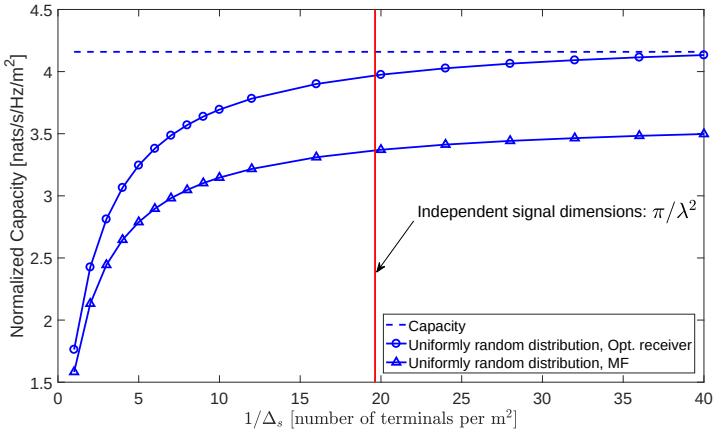


Figure 8: The space-normalized capacity $\hat{\mathcal{C}}$ with randomly located terminals on a plane with both length and width equal to 20 m. We assume that $A = B = \infty$, $N_0 = 1$, $\hat{P} = 10$ and $\lambda = 0.4$ m.

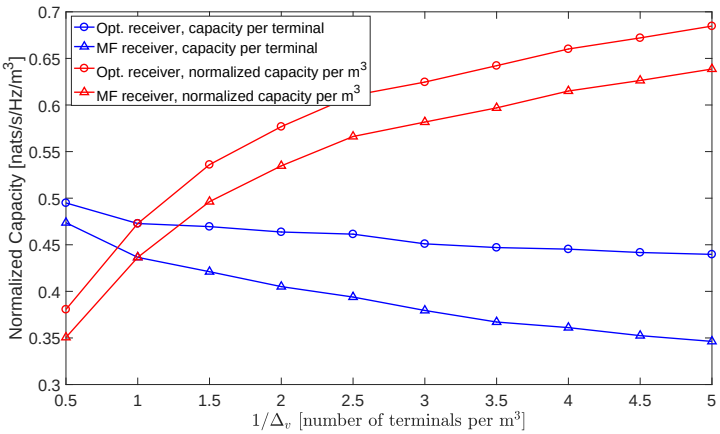


Figure 9: The space-normalized capacity $\hat{\mathcal{C}}$ of randomly located terminals in a room with length, width and height are all equal to 4 m. We assume that $A = 2$ m, $B = 1$ m, $N_0 = 1$, $\hat{P} = 10$ or $P = 10$ and $\lambda = 0.5$ m.

to be $2/\lambda$ per meter (m) for one-dimensional terminal-deployment, and π/λ^2 per m^2 for two and three dimensional cases. We have also analyzed the optimal sampling lattice for designing the LIS in a practical system based on sampling theory and shown that, the hexagonal lattice is optimal for minimizing the surface-area of a LIS under the constraint that one independent signal dimension should be obtained per spent antenna. In addition, we shown through numerical results that the LIS provides robust performances when the number of terminals increases and is highly effective in interference suppressing, which makes it a promising direction of research for data-transmission in communication systems beyond massive-MIMO.

Appendix A: Argumentations of Property 1

We first define a function φ as

$$\varphi(x) = \left(1+x^2\right)^{-\frac{3}{4}} \exp\left(-\frac{2\pi\jmath}{\lambda}\sqrt{1+x^2}\right). \tag{51}$$

Then the function $g(\Delta)$ can be written as

$$g(\Delta) = \int_{-\infty}^{\infty} \varphi(x)\varphi^*(x+\Delta)\mathrm{d}x = \varphi(\Delta) \star \varphi^*(\Delta). \tag{52}$$

To show that $g(\Delta)$ is close to a sinc-function, we first need to show that the Fourier transform of $\varphi(x)$ is close to a brick-shape. The Fourier transform $\Phi(f)$ is

$$\begin{aligned}\Phi(f) &= \int_{-\infty}^{\infty} \varphi(x)\exp(-2\pi\jmath f x)\,\mathrm{d}x \\ &= 2\int_{0}^{\infty} \left(1+x^2\right)^{-\frac{3}{4}} \exp\left(-\frac{2\pi\jmath}{\lambda}\sqrt{1+x^2}\right)\cos\left(2\pi f x\right)\mathrm{d}x. \end{aligned}\tag{53}$$

Noticing that the following Fourier cosine transforms (FCTs) hold

$$\int_0^{\infty} \left(1+x^2\right)^{-\frac{1}{2}} \exp\left(-\frac{2\pi\jmath}{\lambda}\sqrt{1+x^2}\right)\cos\left(2\pi f x\right)\mathrm{d}x = K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right), \tag{54}$$

$$4\int_0^{\infty} K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)\cos\left(2\pi f x\right)\mathrm{d}f = \left(1+x^2\right)^{-\frac{1}{2}} \exp\left(-\frac{2\pi\jmath}{\lambda}\sqrt{1+x^2}\right), \tag{55}$$

we then have

$$\begin{aligned}\Phi(f) &= 2\int_0^{\infty} \left(1+x^2\right)^{-\frac{3}{4}} \exp\left(-\frac{2\pi\jmath}{\lambda}\sqrt{1+x^2}\right)\cos\left(2\pi f x\right)\mathrm{d}x \\ &= 8\int_0^{\infty}\int_0^{\infty} \left(1+x^2\right)^{-\frac{1}{4}} K_0\left(2\pi\sqrt{\xi^2-\lambda^{-2}}\right)\cos\left(2\pi\xi x\right)\cos\left(2\pi f x\right)\mathrm{d}\xi\mathrm{d}x \\ &= 4\int_0^{\infty} K_0\left(2\pi\sqrt{\xi^2-\lambda^{-2}}\right)\left(\Phi_c\big(|f+\xi|\big)+\Phi_c\big(|f-\xi|\big)\right)\mathrm{d}\xi \\ &= 4\left(K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)\star \Phi_c\big(|f|\big)\right), \end{aligned}\tag{56}$$

where $\Phi_c(f)$ is the FCT of $\left(1+x^2\right)^{-\frac{1}{4}}$, which is an even functions and for $f \geq 0$ equal to

$$\Phi_c(f) = \int_0^{\infty} \left(1+x^2\right)^{-\frac{1}{4}} \cos(2\pi f x)\mathrm{d}x = \frac{(\pi/f)^{\frac{1}{4}}}{\Gamma\left(\frac{1}{4}\right)} K_{\frac{1}{4}}(2\pi f). \tag{57}$$

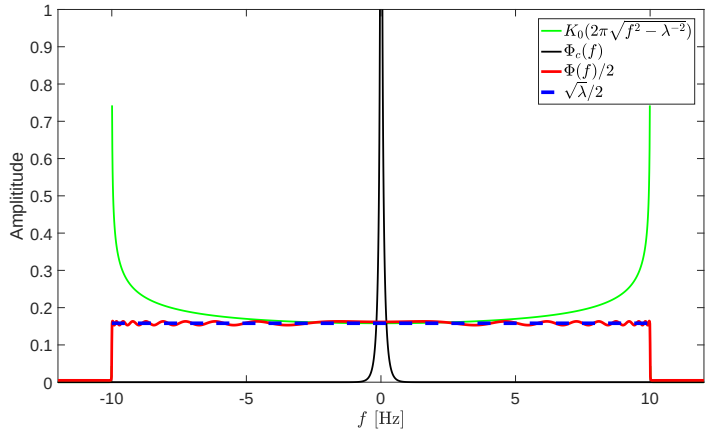


Figure 10: The function values of $K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)$ and $\Phi_c(f)$ and the sinc-function approximation of $\Phi(f)$ for $\lambda=0.1$. Note that $f=0$ is a singularity point for $\Phi_c(f)$ and $f=\pm\frac{1}{\lambda}$ are singularity points for $K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)$.

The functions $K_0(f)$and $K_{\frac{1}{4}}(f)$ are the modified Bessel function of the second kind [27]. A closed-from expression of the convolution in (56) seems out of reach and we have to seek an approximation of it. Firstly, noticing that the amplitude of the modified Bessel function $K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)$ is lower-bounded by a rectangular function as

$$\left|K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)\right| \approx \begin{cases} \frac{\sqrt{\lambda}}{2}, & -\frac{1}{\lambda} < f < \frac{1}{\lambda} \\ 0, & \text{otherwise.} \end{cases} \tag{58}$$

Secondly, $\Phi_c(f)$ can be approximated with a Dirac delta-function when λ is small[10] as

$$\Phi_c(f) \approx \frac{1}{2}\delta(f). \tag{59}$$

Then, it holds that

$$\begin{aligned} \Phi(f) &\approx 2\int_{-\infty}^{\infty} K_0\left(2\pi\sqrt{\xi^2-\lambda^{-2}}\right)\delta(f+\xi)\mathrm{d}\xi \\ &= 2K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right). \end{aligned} \tag{60}$$

From (52), the Fourier transform of $g(\Delta)$, which is denoted as $G(f)$, equals

$$G(f) = |\Phi(f)|^2 \approx \begin{cases} \lambda, & -\frac{1}{\lambda} < f < \frac{1}{\lambda} \\ 0, & \text{otherwise} \end{cases} \tag{61}$$

[10]In order to approximate $\Phi_c(f)$ by a Dirac delta-function in (59), the bandwidth of $K_0\left(2\pi\sqrt{f^2-\lambda^{-2}}\right)$ should be much larger than that of $\Phi_c(f)$, that is to say, λ should not be too small. As shown in the Fig. 11, the sinc-function approximation of $\Phi(f)$ works well with λ up to 1 m.

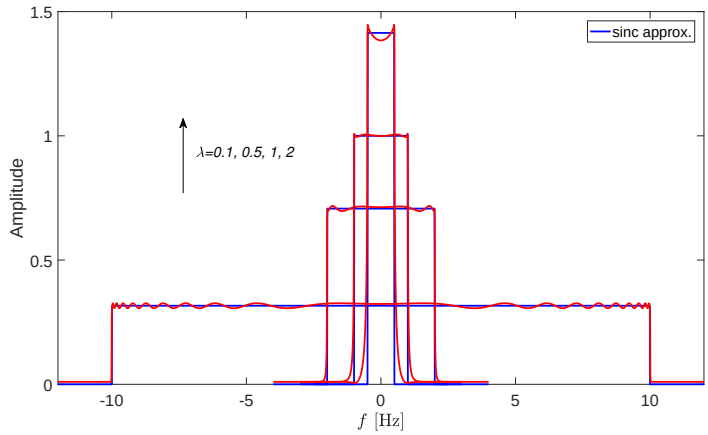


Figure 11: The numerical integrals and approximations with the sinc-function of $\Phi(f)$ for different values of λ.

which is also a rectangular function and the inverse Fourier transform is

$$g(\Delta) \approx 2\text{sinc}\left(\frac{2\pi\Delta}{\lambda}\right).$$

In Fig. 11, we plot the numerical computation of $\Phi(f)$ and the approximations, where the capacities given by the PSD of $|\Phi(f)|^2$ computed as $\int_{-\infty}^{\infty} \log\big(1+|\Phi(f)|^2\big)\mathrm{d}f$, are equal to $[1.9269, 1.6383, 1.4044, 1.1334]$ nats/s/Hz for $\lambda = [0.1, 0.5, 1, 2]$ m, while the capacities with sinc-function based approximations computed as $\int_{-\infty}^{\infty} \log\big(1+|\text{sinc}(f)|^2\big)\mathrm{d}f$, are equal to $[1.9053, 1.6178, 1.3794, 1.0876]$ nats/s/Hz, which are close and slightly smaller.

Appendix B: Proof of Property 2

From (10), it holds that

$$\begin{aligned} g_{k,\ell} &= P \iint\limits_{(x,y)\in\mathcal{S}} s^*_{x_k,y_k,z_k}(x,y)s_{x_\ell,y_\ell,z_\ell}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \frac{z_0 P}{4\pi} \int_{-B}^{B}\int_{-\infty}^{\infty} \frac{1}{(\eta_k\eta_\ell)^{\frac{3}{4}}} \exp\left(\frac{2\pi j(\sqrt{\eta_k}-\sqrt{\eta_\ell}}{\lambda}\right)\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{62}$$

where the metrics η_k and η_ℓ equal

$$\eta_k = z_0^2 + y^2 + (x-x_k)^2, \tag{63}$$

$$\eta_\ell = z_0^2 + y^2 + (x-x_\ell)^2. \tag{64}$$

Using Property 1, it can be shown that

$$\int_{-\infty}^{\infty} \frac{1}{(\eta_k \eta_\ell)^{\frac{3}{4}}} \exp\left(\frac{2\pi j(\sqrt{\eta_k}-\sqrt{\eta_\ell}}{\lambda}\right) \mathrm{d}x = \frac{2}{z_0^2+y^2} \mathrm{sinc}\left(\frac{2(x_k-x_\ell)}{\lambda}\right). \tag{65}$$

Inserting (65) back into (62) yields

$$g_{k,\ell} = \frac{z_0 P}{2\pi} \mathrm{sinc}\left(\frac{2(x_k-x_\ell)}{\lambda}\right) \int_{-B}^{B} \frac{1}{z_0^2+y^2} \mathrm{d}y = \frac{P}{\pi} \tan^{-1}\left(\frac{B}{z_0}\right) \mathrm{sinc}\left(\frac{2(x_k-x_\ell)}{\lambda}\right), \tag{66}$$

which completes the proof.

Appendix C: Proof of Property 3

We first define another auxiliary parameter $\tilde{\theta} = 1-\beta\theta = \alpha\theta$. From the definition of $G(f)$ in (61), the capacity (26) can be split into two parts. In a first part, $G(f)$ is folded by β times with amplitude $\beta\theta\zeta P$ and the integration interval length being $\theta - \tilde{\theta}$, and in a second part, $G(f)$ is folded by $\beta+1$ times with amplitude $(\beta+1)\theta\zeta P$ and the integration interval length being $\tilde{\theta}$. Hence, the capacity (26) equals

$$\mathcal{C} = \frac{1}{\theta}\left((\theta-\tilde{\theta})\log\left(1+\frac{\beta\theta\zeta P}{N_0}\right) + \tilde{\theta}\log\left(1+\frac{(\beta+1)\theta\zeta P}{N_0}\right)\right). \tag{67}$$

By the definition of α, β in (28) and utilizing (23) yields the capacity stated in Property 3.

Appendix D: Proof of Property 4

With only the MF procedure, the capacity with ISI present is in (31), where the interference can be expressed as

$$I = \frac{1}{\zeta P} \sum_{\ell=-\infty, \ell\neq 0}^{\infty} |g_\ell|^2 = \frac{1}{\theta\zeta P} \int_{-\frac{\theta}{2}}^{\frac{\theta}{2}} |G(f)|^2 \, \mathrm{d}f - \zeta P. \tag{68}$$

The second equality in (68) is from Parseval's identity applied to $G(f)$ in (61). Following the same arguments of $G(f)$ as proving Property 3, the interference power can be written as

$$\begin{aligned} I &= \frac{1}{\theta\zeta P}\left((\theta-\tilde{\theta})(\beta\theta\zeta P)^2 + \tilde{\theta}((\beta+1)\theta\zeta P)^2\right) - \zeta P \\ &= \theta\zeta P\left(\theta\beta^2 + 2\tilde{\theta}\beta + \tilde{\theta}\right) - \zeta P. \end{aligned} \tag{69}$$

As $\tilde{\theta} = \alpha\theta$, inserting it back into (69) yields the expression of I in (32).

Appendix E: Proof of Property 6

The proof considers, without loss of generality, $\lambda = 1$. A simple scaling gives the result for arbitrary λ. In two dimensions, the fundamental cell is always a centrally-symmetric hexagon (possibly degenerating into a rectangle) inscribed in a circle whose radius, is the circumcenter of a triangle with vertices $0, \boldsymbol{v}, \boldsymbol{w}$ for some vectors $\boldsymbol{v}, \boldsymbol{w}$ that generate the lattice. So the volume of the lattice generated from $\boldsymbol{S}^{-T}$ is twice the area of a triangle inscribed in a unit circle, and this area is maximized when the triangle is equilateral. This makes the lattice generated from $\boldsymbol{S}^{-T}$, and thus also from $\boldsymbol{S}$, hexagonal.

References

[1] S. Hu, F. Rusek, and O. Edfors, "The potential of using large antenna arrays on intelligent surfaces," *accepted in* IEEE Veh. Technol. Conf. (VTC-Spring), Sydney, 4-7 Jun. 2017, *arXiv preprint: 1702.03128*.

[2] S. Hu, F. Rusek, and O. Edfors, "Cramér-Rao lower bounds for positioning with large intelligent surfaces," *accepted in* IEEE Veh. Technol. Conf. (VTC-Fall), Toronto, Fall, 2017, *arXiv preprint: 1702.03131*.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[4] 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 Process. Mag.*, vol. 30, no. 1, pp. 40-60, Dec. 2012.

[5] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[6] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3899-3911, Jul. 2015.

[7] A. Puglielli, N. Narevsky, P. Lu, T. Courtade, G. Wright, B. Nikolic, and E. Alon, "A scalable massive MIMO array architecture based on common modules," *Proc.* IEEE Int. Conf. Commun. (ICC), workshop on 5G and beyond, May 2015.

[8] L. Atzori, A. Iera, and G. Morabito, "The Internet of things: A survey," *Comput. Netw., Elsevier*, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.

[9] J. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas in Commu.,* vol. 32, no. 6, pp. 1065-1082, Jun. 2014.

[10] A. F. Molisch, *Wireless communications*, the second edition, Wiley-IEEE Press, 2010.

[11] D. P. Petersen and D. Middleton, "Sampling and reconstruction of wave-number-limited functions in N-dimensional Euclidean spaces," *Inf. and Control*, vol. 5, no. 4, pp. 279-323, Dec. 1962.

[12] H. R. Kunsch, E. Agrell, F. A. Hamprecht, "Optimal lattices for sampling," *IEEE Trans. Inf. Theory.*, vol. 51, no.2, pp. 634-647, Jan. 2005.

[13] D. R. Mersereau, "The processing of hexagonally sampled two-dimensional signals," *Proc. IEEE*, vol. 67, no. 6, pp. 930-949, Jun. 1979.

[14] S. M. Kay, *Fundamentals of statistical signal processing volume II: Detection theory,* Prentice-Hall PTR., Upper Saddle River, NJ, 1998.

[15] G. Colavolpe and A. Barbieri, "On MAP symbol detection for ISI channels using the Ungerboeck observation model," *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 720-722, Aug. 2005.

[16] F. Rusek and A. Prlja, "Optimal channel shortening of MIMO and ISI channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 810-818, Feb. 2012.

[17] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "A low-complexity channel shortening terminal with diversity support for evolved 2G device," IEEE Int. Conf. Commun. (ICC), Kuala Lumpur, Malaysia, May 2016, pp. 1-7.

[18] S. Hu and F. Rusek, "A soft-output MIMO detector with achievable information rate based partial marginalization," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1622-1637, Mar. 2017.

[19] S. Hu, X. Gao, and F. Rusek, "Linear precoder design for MIMO-ISI broadcasting channels under channel shortening detection," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1207-1211, Sep. 2016.

[20] C. E. Shannon and W. Weaver, *The mathematical theory of communication,* The University of Illinois Press., Urbana, Illinois, 1949.

[21] J. Salz and J. H. Winters, "Effect of fading correlation on adaptive arrays in digital mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 4, pp. 1049-1057, Nov. 1994.

[22] J. J. Benedetto and G. Zimmermann, "Sampling multipliers and the Poisson summation formula," *J. Fourier Anal. Appl.,* vol. 3. no. 5, pp. 505-523, Sep. 1997.

[23] J. E. Mazo, "Faster-than-Nyquist signaling", *The Bell Syst. Tech. J.,* vol. 54. no. 8, pp. 1451-1462, Oct. 1975.

[24] J. B. Anderson, F. Rusek, and V. Öwall, "Faster-than-Nyquist signaling", *Proc. IEEE*, vol. 101, no. 8, pp. 1817-1830, Aug. 2013.

[25] R. Tolimieri, M. An, and C. Lu, *Mathematics of multidimensional Fourier transform algorithms*, Springer-Verlag, New York, 1993.

[26] E. W. Hansen, "Fast Hankel transform algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 3, pp. 666-671, Jun. 1985.

[27] A. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*, the ninth reprint with additional corrections, Dover Publications, INC., New York, Dec. 1972.

[28] M. Born and E. Wolf, *Principles of optics: Electromagnetic theory of propagation, interference and diffraction of light*, the sixth edition, Pergamon Press, Oxford, England, 1980.

[29] W. Hirt, *Capacity and information rates of discrete-time channels with memory*, Ph.D thesis, no. ETH 8671, Inst. Signal and Inf. Process., Swiss Federal Inst. Technol., Zürich, 1988.

[30] H. R. Knsch, E. Agrell, and F. A. Hamprecht, "Optimal lattices for sampling," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 634-647, Feb. 2005.

[31] J. Chen and P. H. Siegel, "Information rates of two-dimensional finite state ISI channels," IEEE Int. Symp. Inf. Theory (ISIT), Yokohama, Japan, Jul. 2003, pp. 118.

[32] A. S. Y. Poon, R. W. Brodersen, and D. N. C. Tse, "Degrees of freedom in multiple-antenna channels: A signal space approach," *IEEE Trans. Inf. Theory*, vol 51, no. 2, pp. 523-536, Feb. 2005.

[33] A. Alkhateeb, G. Leus, and R. W. Heath Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.,* vol. 14, no. 11, pp. 6481-6494, Nov. 2015.

[34] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE,* vol. 102, no. 3, pp. 366-385, Mar. 2014.

[35] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecom.*, vol. 10, no. 6, pp. 585-595, Nov. 1999.

[36] I. E. Telatar and D. N. C. Tse, "Capacity and mutual information of wideband multipath fading channels," *IEEE Trans. Inf. Theory,* vol. 46, no. 4, pp. 1384-1400, Jul. 2000.

[37] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.,* vol. 6, no. 3, pp. 311-35, Mar. 1998

[38] E. Torkildson, U. Madhow, and M. Rodwell, "Indoor millimeter wave MIMO: Feasibility and performance", *IEEE Trans. Wireless Commun.,* vol. 10, no. 12, pp. 4150-4160, Dec. 2011.

[39] A. Ozgur, O. Lévêque, and D. Tse, "Spatial degrees of freedom of large distributed MIMO systems and wireless ad hoc networks", *IEEE J. Sel. Areas in Commun.,* vol. 31, no. 2, pp. 202-214, Jan. 2013.

[40] M. D. Hahm, E. G. Friedman, and E. L. Titlebaum. "Analog vs. Digital: A comparison of circuit implementations for low-power matched filters," IEEE Int. Symp. Circuits and Syst., vol. 4, 1996.

[41] I. F. Akyildiz, J. M. Jornet, and C. Han, "Terahertz band: Next frontier for wireless communications", *Physical Commun.,* vol. 12, pp. 16-32, Sep. 2014.

[42] A. Sayeed and N. Behdad, "Continuous aperture phased MIMO: basic theory and applications," *Proc.* Annu. Allerton Conf. Commun., Control, and Comput., Sep. 2010, pp. 1196-1203.

[43] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.,* vol. 61, no. 7, pp. 3814-3827, Jul. 2013.

[44] Y. Zeng, R. Zhang, and Z. N. Chen, "Electromagnetic lens-focusing antenna enabled massive MIMO: performance improvement and cost reduction," *IEEE J. Sel. Areas Commun.,* vol. 32, no. 6, pp. 1194-1206, Jun. 2014.

[45] Y. Zeng and R. Zhang, "Millimeter wave MIMO with lens antenna-array: a new path division multiplexing paradigm," *IEEE Trans. Commun.* vol. 64, no. 4. pp. 1557-1571, Apr. 2016.

Paper VIII

Beyond Massive-MIMO: The Potential of Positioning with Large Intelligent Surfaces

We consider the potential for positioning with a system where antenna arrays are deployed as a large intelligent surface (LIS), which is a newly proposed concept beyond massive multi-input multi-output (MIMO) where future man-made structures are electronically active with integrated electronics and wireless communication making the entire environment "intelligent". In a first step, we derive Fisher-information matrix (FIM) and Cramér-Rao lower bound (CRLB) in closed-form for positioning a terminal located perpendicular to the center of the LIS, whose location we refer to as being on the central perpendicular line (CPL) of the LIS. For a terminal that is not on the CPL, closed-form expressions of the FIM and CRLB seem out of reach, and we alternatively find approximations which are shown to be accurate. Under mild conditions, we show that the CRLB for all three Cartesian dimensions (x, y and z) decreases quadratically in the surface-area of the LIS, except for a terminal exactly on the CPL where the CRLB for the z-dimension (distance from the LIS) decreases linearly in the same. In a second step, we analyze the CRLB for positioning when there is an unknown phase φ presented in the analog circuits of the LIS. We then show that the CRLBs are dramatically degraded for all three dimensions but decrease in the third-order of the surface-area. Moreover, with an infinitely large LIS the CRLB for the z-dimension with an unknown φ is 6 dB higher than the case without phase uncertainty, and the CRLB for estimating φ converges to a constant that is independent of the wavelength λ. At last, we extensively discuss the impact of centralized and distributed deployments of LIS, and show that a distributed deployment of LIS can enlarge the coverage for positioning and improve the overall performance.

S. Hu, F. Rusek, and O. Edfors,
"Beyond massive-MIMO: The potential of positioning with Large Intelligent Surfaces,"
accepted in *IEEE Trans. Signal Process.*, Dec. 2017.

1 Introduction

A Large Intelligent Surface (LIS) is a newly proposed concept in wireless communication that is envisioned in [1, 2], where future man-made structures are electronically active with integrated electronics and wireless communication making the entire environment "intelligent". We foresee a practical implementation of LIS as a compact integration of a vast amount of tiny antenna-elements with reconfigurable processing networks. Antennas on the surface cooperate to transmit and sense signals, both for communication and other types of functionality. Machine learning [3, 34] can bring intelligence in the systems both for autonomous operation of the system and for new functionality. One such application is depicted in [20, Fig.1], where three different terminals are communicating to LIS in an outdoor and indoor scenarios, respectively.

The LIS concept can be seen as an extension of earlier research in several other fields. One strong relation is to the massive-MIMO concept [4–6, 9], where large arrays comprising hundreds of antennas are used to achieve massive gains in spectral and energy efficiencies. The natural limit of this evolution is that the LISs in an environment act as transmitting and receiving structures, which allows for an unprecedented focusing of energy in the three-dimensional space which enables, besides unprecedented data-rates, wireless charging and remote sensing with extreme precision. This makes it possible to fulfill the most grand visions in 5G communication [7] and Internet of Things (IoT) [8] systems for providing connections to billions of devices. However, as LIS scales up beyond the traditional antenna array and implies a clean break with the traditional access-point/base-station concept, there are substantial differences between the envisioned LIS and traditional massive MIMO systems: Firstly, LIS in its fundamental form uses the whole contiguous surface for transmitting and receiving; Secondly, in contrast to traditional aperture antennas where the actual physical structure determines the electromagnetic radiation pattern of transmitted and received signals, with a LIS we can control the electromagnetic field on the entire surface, and adapt both transmission and reception across the entire surface.

In [1] we carried out a first analysis on information-transfer capabilities of the LIS, and show that the number of signal-space dimensions per square-meter (m^2) deployed surface-area is π/λ^2, where λ is the wavelength, and the capacity that can be harvested per m^2 surface-area is linear in the average transmit power, rather than logarithmic as in a traditional massive-MIMO deployment. Following [1], in this paper we take a first look at the potential of using LIS for terminal-positioning, where we assume that a terminal to be positioned is equipped with a single-antenna and located in a three-dimensional space in front of the LIS. For analytical tractability, we assume an ideal situation where no scatterers or reflections are present, yielding a perfect line-of-sight (LOS) propagation scenario. Although we do not deal with more complicated geometries, the results are fundamental in understanding the limits of positioning with the LIS.

We first derive the Cramér-Rao lower bounds (CRLBs) for positioning a terminal on the central perpendicular line (CPL) in closed-form, where the CPL is the line perpendicular to the LIS and crossing the LIS at its center point as shown in Fig. 1. For remaining cases, as closed-form expressions seem out of reach and we approximate the Fisher-information matrix (FIM) and CRLB in closed-form, which are shown to be accurate under mild conditions. We also show that, the CRLB in general decreases quadratically in the surface-area of the LIS, except for a terminal on the CPL where the CRLB for z-dimension decreases just linearly in the same. Meanwhile, the impact of wavelength is $\sim\lambda^2$. These scaling laws play in favor of a LIS when compared to other positioning technologies e.g., optical systems [12]. A LIS can compensate for its, comparatively, large wavelength by a much larger aperture.

Besides, we also analyze the CRLB for positioning when there is an unknown phase φ presented in the analog circuits of the LIS, in which case the CRLBs for all dimensions are dramatically increased by φ and in general decrease in the third-order of the surface-area. Therefore, LIS has significant gains over traditional massive MIMO for positioning as LIS has a much larger surface-area. Furthermore, the CRLB for estimating φ is usually significantly large and about $4\pi^2/\lambda^2$ times of the CRLB for the z-dimension. Moreover, for an infinitely large LIS, the CRLB for the z-dimension with unknown φ is 6 dB higher than with known φ, and the CRLB for estimating φ converges to a constant[1].

Then, we also extensively discuss the impact of deployments with a single centralized LIS and multiple distributed smaller LISs constrained to the same total surface-area. We show that, a distributed deployment with splitting the single LIS into 4 small LISs can extend the range of positioning and provide better average CRLB than the centralized deployment under the case that the terminal has a distance to the CPL larger than $\sqrt{6}R$, where R is the radius of the single centralized LIS. Further splitting the 4 small LISs into 16 smaller LISs improves the CRLB, but may also increase the overheads of cooperating among different small LISs.

The rest of the paper is organized as follows. In Sec. II, we describe the signal propagation model for the LIS-terminal link and some common features of FIM computations considered in the paper. In Sec. III, we derive the FIM and CRLB for a terminal on the CPL. In Sec. IV, we discuss a terminal not on the CPL and put forth closed-form approximations of the FIM and CRLB. In Sec. V, we extend the signal model in Sec. II into having a common unknown phase φ, caused by, e.g., the front-end circuits of the LIS and the terminal. In Sec. VI, we discuss the impacts on the CRLB of different deployments of the LIS. Numerical results are provided in Sec. VII and Sec. VIII summarizes the paper.

[1]Note that, all CRLBs and their limits considered in this paper can be linearly scaled down by the signal-to-noise ratio (SNR) as a natural result.

Notation

Throughout this paper, boldface letters indicate vectors and boldface uppercase letters designate matrices. Superscripts $(\cdot)^{-1}$, $(\cdot)^*$ and $(\cdot)^{\mathrm{T}}$ stand for the inverse, complex conjugate, and transpose, respectively. In addition, $\mathcal{R}\{\cdot\}$ takes the real part.

2 Signal Model with LIS

Expressed in Cartesian coordinates, we assume that the center of the LIS is located at coordinates $x=y=z=0$ and a terminal is located at positive z-coordinate. The propagation model of a transmitting terminal at location (x_0, y_0, z_0) to the LIS is depicted in Fig. 1, where we assume a perfect LOS propagation scenario and each terminal is assumed to radiate isotropically. Denoting the wavelength as λ, and assuming a narrow-band system and ideal free-space propagation from the terminal to all points at the LIS, the received signal at the surface at location $(x,\ y,\ 0)$ radiated by a terminal at location $(x_0,\ y_0,\ z_0)$ is

$$\hat{s}_{x_0,\,y_0,\,z_0}(x,y) = s_{x_0,\,y_0,\,z_0}(x,y) + n(x,y), \tag{1}$$

where $n(x,y)$ is modeled as zero-mean white Gaussian noise with spectral density N_0, and the noiseless signal $s_{x_0,\,y_0,\,z_0}(x,y)$ is stated in Property 1.

Property 1. *The noiseless signal $s_{x_0,\,y_0,\,z_0}(x,y)$ can be modeled as*

$$s_{x_0,\,y_0,\,z_0}(x,y) = \frac{\sqrt{z_0}}{2\sqrt{\pi}\eta^{\frac{3}{4}}} \exp\left(-\frac{2\pi j\sqrt{\eta}}{\lambda}\right), \tag{2}$$

where the metric

$$\eta = z_0^2 + (y-y_0)^2 + (x-x_0)^2. \tag{3}$$

Proof. The noiseless signal received at location $(x,y,0)$ on the LIS and at time t as shown in Fig. 1 reads

$$s_{x_0,\,y_0,\,z_0}(x,y) = \sqrt{P_L \cos\phi(x,y)}\, s(t) \exp(-2\pi j f_c \Delta_t(x,y))\,, \tag{4}$$

where P_L denotes the free-space attenuation, $\phi(x,y)$ is angle-of-arrival (AoA) of the transmitted baseband signal $s(t)$ at $(x,y,0)$, and f_c is the carrier-frequency. The transmit-time from the terminal to $(x,y,0)$ equals $\Delta_t(x,y) = \sqrt{\eta}/c$, where c is the speed-of-light. Since we are considering a narrow-band system, the signal $s(t)$ can be assumed to be the same at all locations $(x,y,0)$ of the LIS, hence we can let $s(t)=1$ and remove it from (4). Further, as the free-space attenuation $P_L = 1/(4\pi\eta)$ and $\cos\phi(x,y) = z_0/\sqrt{\eta}$, inserting them back into (4) yields (2). □

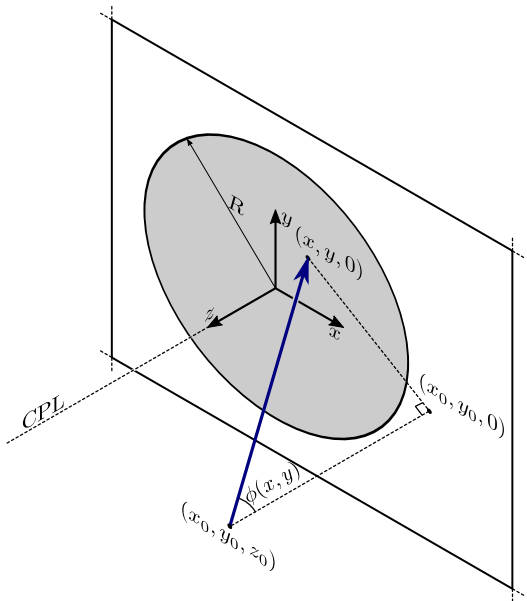


Figure 1: The radiating model of transmitting signal to the LIS. In contrary to a conventional aperture antenna where the physical structure of the antenna controls the radiation of signal, with LIS we can control the electromagnetic field over the entire surface. Therefore, both the near-field (spherical wave-fronts) and far-field (circular wave-fronts) effects are introduced by the integration across the whole disk-shaped LIS when processing the received signal such as for calculating the Fisher information in (8) or data detection in [1].

To analyze the CRLB for positioning, we denote the first-order derivatives of $s_{x_0,\,y_0,\,z_0}(x,y)$ with respect to variables x_0, y_0 and z_0 as Δs_1, Δs_2 and Δs_3, respectively, which are equal to

$$\Delta s_1 = \frac{\sqrt{z_0}\,(x-x_0)}{2\sqrt{\pi}}\left(\frac{3}{2}\eta^{-\frac{7}{4}}+\frac{2\pi j}{\lambda}\eta^{-\frac{5}{4}}\right)\exp\left(-\frac{2\pi j\sqrt{\eta}}{\lambda}\right), \tag{5}$$

$$\Delta s_2 = \frac{\sqrt{z_0}\,(y-y_0)}{2\sqrt{\pi}}\left(\frac{3}{2}\eta^{-\frac{7}{4}}+\frac{2\pi j}{\lambda}\eta^{-\frac{5}{4}}\right)\exp\left(-\frac{2\pi j\sqrt{\eta}}{\lambda}\right), \tag{6}$$

$$\Delta s_3 = \frac{z_0^{\frac{3}{2}}}{2\sqrt{\pi}}\left(\frac{1}{2z_0^2}\eta^{-\frac{3}{4}}-\frac{3}{2}\eta^{-\frac{7}{4}}-\frac{2\pi j}{\lambda}\eta^{-\frac{5}{4}}\right)\exp\left(-\frac{2\pi j\sqrt{\eta}}{\lambda}\right). \tag{7}$$

From [14, Chapter 15], as $\hat{s}_{x_0,\,y_0,\,z_0}(x,y)$ is Gaussian with mean $s_{x_0,\,y_0,\,z_0}(x,y)$ and variance N_0, the elements of the FIM are given by the following double integrals[2]

$$I_{ij} = \frac{2}{N_0}\iint_{x,y}\mathcal{R}\{\Delta s_j\,(\Delta s_i)^*\}\,\mathrm{d}x\mathrm{d}y, \tag{8}$$

where the integrals are taken over the area of the LIS, which we assume to have a disk-

[2]The integrals in (8) are due to the additive property of Fisher-information, which can be obtained by sampling the continuous signal $\hat{s}_{x_0,\,y_0,\,z_0}(x,y)$ with Nyquist frequency[14], and the band-limited property of $s_{x_0,\,y_0,\,z_0}(x,y)$ can be seen from [1].

shape[3] with radius R. As CRLB scales down linearly in SNR, we set $N_0 = 2$ throughout the paper to eliminate the scaling factor in (8).

Further, we define three functions $g_1(n)$, $g_2(n)$ and $g_3(n)$ which are necessary to compute the CRLB based on (5)-(8),

$$g_1(n) = \iint_{x^2+y^2\leq R^2} x^2 \eta^{-\frac{n}{2}} \mathrm{d}x\mathrm{d}y, \tag{9}$$

$$g_2(n) = \iint_{x^2+y^2\leq R^2} y^2 \eta^{-\frac{n}{2}} \mathrm{d}x\mathrm{d}y, \tag{10}$$

$$g_3(n) = \iint_{x^2+y^2\leq R^2} \eta^{-\frac{n}{2}} \mathrm{d}x\mathrm{d}y. \tag{11}$$

In general, closed-form expressions of $g_1(n)$, $g_2(n)$ and $g_3(n)$ are out of reach, except for the case that $x_0 = y_0 = 0$, i.e., the terminal is on the CPL, and it holds that

$$g_1(n) = g_2(n) = \frac{\pi}{n^2-6n+8}\left(2z_0^{4-n} - \left(R^2+z_0^2\right)^{1-\frac{n}{2}}\left(nR^2 - 2R^2 + 2z_0^2\right)\right), \tag{12}$$

$$g_3(n) = \frac{z_0^{2-n} - \left(R^2+z_0^2\right)^{1-\frac{n}{2}}}{n-2}. \tag{13}$$

For a terminal that is not on the CPL, to analyze the properties of CRLB, we will use effective approximations for the FIM and CRLB with the results obtained for the CPL case.

2.1 Related Work

The CRLB for terminal-positioning has been analyzed thoroughly in the literature for traditional antenna-array systems such as in [21–30], and the positioning performance has been evaluated based on various metrics such as AoA [23], time-of-arrival (ToA) [27], time-difference-of-arrival (TDoA) [25], received-signal-strength (RSS) [22], hybrid AoA/ToA [32], and hybrid AoA/TDoA [33]. These positioning methods, both for LOS and non-line-of-sight (NLOS) [35–37] environments, can be directly applied in massive MIMO systems [17–19].

Due to the increased size of measurement in massive MIMO systems, more accurate positioning can be expected, and advanced positioning techniques such as fingerprinting based

[3]The assumption of assuming a LIS with a disk-shape is solely for the convenience of derivations. Other shapes such as rectangular, triangular, or ring shapes can be analyzed in similar ways. Moreover, when the terminal is in the far-field, i.e., $R \ll z_0$, the shape of the LIS is irrelevant and can be regarded as a disk-shape with equal surface-area.

positioning using convolutional neural network (CNN) [34] and Gaussian process regression [17] can be applied. As LIS further extends the antenna-arrays in massive MIMO systems into contiguous surfaces, such advantages can be expected and strengthened in LIS systems. One substantial issue with the LIS is the fundamental limits of terminal-positioning in relation to the LIS surface-area deployed which we are addressing in this work.

Note that, with LIS we derive the fundamental positioning performances under the assumption of a general λ. Practical and implementation-friendly version of LIS that approximates transmission and reception across a contiguous surface with an antenna-array spread across the same surface structure. This is due to the fact that, the integration of received signal across the LIS can be approximated by a summation over discrete signals received by a large antenna-array system. However, such a discrete approximation of LIS with a large antenna-array system requires sampling the LIS densely enough, which depends on the spatial spectrum of received baseband signals at the LIS [20] and is beyond a traditional massive MIMO concept. With such an approximated large antenna-array system originating from the LIS, CRLBs for terminal-position in relation to the deployed surface-area are not adequately addressed in literatures. Nevertheless, the CRLBs derived with the LIS can also be regarded as fundamental limits (lower bounds) for terminal-positioning with a traditional large antenna-array system that has the same deployed surface-area, due to what has been mentioned earlier, with LIS the whole contiguous surface is used for transmitting and receiving and we can control the electromagnetic field on the entire surface.

2.2 Limitations

In this paper, we limit the CRLB analysis with LIS to the perfect LOS case. Under NLOS environments, the signals reach the LIS comprises multi-path components (MPCs) arising from reflections and scatterings. A complete analysis of CRLB for NLOS case would clearly be of major importance, but is out of scope for this paper and should be pursued in future research.

Some approaches for dealing with NLOS in tradition antenna-array systems can be found in e.g., [28–30]. Nevertheless, as shown in [28] and [30], under the assumption that no prior knowledge for the terminal position or NLOS delays is available, the contribution of NLOS signals ought to be completely ignored. That is, the CRLB for NLOS geolocation are equivalent to that for the case that only the LOS signals are received by the LIS. On the other hand, when there is prior information about the geometric relationship of multi-path propagation, the MPCs from a reflecting object can be regarded as direct paths from the virtual terminal behind such a reflecting object [31]. Therefore, in both cases, the CRLB obtained for LOS case is substantial for understanding the terminal-positioning performance with the LIS for NLOS cases.

3 CRLB of a Terminal on the CPL

In this section, we analyze the CRLB for terminals along the CPL with coordinates $(0, 0, z_0)$. A nice property is that, the CRLB for all dimensions are in closed-form, i.e., the integrals (8) can be efficiently solved using (12) and (13). We denote the Fisher-information and CRLB for a terminal with coordinates (x_0, y_0, z_0) and a LIS with radius R as $I_i([x_0, y_0, z_0], R)$ and $C_i([x_0, y_0, z_0], R)$, where the suffix $i = x, y, z$ represents the x, y, and z dimension, respectively. When suffix i has multiple variables, we mean that all these dimensions contained in i are of the same value. For instance, $I_{x,y}([x_0, y_0, z_0], R)$ denotes the Fisher-information for both x and y dimensions whenever they are equal.

We denote an useful parameter

$$\tau = (R/z_0)^2 \,, \tag{14}$$

which measures the relative surface-area normalized by the squared distance of the considered terminal position to the LIS. For a terminal in the far-field (in relation to the radius R), the value of τ is small, and for a terminal close to the LIS, τ becomes large.

3.1 CRLB for Three Cartesian Dimensions

Theorem 1. *The FIM for a terminal with coordinates* $(0, 0, z_0)$ *is diagonal and the Fisher-information for each Cartesian dimension is*

$$I_{x,y}([0, 0, z_0], R) = \frac{1}{30 z_0^2} f_1(\tau) + \frac{2\pi^2}{3\lambda^2} f_2(\tau), \tag{15}$$

$$I_z([0,\ 0,\ z_0], R) = \frac{1}{40 z_0^2} f_3(\tau) + \frac{2\pi^2}{3\lambda^2} f_4(\tau), \tag{16}$$

where the functions $f_1(\tau)$*,* $f_2(\tau)$*,* $f_3(\tau)$*, and* $f_4(\tau)$ *obtained with (12)-(13) are defined as*

$$f_1(\tau) = 1 - \frac{1 + 2.5\tau}{(1+\tau)^{\frac{5}{2}}}, \tag{17}$$

$$f_2(\tau) = 1 - \frac{1 + 1.5\tau}{(1+\tau)^{\frac{3}{2}}}, \tag{18}$$

$$f_3(\tau) = 13 - \frac{13 + 5\tau^2}{(1+\tau)^{\frac{5}{2}}}, \tag{19}$$

$$f_4(\tau) = 1 - \frac{1}{(1+\tau)^{\frac{3}{2}}}. \tag{20}$$

respectively. Then, the CRLB for each dimension can be computed according to

$$C_i([0,\ 0,\ z_0], R) = I_i^{-1}([0,\ 0,\ z_0], R),\ i = x, y, z. \tag{21}$$

Proof. See Appendix A. □

From Theorem 1, the following conclusions can be derived. Firstly, when the terminal is close to the LIS, the Fisher-information is infinitely large for all dimensions, and the CRLB $C_i([0,\ 0,\ z_0], R)$ becomes 0, while under the case $R \ll z_0$, the CRLB is ∞. These observations are consistent with the nature of the problem at hand.

Secondly, in order to get a direct view of the CRLB in relation to surface-area of the LIS, we assume $\lambda \ll z_0$ (which in general holds as λ is the wavelength). Then, the terms of the Fisher-information comprising $f_1(\tau)$ and $f_3(\tau)$ in (15) and (16) can be omitted, and the CRLBs can be approximated as

$$C_{x,y}([0,\ 0,\ z_0], R) \approx \frac{3\lambda^2}{2\pi^2 f_2(\tau)}, \tag{22}$$

$$C_z([0,\ 0,\ z_0], R) \approx \frac{3\lambda^2}{2\pi^2 f_4(\tau)}. \tag{23}$$

respectively. As it can been seen that, the CRLB for all dimensions are uniquely decided by λ and τ. Hence, when z_0 is increased by a factor, the radius R of the LIS also has to increase by the same factor in order to have unaltered CRLBs. Another interesting fact is that, the CRLBs for x and y dimensions are higher than that for z-dimension due to

$$f_2(\tau) < f_4(\tau), \tag{24}$$

which can be seen directly from (18) and (20) using the fact $\tau > 0$.

Lastly, when the surface radius R is much larger than the distance z_0 from the terminal to the LIS, it holds that

$$\lim_{\tau \to \infty} f_2(\tau) = \lim_{\tau \to \infty} f_4(\tau) = 1. \tag{25}$$

Therefore, the asymptotic CRLBs in (22) and (23) are identical and equal to

$$\lim_{\tau \to \infty} C_{x,y,z}([0,\ 0,\ z_0], R) = \frac{3\lambda^2}{2\pi^2}, \tag{26}$$

for all three dimensions and depend solely on the wavelength λ, which represents a fundamental lower limit to positioning precision.

In practical scenarios, a more interesting case is $R \ll z_0$, and then we have the following

approximations by using Taylor expansions [15] at $\tau = 0$,

$$f_1(\tau) = \frac{15}{8}\tau^2 + o\big(\tau^2\big)\,, \tag{27}$$

$$f_2(\tau) = \frac{3}{8}\tau^2 + o\big(\tau^2\big)\,, \tag{28}$$

$$f_3(\tau) = \frac{65}{2}\tau + o(\tau)\,, \tag{29}$$

$$f_4(\tau) = \frac{3}{2}\tau + o(\tau)\,. \tag{30}$$

respectively. From Theorem 1, the CRLBs for different dimensions are equal to

$$C_{x,y}([0,\,0,\,z_0],R) = 16\tau^{-2}\left(\frac{1}{z_0^2} + \frac{4\pi^2}{\lambda^2}\right)^{-1} + o\big(\tau^{-2}\big)\,, \tag{31}$$

$$C_{z}([0,\,0,\,z_0],R) = 16\tau^{-1}\left(\frac{13}{z_0^2} + \frac{16\pi^2}{\lambda^2}\right)^{-1} + o\big(\tau^{-1}\big)\,. \tag{32}$$

As τ is proportional to R^2, for a terminal on the CPL the CRLBs for x and y dimensions decreases quadratically in the surface-area, while the CRLB for z-dimension decreases linearly in the same. Moreover, if we also assume that $\lambda \ll z_0$ (which usually holds as λ is the wavelength), the CRLBs in (31) and (32) can be effectively approximated as

$$C_{x,y}([0,\,0,\,z_0],R) \approx \frac{4\lambda^2}{\pi^2\tau^2}, \tag{33}$$

$$C_{z}([0,\,0,\,z_0],R) \approx \frac{\lambda^2}{\pi^2\tau}, \tag{34}$$

respectively, which only depend on λ and τ. As will be shown in the next section, when the terminal moves away from the CPL, the CRLBs for all three dimensions degrade dramatically compared to (33) and (34), and decreases quadratically in the surface-area.

4 CRLB of a Terminal not on the CPL

In this section, we consider a terminal with arbitrary coordinates (x_0, y_0, z_0). When $x_0, y_0 \neq 0$, closed-form expressions of the CRLB seem out of reach due to the complicated integrals in (8). Therefore, we seek approximations, tight enough so that insights can still be drawn, of the CRLBs. Using the closed-form expressions of Fisher-information for a terminal on the CPL in Sec. III, the CRLBs for general cases can be well approximated as elaborated next.

4.1 CRLB Approximations for a Terminal with Coordinates (x_0, y_0, z_0)

We first introduce two mild conditions[4],

$$\lambda \ll \frac{z_0^2}{\sqrt{z_0^2 + x_0^2 + y_0^2 + R^2}}, \tag{35}$$

$$2R \ll \frac{z_0^2}{\sqrt{x_0^2 + y_0^2}} + \sqrt{x_0^2 + y_0^2}. \tag{36}$$

As for the cases of interest R is relatively small compared to z_0, and λ is much smaller than z_0, these two conditions are usually satisfied. Letting

$$z_1 = \sqrt{x_0^2 + y_0^2 + z_0^2}, \tag{37}$$

the approximations for FIM and CRLB matrices are stated in Property 2.

Property 2. *Under the conditions (35)-(36), the FIM for a terminal with coordinates* (x_0, y_0, z_0) *can be approximated as*

$$\boldsymbol{I} \approx \begin{bmatrix} \alpha + \frac{\beta\, x_0^2}{z_0^2} & \frac{\beta\, x_0\, y_0}{z_0^2} & \frac{\beta\, x_0}{z_0} \\ \frac{\beta\, x_0\, y_0}{z_0^2} & \alpha + \frac{\beta\, y_0^2}{z_0^2} & \frac{\beta\, y_0}{z_0} \\ \frac{\beta\, x_0}{z_0} & \frac{\beta\, y_0}{z_0} & \beta \end{bmatrix}, \tag{38}$$

where α *and* β *are equal to*

$$\alpha = \frac{z_0}{z_1} I_{x,y}([0, 0, z_1], R), \tag{39}$$

$$\beta = \left(\frac{z_0}{z_1}\right)^3 I_z([0, 0, z_1], R), \tag{40}$$

and $I_{x,y}([0, 0, z_1], R)$, $I_z([0, 0, z_1], R)$ *are the Fisher-information for* x, y *and* z *dimensions for a terminal with coordinates* $(0, 0, z_1)$ *which are stated in Theorem 1. Then, the CRLB matrix reads*

$$\boldsymbol{C} = \boldsymbol{I}^{-1} \approx \begin{bmatrix} \frac{1}{\alpha} & 0 & -\frac{x_0}{\alpha\, z_0} \\ 0 & \frac{1}{\alpha} & -\frac{y_0}{\alpha\, z_0} \\ -\frac{x_0}{a\, z_0} & -\frac{y_0}{\alpha\, z_0} & \frac{1}{\beta} + \frac{x_0^2 + y_0^2}{\alpha z_0^2} \end{bmatrix}. \tag{41}$$

Proof. See Appendix B. □

[4]These two conditions are only used to simplify the expressions (5)-(7). That is, only the terms containing $1/\lambda$ in (5)-(7) are preserved and the remaining terms are omitted since they are negligible compared to other terms comprising $1/\lambda$, which simplifies the calculations of Fisher-information as shown later in Appendix B.

From Property 2, the FIM and CRLBs are approximated in closed-form. As a special case, when $x_0 = y_0 = 0$, i.e., the terminal is on the CPL, the approximation (41) is exact as from Theorem 1. Further, we have the below corollary.

Corollary 1. *Under the conditions (35)-(36), the CRLBs for x and y dimensions are approximately equal, and depend on (x_0, y_0, z_0) through z_0 and $\sqrt{x_0^2 + y_0^2}$. That is, terminals on the circle $x_0^2 + y_0^2 = r^2$ have the same CRLBs for all dimensions for a given distance z_0.*

Applying (33)-(34) to approximate $I_{x,y}([0, 0, z_1], R)$ and $I_z([0, 0, z_1], R)$ in (39)-(40), we have the approximated CRLBs for the non-CPL case stated in Property 3.

Property 3. *Under the case $R \ll z_0$ and with conditions in (35)-(36), the CRLBs for a terminal with coordinates (x_0, y_0, z_0) can be approximated as*

$$C_{x,y} \approx \frac{4\lambda^2 z_1^5}{\pi^2 z_0 R^4}, \tag{42}$$

$$C_z \approx \frac{\lambda^2 z_0^2}{\pi^2 R^2} + \frac{4\lambda^2 (x_0^2 + y_0^2) z_1^5}{\pi^2 z_0^3 R^4}. \tag{43}$$

Compared to the CPL case, with a small R the CRLB for z-dimension is dramatically degraded when the terminal is away from the CPL, that is, $x_0^2 + y_0^2 > 0$. Further, when $\sqrt{x_0^2 + y_0^2} > z_0$, the CRLB for z-dimension becomes even larger than the CRLBs for x and y dimensions. Furthermore, the CRLBs decrease quadratically[5] in the surface-area of the LIS for all three dimensions in this case, which is an important motivation to go beyond the massive MIMO deployment to the LIS, which provides significant gains (quadratical in the surface-area) of the CRLB for positioning a terminal.

5 CRLB with Phase Uncertainty in Analog Circuits of the LIS

In practical scenarios, the front-end circuitry of the LIS and of the terminal is not ideal and presents unknown distortions to the signal model. Using off-line calibration of the LIS [16], the entire LIS can be calibrated up to a common constant which is unknown to the LIS. The terminal has its own distortion, but what comes into play here is the product of the two distortions, which is then a single scalar number. In this paper we will model this distortion as a random phase uncertainty φ since amplitude stability is easier to achieve in practice, see e.g., [16]. Such a presence of the unknown phase φ degrades the CRLB of positioning, and in this section we analyze the ensuing CRLB uncertainty thoroughly. To

[5]This is a consequence of the increasing CRLB for a terminal not on the CPL. As the limits of the CRLB when R is ∞ are the same for a terminal at any position with the same z_0, the CRLB for a terminal located not the CPL must decrease faster than when it is located on the CPL.

simplify the analysis, we take a special interest for a terminal on the CPL, while for the other positions we use numerical simulations.

Note that, the received signal power is not impacted by the presence of the unknown phase φ and there are certain parameter measurement methods that are insensitive to the phase uncertainty such as in [39]. However, utilizing only the received power and ignoring the phase differences among received signals across the LIS can significantly degrade the position performance. As a direct consequence, if the phases are removed from the signal model (2), the CRLBs are then independent from λ, whereas in previous obtained results we have shown that the CRLBs are linearly scaled down by λ^2.

With an unknown phase φ, the noiseless signal in (2) is modified to

$$\tilde{s}_{x_0,\,y_0,\,z_0}(x,y) = \frac{\sqrt{z_0}}{2\sqrt{\pi}\eta^{3/4}} \exp\!\left(j\left(-\frac{2\pi\sqrt{\eta}}{\lambda} - \varphi\right)\right). \tag{44}$$

Similarly, we denote the first-order derivatives with respect to variables x_0, y_0, z_0 and φ as $\Delta\tilde{s}_1$, $\Delta\tilde{s}_2$, $\Delta\tilde{s}_3$ and $\Delta\tilde{s}_4$, respectively, which are

$$\Delta\tilde{s}_i = \Delta s_i \exp\left(-j\varphi\right),\ 1 \le i \le 3, \tag{45}$$

$$\Delta\tilde{s}_4 = -j\tilde{s}_{x_0,\,y_0,\,z_0}(x,y), \tag{46}$$

where Δs_i are given in (5)-(7). As the received signal $\hat{s}_{x_0,\,y_0,\,z_0}(x,y)$ is still Gaussian with mean $\tilde{s}_{x_0,\,y_0,\,z_0}(x,y)$ and variance N_0, the elements of FIM are still given by the double integrals in (8). However, compared to the case without φ, in this case the FIM is 4-dimensional and the CRLBs for all three Cartesian dimensions are degraded. We then state the FIM for the non-CPL case in Theorem 2.

Theorem 2. *With an unknown phase φ considered in (44), the FIM equals*

$$\boldsymbol{I} = \begin{bmatrix} \boldsymbol{I}_0 & \boldsymbol{i}^{\mathrm{T}} \\ \boldsymbol{i} & I_{44} \end{bmatrix}, \tag{47}$$

where $\boldsymbol{I}_0$ is the Fisher-information for x, y and z dimensions for the case with known phase φ, and the vector $\boldsymbol{i}$ comprises the cross-terms of Fisher-information between the x, y, z dimensions and the phase φ, which equals

$$\boldsymbol{i} = \begin{bmatrix} I_{14} & I_{24} & I_{34} \end{bmatrix} = \frac{z_0 g_3(4)}{\lambda} \begin{bmatrix} x_0 & y_0 & z_0 \end{bmatrix}. \tag{48}$$

Further, the Fisher-information for the unknown φ equals

$$I_{44} = \frac{z_0}{4\pi} g_3(3), \tag{49}$$

where $g_3(n)$ is the integral defined in (11).

Proof. See Appendix C. □

From Theorem 2, if we know the CRLB matrix $\boldsymbol{C}_0 = (\boldsymbol{I}_0)^{-1}$ for x, y and z dimensions for the case with known φ, the CRLB matrix with φ can be computed as

$$\boldsymbol{C} = \frac{1}{I_{44} - \boldsymbol{i}\boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}}} \begin{bmatrix} \boldsymbol{C}_0\left(I_{44} - \boldsymbol{i}\boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}}\right) + \boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}}\boldsymbol{i}\boldsymbol{C}_0 & -\boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}} \\ -\boldsymbol{C}_0\boldsymbol{i} & 1 \end{bmatrix}. \tag{50}$$

As can be seen from (50), the CRLB for estimating φ equals

$$C_\varphi = \frac{1}{I_{44} - \boldsymbol{i}\boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}}}, \tag{51}$$

and the CRLB matrix for the three Cartesian dimensions becomes

$$\tilde{\boldsymbol{C}}_0 = \boldsymbol{C}_0 + \boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}}\boldsymbol{i}\boldsymbol{C}_0 C_\varphi. \tag{52}$$

Hence, from (52) the CRLBs are dramatically degraded due to the presence of φ for the three Cartesian dimensions with the additional term $\boldsymbol{C}_0\boldsymbol{i}^{\mathrm{T}}\boldsymbol{i}\boldsymbol{C}_0 C_\varphi$. However, as φ plays no role in the FIM in (47), we have the corollary below.

Corollary 2. *The Fisher-information and CRLB for all three Cartesian dimensions and the phase φ are independent of the true value of φ.*

Since in general we cannot get $g_3(n)$ in closed-from, we start with analyzing the FIM for a terminal on the CPL, which from Theorem 2 equals

$$\boldsymbol{I} = \begin{bmatrix} I_{11} & 0 & 0 & 0 \\ 0 & I_{22} & 0 & 0 \\ 0 & 0 & I_{33} & I_{34} \\ 0 & 0 & I_{34} & I_{44} \end{bmatrix}. \tag{53}$$

Hence, the CRLBs for x and y dimensions remain the same with the unknown φ, and the CRLBs for z-dimension and phase φ are equal to

$$C_z = \frac{I_{44}}{I_{33}I_{44} - I_{34}^2}, \tag{54}$$

$$C_\varphi = \frac{I_{33}}{I_{33}I_{44} - I_{34}^2}. \tag{55}$$

On the CPL, we can reach expressions for I_{34} and I_{44} in closed-from, and with I_{ii} ($1 \leq i \leq 3$) computed in Theorem 1, the CRLB for all dimensions are stated in the below property.

Property 4. *With an unknown phase φ, for a terminal on the CPL the CRLBs for x and y dimensions remain the same as with known φ, while the CRLBs for z-dimension and phase φ are equal to*

$$C_z = \left(\frac{1}{10z_0^2} f_5(\tau) + \frac{\pi^2}{6\lambda^2} f_6(\tau)\right)^{-1}, \tag{56}$$

$$C_\varphi = \left(\frac{1}{2} f_7(\tau) + \left(\frac{\lambda^2}{10\pi^2 z_0^2 f_8(\tau)} + \frac{8}{3 f_9(\tau)}\right)^{-1}\right)^{-1}, \tag{57}$$

where the functions $f_5(\tau)$, $f_6(\tau)$, $f_7(\tau)$, $f_8(\tau)$ and $f_9(\tau)$ obtained with (12)-(13) are defined as

$$f_5(\tau) = 1 - \frac{1 + 1.25\tau^2}{(1+\tau)^{\frac{5}{2}}}, \tag{58}$$

$$f_6(\tau) = 1 - \frac{4 - 3\sqrt{1+\tau} + 3\tau}{(1+\tau)^{\frac{3}{2}}}, \tag{59}$$

$$f_7(\tau) = 1 - \frac{1}{\sqrt{1+\tau}}, \tag{60}$$

$$f_8(\tau) = \frac{\tau^2\sqrt{1+\tau}}{4 + 5\tau^2 - 4(1+\tau)^{\frac{5}{2}}}, \tag{61}$$

$$f_9(\tau) = \frac{\tau^2}{\sqrt{1+\tau} - (1+\tau)^2}. \tag{62}$$

Proof. See Appendix D. □

Using Property 4, when $\tau \to \infty$ it holds that

$$\lim_{\tau\to\infty} f_5(\tau) = \lim_{\tau\to\infty} f_6(\tau) = 1, \tag{63}$$

and the CRLB limit for z-dimension is

$$\lim_{\tau\to\infty} C_z = \frac{6\lambda^2}{\pi^2}, \tag{64}$$

which is 4 times of the CRLB for z-dimension with known φ, hence, the unknown phase causes 6 dB degradation of the positioning precisions for z-dimension for a terminal on the CPL. Further, as it also holds that

$$\lim_{\tau\to\infty} f_7(\tau) = 1, \tag{65}$$

$$\lim_{\tau\to\infty} f_8(\tau) = -\frac{1}{4}, \tag{66}$$

$$\lim_{\tau\to\infty} f_9(\tau) = -1, \tag{67}$$

the CRLB limit for phase φ equals

$$\lim_{\tau\to\infty} C_\varphi = \left(\frac{1}{2} - \left(\frac{8}{3} + \frac{\lambda^2}{10\pi^2 z_0^2}\right)^{-1}\right)^{-1}, \tag{68}$$

which becomes a constant when $\lambda \ll z_0$,

$$\lim_{\tau\to\infty} C_\varphi = 8. \tag{69}$$

Therefore, in order to estimate φ, the SNR should be extremely high regardless of the wavelength λ and surface-area of the LIS.

To see the trends at small τ, we also use Taylor expansions at $\tau = 0$ which results in

$$f_5(\tau) = \frac{5}{2}\tau + o(\tau), \tag{70}$$

$$f_6(\tau) = \frac{1}{8}\tau^3 + o\left(\tau^3\right), \tag{71}$$

$$f_7(\tau) = \frac{1}{2}\tau + o(\tau), \tag{72}$$

$$f_8(\tau) = -\frac{1}{10}\tau + o(\tau), \tag{73}$$

$$f_9(\tau) = -\frac{2}{3}\tau + o(\tau). \tag{74}$$

From Property 4 and using (70)-(74), when τ is sufficiently small we have the approximations

$$C_z \approx \frac{48\lambda^2}{\pi^2\tau^3}\left(1 + \frac{12\lambda^2}{\pi^2 z_0^2 \tau^2}\right)^{-1}, \tag{75}$$

$$C_\varphi \approx \frac{4}{\tau\lambda^2}\left(\lambda^2 + 4\pi^2 z_0^2\right). \tag{76}$$

An interesting fact is that, unlike the case with known φ where the CRLB for z-dimension decreases linearly in the surface-area, in the presence of an unknown φ the slope of the CRLB for z-dimension in relation to the surface-area (both are in logarithmic domain) varies between 1 and 3. This can be seen from (75) as we have the two cases:

- When $\frac{2\sqrt{3}\lambda}{\pi z_0} \ll \tau \ll 1$, it holds that

$$C_z \approx \frac{48\lambda^2}{\pi^2\tau^3}, \tag{77}$$

  which decreases in the third-order of the surface-area of the LIS.

- When $0<\tau\ll\frac{2\sqrt{3}\lambda}{\pi z_0}$, it holds that

$$C_z\approx\frac{4z_0^2}{\tau},\tag{78}$$

which decreases linearly in the surface-area of the LIS.

Remark 1. *Note that, the CRLB for z-dimension in (78) is independent of λ, which is different from the CRLB with known phase as in (34). Therefore, with phase uncertainty, decreasing the wavelength is not beneficial for improving the CRLB for estimating the distance z_0.*

Moreover, when τ is sufficiently small and $\lambda\ll z_0$ holds, the CRLB for phase φ is significantly larger than that for z-dimension since

$$\frac{C_\varphi}{C_z}\approx\frac{4\pi^2}{\lambda^2}.\tag{79}$$

In Fig. 2, we depict the CRLBs for all three Cartesian dimensions with and without φ, derived in Theorem 1 and Property 4 for the LIS, respectively, where we let $z_0=4$ m and $\lambda=0.1$ m. Assuming that the distance between two adjacent antenna-elements is $\lambda/2$, then the total number of antenna-elements deployed in a traditional massive MIMO system with a surface-area πR^2 equals

$$N=\frac{4\pi R^2}{\lambda^2}=\frac{4\pi\tau z_0^2}{\lambda^2}\approx 2\tau\times 10^4.\tag{80}$$

A typical traditional massive-MIMO array comprises $N=200$ antennas and yields $\tau=0.01$. As seen from Fig. 2, a LIS system with the same surface-area (whose CRLBs are lower bounds of those that a traditional massive MIMO can achieve with the same surface-area) falls just short of reaching the cubic slope, whereas a LIS that increases the surface-area 10-20 fold with $\tau>0.1$ reaches the cubic slope. This clearly shows the benefit of increasing the surface-area from a traditional massive MIMO to a LIS with much larger surface-area that can fall in the cubic slope for achieving better terminal-positioning performances.

In Fig. 3, we depict the CRLBs for z-dimension with unknown phase φ. As can be seen, the approximations in (75)-(76) are well aligned with the exact forms obtained in Property 4 when $\tau<0.02$. Moreover, in this case the CRLBs for estimating φ is around $4\pi^2/\lambda^2=4000$ times of the CRLB for z-dimension which is shown in (79).

Following the similar discussion in Sec. IV and utilizing the approximations in Property 2 and (51)-(52), the CRLBs for a terminal not on the CPL can also be approximated. However, the derivations are relatively long and the conclusions are similar as those drawn for the case with a terminal on the CPL. Loosely speaking, the Fisher-information terms comprised in

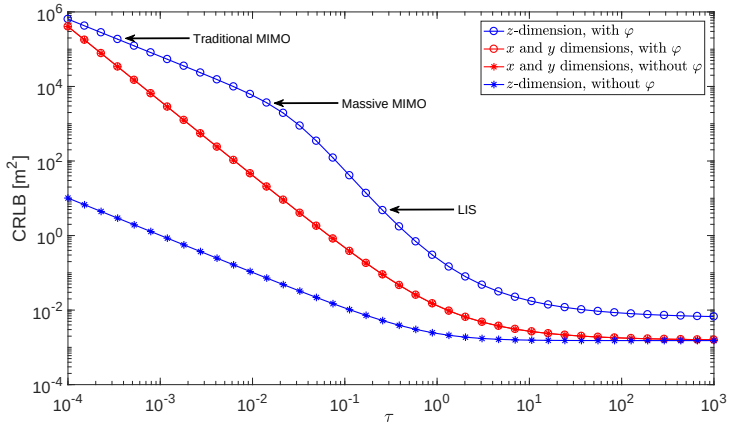


Figure 2: The exact CRLB for x, y and z dimensions for terminals along the CPL obtained in Theorem 1 and Property 4. As can be seen, with LIS the CRLB is in the fast-decreasing region compared to a traditional massive-MIMO system whose surface-area is relatively small, which shows the potential gains of the LIS.

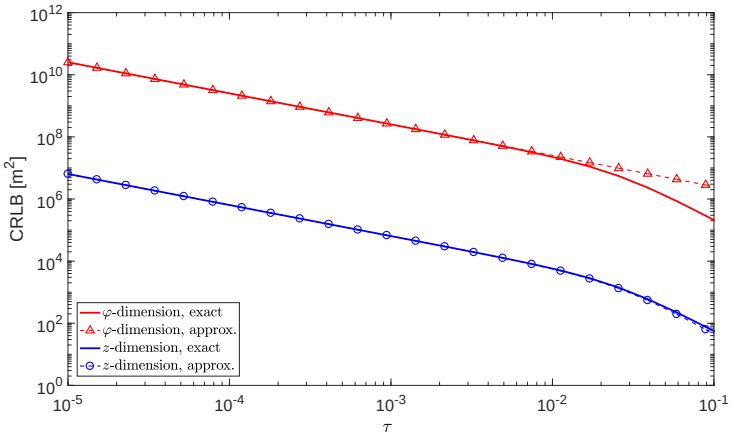


Figure 3: The exact CRLB in (56)-(57) and approximated CRLB in (75)-(76) for z-dimension and phase φ for terminals on the CPL, which are well aligned for small values of τ.

vector $\boldsymbol{i}$ increases linearly in the surface-area and $\boldsymbol{C}_0$ decreases quadratically in the surface-area for all dimensions as explained in Sec. IV. Further, as $\boldsymbol{C}_\varphi$ decreases linearly in the surface-area, the CRLBs for x, y and z dimensions then decrease in the third-order of the surface-area from (52), which is the same as for the CPL case. Furthermore, as τ grows large, the limits of the CRLBs for x and y dimensions remain the same for the case with known φ since all positions can be approximated as on the CPL in the far-field, while for z-dimension the limit of CRLB is 6 dB higher than that with known φ as shown in (64).

6 Deployment of the LIS

In this section we consider different deployments of the LIS on a large surface with size $W \times H$ where W, H are the width and length, respectively. In particular, we consider the centralized-deployment (a) and distributed-deployments (b), (c) as depicted in Fig. 4. For simplicity, we assume $R,\ \lambda \ll z_0$ and consider the CRLBs for a terminal on the CPL with coordinates $(0, 0, z_0)$ without phase-uncertainty in the received signal, that is, positioning a terminal in the far-field.

For the centralized deployment (a), the CRLBs for all three dimensions are given in (33) and (34). With a distributed deployment (b), the LIS is split into four small LISs centered at $(\pm W/4, \pm H/4)$, each with radius $R/2$. Using Property 2, the symmetry of the LIS, and the approximations in (33)-(34), the sum of the FIMs corresponding to the four small LISs can be shown to be diagonal, and the Fisher-information for the x, y and z dimensions are equal to

$$I_{x,y} \approx \frac{\pi^2 z_0 R^4}{16\lambda^2 (z_0^2 + D^2)^{5/2}} + \frac{\pi^2 D^2 z_0 R^2}{2\lambda^2 (z_0^2 + D^2)^{5/2}}, \tag{81}$$

$$I_z \approx \frac{\pi^2 R^2 z_0^3}{\lambda^2 (z_0^2 + D^2)^{5/2}}, \tag{82}$$

respectively, where D equals

$$D = \frac{\sqrt{W^2 + H^2}}{4}. \tag{83}$$

Assuming $D \ll z_0$, the Fisher-information can further be approximated as

$$I_{x,y} \approx \frac{\pi^2 R^4}{4\lambda^2 z_0^4} \left(\frac{1}{4} + \frac{2D^2}{R^2} \right), \tag{84}$$

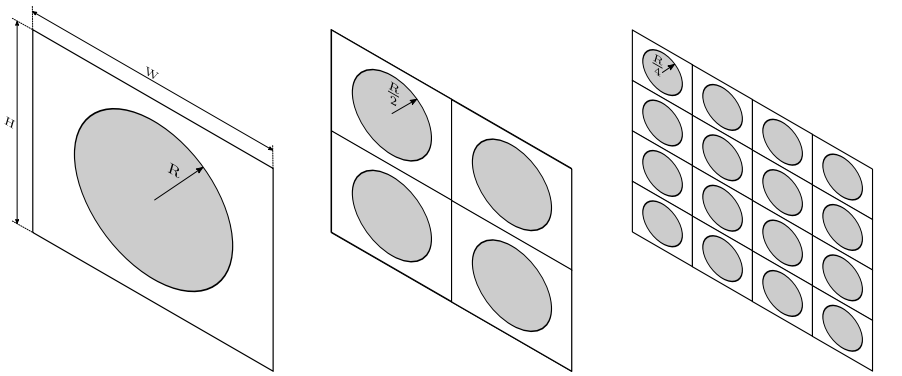
$$I_z \approx \frac{\pi^2 R^2}{\lambda^2 z_0^2}. \tag{85}$$

Comparing (33) to (84), it can be seen that the CRLB for x and y dimensions with the distributed deployment (b) is lower than that with the centralized deployment (a) only if

$$\frac{1}{4} + \frac{2D^2}{R^2} > 1, \tag{86}$$

or equivalently,

$$\sqrt{W^2 + H^2} > \sqrt{6} R. \tag{87}$$



a. Centralized: Single LIS. b. Distributed: 4 half-scale LISs. c. Well distributed: 16 quarter-scale LISs.

Figure 4: Different deployments of the LIS in a surface with width W and length H. Note that the total surface-area is the same for different deployments, and each of the small LISs has the same properties of the single large LIS.

That is to say, in the far-field with the distributed deployment (b), the CRLBs for x and y dimensions are improved if the four small LISs are deployed sufficiently far apart in relation to radius R. Otherwise, the centralized deployment (a) provides lower CRLBs for x and y dimensions than that for the distributed deployment. However, the CRLB for z-dimension remains the same for both deployments. Further, when $R \ll D$, the Fisher-information in (84) becomes

$$I_{x,y} \approx \frac{\pi^2 D^2 R^2}{2\lambda^2 z_0^4}, \tag{88}$$

which shows that, the CRLBs for x and y dimensions are not only improved, but also decreases linearly in the surface-area of the LIS with a distributed deployment rather than quadratically.

Following the same principle, one can split the LIS into more small pieces and obtain an ultra-densely distributed deployment such as in (c) of Fig. 4. In general, with a distributed deployment, the overall positioning performance is more robust than a centralized deployment, and the average positioning performance is improved which we show later with numerical simulations.

7 Numerical Results

In this section, numerical results are provided to illustrate the theories and conclusions that we have developed in previous sections. As explained earlier, in all tests we set the noise spectral-density to $N_0 = 2$, and without explicitly pointed out, the unit for the coordinates of the terminal, the wavelength λ and the radius R of the LIS are all in meter, while the unit for CRLB is m^2.

7.1 Exact-CRLB Evaluations

We first evaluate the CRLB for terminals both on and away from the CPL as discussed in Sec. III and Sec. IV. As only the radius $\sqrt{x_0^2+y_0^2}$ matters as shown in Corollary 1, we illustrate with offsets only in x-dimension. In Fig. 5 and Fig. 6, we test with $R=1$, $\lambda=0.1$, $y_0=0$, $x_0=2$, 4, 8, and $z_0=4$, 6, respectively, and some interesting results can be observed.

Firstly, as shown in Fig. 5, when τ is small the CRLBs for x and y dimensions decrease quadratically in the surface-area of the LIS, while as shown in Fig. 6, the CRLB for z-dimension decreases only linearly in that. This is well aligned with the results in (33) and (34). Secondly, the CRLB for z-dimension increases dramatically when the terminal is away from the CPL. Furthermore, as long as $x_0 \neq 0$, the CRLB for z-dimension also decreases quadratically in the surface-area. These phenomenons are well predicted by Property 2. Lastly, it can been seen that, as $R \to \infty$ the CRLB converges to a limit $3\lambda^2/(2\pi^2) = 1.5\times10^{-3}$ for all dimensions as shown in (26).

7.2 CRLB-Approximation Accuracies

Next, we evaluate the CRLB approximations for a terminal not on the CPL as discussed in Sec. IV. We compare the numerical integration results[6] of CRLB and their approximations using (42)-(43) in Property 2. We test with $R=0.5$, $\lambda=0.1$, $z_0=8$, and $x_0=y_0$ varying from 1 to 8.

The CRLBs and the normalized approximation errors that are computed as the normalized CRLB differences between the numerical integrations and the approximations are both shown in Fig. 7. As can be seen, the approximations given by Property 2 perform well, with normalized errors less than 0.5% for the x and y dimensions, and close to 1% for z-dimension.

In Fig. 8, we repeat the tests in Fig. 5 and Fig. 6 with numerical integrations, but setting $x_0 = r\cos\psi$ and $y_0 = r\sin\psi$, with $r=4$ and ψ changing over $[0, 2\pi]$. The CRLBs in all dimensions are normalized with those obtained at coordinates $(4, 0, 8)$. As can be seen, the CRLB for the z-dimension is identical for any angle ψ, while the CRLBs for x and y dimensions are almost identical. These observations corroborate Corollary 1.

[6]For numerical computation of the CRLB, we use the Matlab built-in function 'integral' to calculate the integrals in the CRLB matrix directly, which has an absolute error of 10^{-10} and relative error of 10^{-6}.

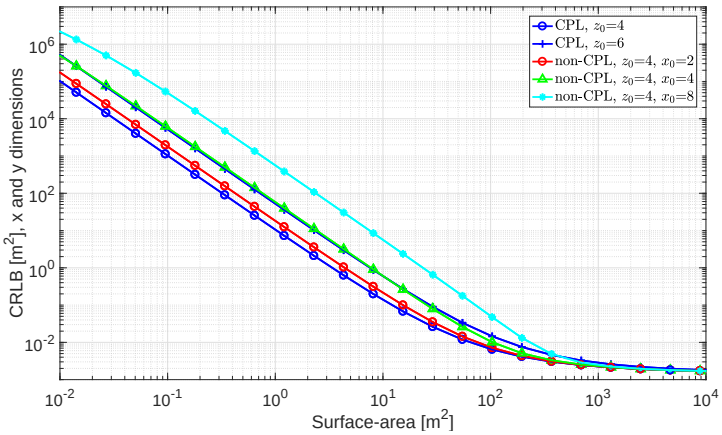


Figure 5: CRLB for x and y dimensions, and the CRLBs for y-dimension are almost overlapped with those for x-dimension.

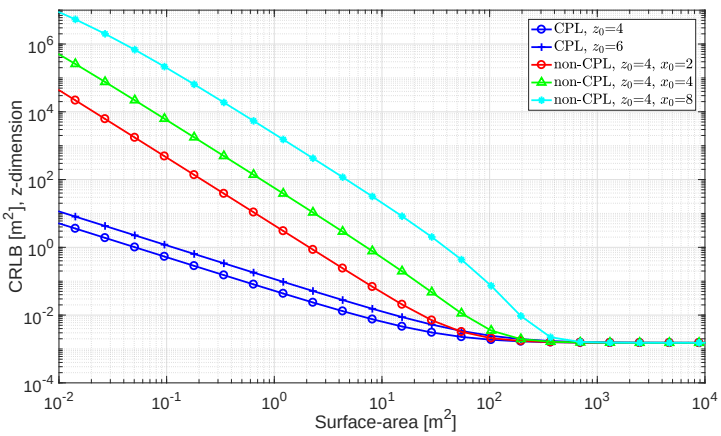


Figure 6: CRLB for z-dimension with the same tests in Fig. 5.

7.3 CRLB with an Unknown Phase φ

Next, we evaluate the CRLB for positioning with an unknown phase φ presented as discussed in Sec. V. As can be seen in Fig. 9, when the terminal is away from the CPL, the CRLBs for all dimensions are increased and the curves have similar shapes. For all three Cartesian dimensions, the CRLB starts to decrease in the third-order of the surface-area when R is larger than a certain threshold as explained in (115). More interestingly, the CRLBs for x and y dimensions are lower than that for z-dimension when there is an unknown phase φ present in the signal model. Furthermore, the behaviors of CRLB for a terminal not on the CPL is slightly different from the case located on the CPL. As can be seen, when R is small, the CRLB decreases first quadratically in the surface-area instead of linearly, which is mainly because that the CRLB converges to the case with known φ, since the CRLB is so large that the impact of an unknown φ is negligible. The CRLB for phase φ is much higher than for the other Cartesian dimensions, and is around $\frac{4\pi^2}{\lambda^2}$ times of the

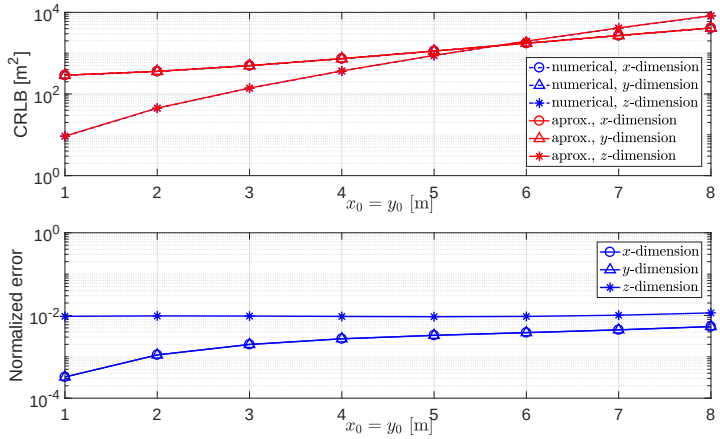


Figure 7: CRLB computed with numerical integrations and their approximations using (42)-(43) in Property 2, and the normalized approximation errors.

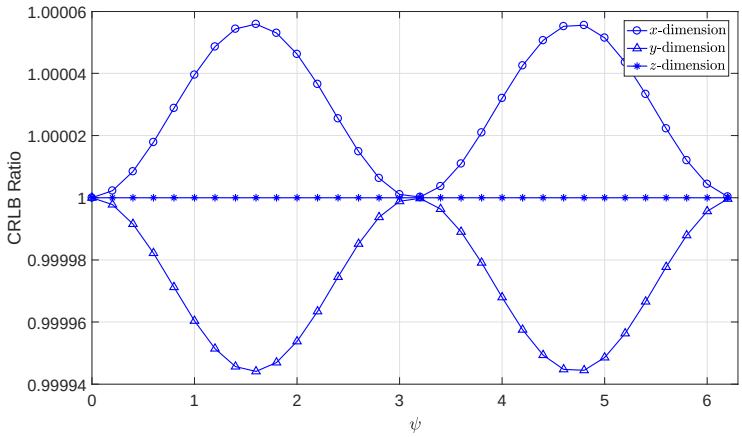


Figure 8: The CRLB differences for terminals on a circle that is parallel to the LIS with center $(0, 0, 4)$ and radius $r = 4$.

CRLB for z-dimension as shown in (79), which basically means that the estimation of φ is highly inaccurate unless at a very high SNR.

7.4 CRLB with Centralized and Distributed Deployments of the LIS

Finally, we evaluate the CRLB with the centralized and distributed deployments as discussed in Sec. VI. We set $W = H = 4$ and $z_0 = 8$. All curves are obtained with numerical integrations without any approximations. We compare the CRLB with different deployments depicted in Fig. 4, that is, a single LIS, 4 small LISs, and 16 smaller LISs, with the same total surface-area.

As shown in Fig. 10 for a terminal on the CPL, when (87) is fulfilled, i.e., $R \leq \sqrt{\frac{W^2+H^2}{6}} = 2.31$, the distributed deployments with 4 and 16 small LISs render lower CRLBs than the

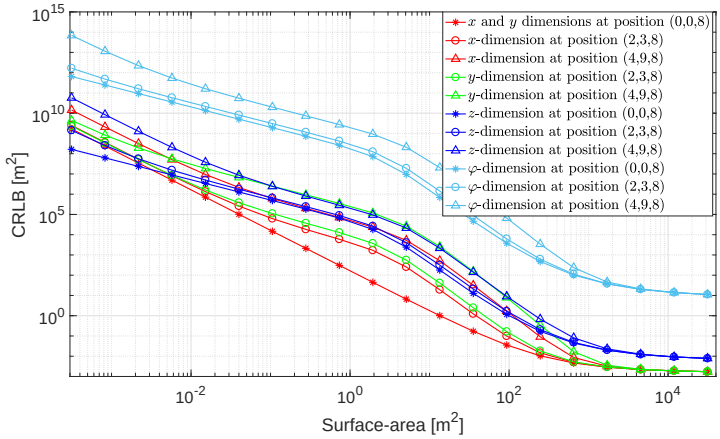


Figure 9: The CRLB evaluated with unknown phase φ for a terminal with different locations, both on and away from the CPL.

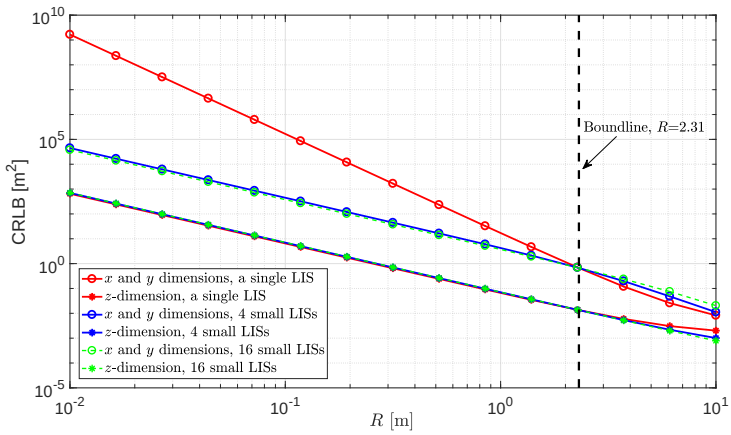


Figure 10: The CRLB with different deployments of the LIS for a terminal on the CPL with $z_0 = 8$ and different radius R.

centralized deployment for x and y dimensions, while the CRLB for z-dimension remains the same. When R increases beyond the threshold, the distributed deployments become worse for x and y dimensions, although the CRLB for z-dimension is slightly better.

In order to evaluate the average positioning performance, we draw 1000 terminals with coordinates of x and y dimensions uniformly distributed in [-2, 2], and $z_0 = 12$ for all terminals. In Fig. 11 we plot the average CRLB for different dimensions. As can be seen, the average CRLB for all three dimensions are significantly improved with the distributed deployments. The average CRLB with 4 small LISs, each small LIS has a radius 0.005, can achieve the same average CRLB for a single LIS with $R = 0.2$, that is, the surface-area needed for the distributed deployment is only 0.25% of that for a centralized deployment when R is small. As R increases, different deployments converge to each other as expected. Further splitting the 4 small LISs into 16 smaller LISs provides marginal gains, but a likely cost of more stringent hardware requirements to achieve phase calibration and cooperation

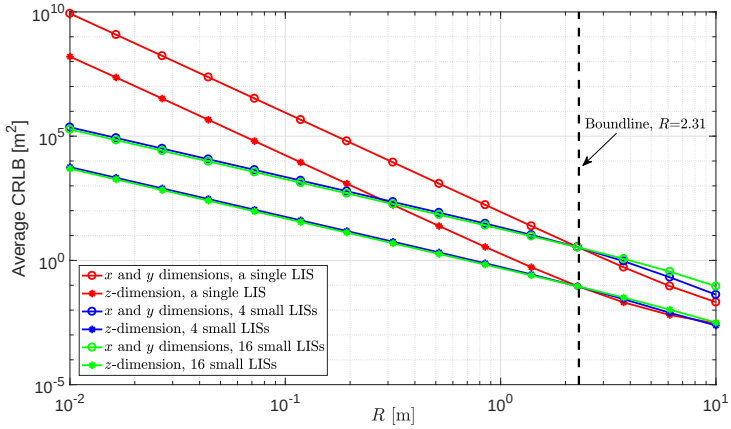


Figure 11: The average CRLB with different deployments of the LIS for 1000 uniformly distributed terminal locations, and with $z_0 = 8$ and different radius R.

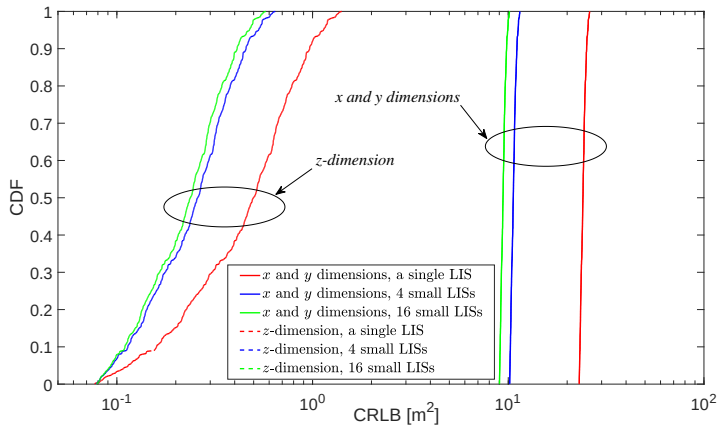


Figure 12: The CDF of CRLB with different deployments of the LIS for 1000 uniformly distributed terminals with $R = 1.39$.

among the small LISs.

The cumulative distribution functions (CDF) of the CRLBs are plotted in Fig. 12, where we can see that the CRLBs for all three Cartesian dimensions with a distributed deployment comprising 4 small LISs are significantly improved compared to a single centralized LIS. The CRLBs for x and y dimensions are relatively larger than that for z-dimension, however, the values of CRLB are also more concentrated than those for z-dimension. With 4 small LISs, the values of CRLB also become concentrated, which means that the overall positioning performance is improved with a distributed deployment of the LIS.

8 Summary

In this paper, we have derived the Fisher-information matrix (FIM) and Cramér-Rao lower bounds (CRLB) for positioning with large intelligent surfaces (LIS). For a terminal on the central perpendicular line (CPL), the CRLBs are derived in closed-form. For other positions we alternatively provide approximations in closed-form to compute the Fisher-information and CRLB which are shown to be accurate. We have also shown that, under mild conditions the CRLBs for x and y dimensions decrease quadratically in the surface-area of the deployed LIS. For z-dimension, the CRLB decreases linearly in the surface-area for a terminal on the CPL. When the terminal is away from the CPL, the CRLBs for all Cartesian dimensions increase dramatically and decrease quadratically in the surface-area of the LIS.

Furthermore, we have analyzed the CRLB for positioning in the presence of a random unknown phase φ in the received signal model. We have shown that, the CRLBs are dramatically increased by the unknown phase, and in general the CRLBs for all dimensions decrease in the third-order of the surface-area, provided that the surface-area exceeds a certain threshold. We have also shown that, for an infinitely large LIS, the CRLB for z-dimension with an unknown phase is 6 dB higher than that with a known φ, and the CRLB for estimating φ converges to a constant independent of the wavelength λ.

Additionally, we compare centralized and distributed deployments of the LIS and show that, the distributed deployments have the potential to extend the coverage of terminal-positioning and can provide better average CRLBs for all dimensions.

Appendix A: Proof of Theorem 1

For a terminal on the CPL, we have $x_0 = y_0 = 0$, and then the first-order derivatives with respect to x and y are equal to

$$\Delta s_1 = \frac{\sqrt{z_0}x}{2\sqrt{\pi}} \left(\frac{3}{2}\eta^{-\frac{7}{4}} + \frac{2\pi j}{\lambda}\eta^{-\frac{5}{4}} \right) \exp\left(-\frac{2\pi j\sqrt{\eta}}{\lambda} \right), \tag{89}$$

$$\Delta s_2 = \frac{\sqrt{z_0}y}{2\sqrt{\pi}} \left(\frac{3}{2}\eta^{-\frac{7}{4}} + \frac{2\pi j}{\lambda}\eta^{-\frac{5}{4}} \right) \exp\left(-\frac{2\pi j\sqrt{\eta}}{\lambda} \right), \tag{90}$$

where $\eta = z_0^2 + y^2 + x^2$, and the first-order derivative with respect to z is in (7). Since η is an even function with respect to x and y, the cross-terms of different dimensions in the FIM vanish, and we obtain a diagonal FIM with diagonal elements being

$$I_{ii} = \iint_{x^2+y^2 \le R^2} |\Delta s_i|^2 \mathrm{d}x\mathrm{d}y. \tag{91}$$

Calculating (91) directly yields

$$I_{11} = I_{22} = \frac{z_0}{4\pi}\left(\frac{9}{4}g_1(7) + \frac{4\pi^2}{\lambda^2}g_1(5)\right), \tag{92}$$

$$I_{33} = \frac{z_0^3}{4\pi}\left(\frac{1}{4z_0^4}g_3(3) + \left(\frac{4\pi^2}{\lambda^2} - \frac{3}{2z_0^2}\right)g_3(5) + \frac{9}{4}g_3(7)\right), \tag{93}$$

Utilizing the results in (12) and (13) and after some manipulations, the Fisher-information for different dimensions are then in (15) and (16).

Appendix B: Proof of Property 2

Using the condition in (35), the derivatives in (5)-(7) can be approximated as

$$\Delta s_1 \approx \frac{\sqrt{\pi z_0}\,(x - x_0)\,j}{\lambda}\eta^{-\frac{5}{4}}\exp\left(-\frac{2\pi j}{\lambda}\sqrt{\eta}\right), \tag{94}$$

$$\Delta s_2 \approx \frac{\sqrt{\pi z_0}\,(y - y_0)}{\lambda}\eta^{-\frac{5}{4}}\exp\left(-\frac{2\pi j}{\lambda}\sqrt{\eta}\right), \tag{95}$$

$$\Delta s_3 \approx -\frac{\sqrt{\pi}z_0^{\frac{3}{2}}j}{\lambda}\eta^{-\frac{5}{4}}\exp\left(-\frac{2\pi j}{\lambda}\sqrt{\eta}\right), \tag{96}$$

with approximation errors go to zero if $\frac{\lambda\sqrt{z_1^2+R^2}}{z_0^2}$ goes to zero.

On the other hand, denoting the metric corresponds coordinates $(0, 0, z_1)$ as

$$\eta_1 = z_1^2 + x^2 + y^2, \tag{97}$$

then we have

$$\eta = z_0^2 + (y - y_0)^2 + (x - x_0)^2 = \eta_1(1 - \varepsilon), \tag{98}$$

where it holds from (36) that

$$|\varepsilon| = \left|\frac{2(xx_0 + yy_0)}{\eta_1}\right| \leq \frac{2R\sqrt{x_0^2 + y_0^2}}{z_1^2} \ll 1. \tag{99}$$

Therefore, we can approximate η by η_1.

Now we first consider approximating the Fisher-information for the z-dimension with coordinates (x_0, y_0, z_0) using that obtained with coordinates $(0, 0, z_1)$. Based on the derivative for the z-dimension in (96), directly computing $\Delta s_3\,(\Delta s_3)^*$ with coordinates

(x_0, y_0, z_0) yields

$$\begin{aligned}\Delta s_3 \left(\Delta s_3\right)^*\big|_{(x_0,y_0,z_0)} &\approx \frac{\pi z_0^3}{\lambda^2}\eta^{-\frac{5}{2}} \\ &\approx \left(\frac{z_0}{z_1}\right)^3 \frac{\pi z_1^3}{\lambda^2}{\eta_1}^{-\frac{5}{2}} \\ &= \left(\frac{z_0}{z_1}\right)^3 \Delta s_3 \left(\Delta s_3\right)^*\big|_{(0,0,z_1)}. \end{aligned} \tag{100}$$

Therefore, the Fisher-information for the z-dimension with coordinates (x_0, y_0, z_0) can be approximated as

$$I_{33} \approx \beta = \left(\frac{z_0}{z_1}\right)^3 I_z([0, 0, z_1], R). \tag{101}$$

Similarly, computing $\Delta s_1 \left(\Delta s_3\right)^*$ based on (94) and (96) yields

$$\Delta s_1 \left(\Delta s_3\right)^* \approx \left(-\frac{\pi x z_0^2}{\lambda^2}\eta^{-\frac{5}{2}} + \frac{\pi x_0 z_0^2}{\lambda^2}\eta^{-\frac{5}{2}}\right). \tag{102}$$

As the first term $-\frac{\pi x z^2}{\lambda^2}\eta^{-\frac{5}{2}}$ is an odd function in x, the integral over it (with respect to x and y) is thusly zero. Then, by directly comparing the remaining term in (102) to (100) and after integrating over x and y, it can be shown that

$$I_{13} = \frac{x_0}{z_0} I_{33} \approx \frac{x_0}{z_0}\beta. \tag{103}$$

Next we approximate the Fisher-information for the x-dimension. At first, note that with approximating η by η_1, the noiseless signal (2) can be written as

$$s_{x_0,y_0,z_0}(x, y) \approx \sqrt{\frac{z_0}{z_1}} s_{0,0,z_1}(x, y). \tag{104}$$

Then, using (94) we have

$$\Delta s_1 \left(\Delta s_1\right)^* \approx \frac{\pi z_0 (x^2 - 2x x_0 + x_0^2)}{\lambda^2}\eta^{-\frac{5}{2}}, \tag{105}$$

Based on (104), the integrals of the first term $\frac{\pi z_0 x^2}{\lambda^2}\eta^{-\frac{5}{2}}$ in (105) can be approximated by the Fisher-information for the x-dimension with coordinates $(0, 0, z_1)$. The second term

$\frac{xx_0}{\lambda^2}\eta^{-\frac{5}{2}}$ is an odd function in x and the integral over it is zero. At last, comparing the last term $\frac{\pi z_0 x_0^2}{\lambda^2}\eta^{-\frac{5}{2}}$ to (100) yields

$$\begin{aligned} I_{11} &\approx \frac{z_0}{z_1} I_{x,y}([0,0,z_1],R) + \frac{x_0^2}{z_0^2}\beta \\ &= \alpha + \frac{x_0^2}{z_0^2}\beta. \end{aligned} \tag{106}$$

Utilizing the symmetry between x and y dimensions, from (103) and (106) it can also be shown that

$$I_{23} \approx \frac{y_0}{z_0}\beta, \tag{107}$$

$$I_{22} \approx \alpha + \frac{y_0^2}{z_0^2}\beta. \tag{108}$$

Finally, based on (94) and (95) it holds that

$$\Delta s_1 \left(\Delta s_2\right)^* = \frac{\pi z_0 (xy - xx_0 - yy_0 + x_0 y_0)}{\lambda^2}\eta^{-\frac{5}{2}}, \tag{109}$$

and the integrals of the first three terms in (109) vanish as they are odd functions either in x or y. Then, comparing the last term to (100), it can be shown that

$$I_{12} \approx \frac{x_0 y_0}{z_0^2}\beta. \tag{110}$$

By noting that the FIM is symmetric and assembling all the elements I_{ij}, the FIM is in (38), which completes the proof.

Appendix C: Proof of Theorem 2

First of all, since the unknown phase φ only appears in the exponential terms of the first-order derivatives, it does not appear in the FIM, and the Fisher-information for x, y, and z dimensions remain the same. Next, we compute the cross-terms between φ-dimension and the other dimensions. Based on the derivatives in (45)-(46) and the definition of Fisher-information in (8), it can be shown that,

$$I_{14} = \iint_{x^2+y^2\le R} \mathcal{R}\{\Delta s_1 \left(\Delta s_4\right)^*\}\,\mathrm{d}x\mathrm{d}y = \frac{x_0 z_0}{2\lambda} g_3(4), \tag{111}$$

$$I_{24} = \iint_{x^2+y^2\le R} \mathcal{R}\{\Delta s_1 \left(\Delta s_4\right)^*\}\,\mathrm{d}x\mathrm{d}y = \frac{y_0 z_0}{2\lambda} g_3(4), \tag{112}$$

$$I_{34} = \iint_{x^2+y^2\le R} \mathcal{R}\{\Delta s_3 \left(\Delta s_4\right)^*\}\,\mathrm{d}x\mathrm{d}y = \frac{z_0^2}{2\lambda} g_3(4), \tag{113}$$

and

$$I_{44} = \iint_{x^2+y^2\leq R} |\Delta s_4|^2 \mathrm{d}x\mathrm{d}y = \frac{z_0}{4\pi} g_3(3), \tag{114}$$

where $g_3(n)$ is defined in (11).

Appendix D: Proof of Property 4

Inserting the expressions of I_{33}, I_{34}, I_{44} given in (101), (113), (114) to (54) and (55), we have the CRLBs for z-dimension and phase φ equal to

$$C_z = \left(\frac{z_0^3}{4\pi} \left(\frac{1}{4z_0^4} g_3(3) + \left(\frac{4\pi^2}{\lambda^2} - \frac{3}{2z_0^2} \right) g_3(5) + \frac{9}{4} g_3(7) \right) - \frac{\pi z_0^3}{\lambda^2} \frac{g_3^2(4)}{g_3(3)} \right)^{-1}, \tag{115}$$

and

$$C_\varphi = \left(\frac{z_0}{4\pi} g_3(3) - \frac{\pi z_0}{\lambda^2} \frac{g_3^2(4)}{\left(\frac{1}{4z_0^4} g_3(3) + \left(\frac{4\pi^2}{\lambda^2} - \frac{3}{2z_0^2} \right) g_3(5) + \frac{9}{4} g_3(7) \right)} \right)^{-1}, \tag{116}$$

respectively. For a terminal on the CPL, using the formula of $g_3(n)$ in (13) yields

$$g_3(3) = \frac{1}{z_0} - \frac{1}{R^2 + z_0^2}, \tag{117}$$

$$g_3(4) = \frac{1}{2} \left(\frac{1}{z_0^2} - \frac{1}{\sqrt{R^2 + z_0^2}} \right), \tag{118}$$

$$g_3(5) = \frac{1}{3} \left(\frac{1}{z_0^3} - \frac{1}{(R^2 + z_0^2)^{\frac{3}{2}}} \right), \tag{119}$$

$$g_3(7) = \frac{1}{5} \left(\frac{1}{z_0^5} - \frac{1}{(R^2 + z_0^2)^{\frac{5}{2}}} \right). \tag{120}$$

Inserting (117)-(120) back into (115) and (116), after some manipulations, the CRLBs for z-dimension and phase φ are in (56) and (57), respectively.

References

[1] S. Hu, F. Rusek, and O. Edfors, "The potential of using large antenna arrays on intelligent surfaces," *accepted in* IEEE Veh. Technol. Conf. (VTC-Spring), Sydney, 4-7 Jun. 2017, *arXiv preprint: 1702.03128*.

[2] S. Hu, F. Rusek, and O. Edfors, "Cramér-Rao lower bounds for positioning with large intelligent surfaces," *accepted in* IEEE Technol. Tech. Conf. (VTC-Fall), Toronto, Fall, 2017, *arXiv preprint: 1702.03131*.

[3] X. Huang, "Machine learning and intelligent communications," *Proc.* Int. Conf. Machine Learning and Intelligent Commun. (MLICOM), Shanghai, China, Aug. 27-28, 2016.

[4] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[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 Process. Mag.*, vol. 30, no. 1, pp. 40-60, Dec. 2012.

[6] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[7] J. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?", *IEEE J. Sel. Areas in Commun.*, vol. 32, no. 6, pp. 1065-1082, Jun. 2014.

[8] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey". *Comput. Netw., Elsevier*, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.

[9] A. Puglielli, N. Narevsky, P. Lu, T. Courtade, G. Wright, B. Nikolic, and E. Alon, "A scalable massive MIMO array architecture based on common modules," *Proc.* IEEE Int. Conf. Commun. (ICC), Workshop on 5G and beyond, May 2015.

[10] S. Al-Jazzar, J. Caffery, and H. R. You, "Scattering-model-based methods for TOA location in NLOS environments", *IEEE Trans. Veh. Technol.*, wol. 56, no. 2. pp. 583-593, Mar. 2007.

[11] S. Wu, D. Xu, and H. Wang, "Adaptive NLOS mitigation location algorithm in wireless cellular network," *Wireless Personal Commun.: An Int. J.*, vol. 84, no. 4, pp. 3143-3156, Oct. 2015.

[12] R. Mautz and S. Tilch, "Survey of optical indoor positioning systems," Int. Conf. Indoor Positioning and Indoor Navigation (IPIN), Nov. 2011.

[13] A. F. Molisch, *Wireless communications*, the second edition, Wiley-IEEE Press, 2010.

[14] S. M. Kay, "Fundamentals of statistical signal processing, volume I: Estimation theory," Prentice Hall signal processing series, 1993.

[15] W. Rudin, *Real and complex analysis*, the third edition, New York, McGraw-Hill Book Co., 1987.

[16] J. Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu, and F. Tufvesson, "Reciprocity calibration for massive MIMO: Proposal, modeling and validation", *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3042-3056, May 2017.

[17] V. Savic and E. G. Larsson, "Fingerprinting-based positioning in distributed massive MIMO systems," IEEE Veh. Technol. Conf. (VTC-Fall), Boston, USA, Sep. 6-9, 2015, pp. 1-5.

[18] M. Petersson, "Performance assessment of massive MIMO systems for positioning and tracking of vehicles in open highways," Master thesis, LiTH-ISY-EX–17/5049–SE, Linköping University, 2017.

[19] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "5G position and orientation estimation through millimeter wave MIMO," GLOBECOM Workshops (GC Wkshps), San Diego, USA, Dec. 6-10, 2015, pp. 1-6.

[20] S. Hu, F. Rusek, and O. Edfors, "Beyond Massive-MIMO: The potential of data-transmission with Large Intelligent Surfaces," submitted to *IEEE Trans. Signal Process., arXiv preprint: 1707.02887*, 2017.

[21] Y. Kalkan, "Cramer-Rao bounds for target position and velocity estimations for widely separated MIMO radar," *Radio Eng.*, vol. 22, no .4, pp. 1156-1161, 2013.

[22] A. J. Weiss, "On the accuracy of a cellular location system based on received signal strength measurements," *IEEE Trans. Veh. Technol.*, vol. 52, no. 6, pp. 1508-1518, Jun. 2003.

[23] C. Botteron, A. Host-Madsen, and M. Fattouche, "Cramer-Rao bounds for the estimation of multipath parameters and mobiles' positions in asynchronous DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 52, no. 4, pp. 862-875, Jan. 2004.

[24] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," Annual Joint Conf. of the IEEE Comput. and Commun., 2003, vol. 3, pp. 1734-1743.

[25] N. Vankayalapati, S. Kay, and Q. Ding. "TDOA based direct positioning maximum likelihood estimator and the Cramer-Rao bound," *IEEE Trans. Aerospace and Electron. Syst.*, vol. 50, no, 3, pp. 1616-1635, Dec. 2014.

[26] M. Veletić and M. Šunjevarić, "On the Cramer-Rao lower bound for RSS-based positioning in wireless cellular networks," *AEU-Int. J. Electron. and Commun.*, vol. 68, no. 8, pp. 730-736, Aug. 2014.

[27] Y. Qi, H. Kobayashi, and H. Suda, "On time-of-arrival positioning in a multipath environment," *IEEE Trans. Veh. Technol.*, vol. 55, no .5, pp. 1516-1526, Sep. 2006.

[28] Y. Qi and H. Kobayashi, "Cramer-Rao lower bound for geolocation in non-line-of-sight environment," *Proc.* IEEE Conf. Acoustics, Speech and Signal Process. (CASSP), Orlando, USA, May 2002, pp. 2473-2476.

[29] Y. Qi, H. Kobayashi, and H. Suda, "Analysis of wireless geolocation in a non-line-of-sight environment," *IEEE Trans. Wireless Commun.*, vol. 5, no. 3, pp. 672-681, Mar. 2006.

[30] Y. Shen and M. Z. Win, "Fundamental limits of wideband localization—Part I: A general framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4956-4980, Oct. 2010.

[31] Y. Shen and M. Z. Win, "On the accuracy of localization systems using wideband antenna arrays," *IEEE Trans. Commun.*, vol. 58, no. 1, pp. 270-280, Jan. 2010.

[32] P. Deng and P. Fan, "An AOA assisted TOA positioning system," *Proc.* Int. Conf. Commun. Technol., Beijing, China, 2000, vol. 2, pp. 1501-1504.

[33] L. Cong and W. Zhuang, "Hybrid TDOA/AOA mobile user location for wideband CDMA cellular systems," *IEEE Trans. Wireless Commun.*, vol. 1, no. 3, pp. 439-447, Jul. 2002.

[34] T. J. O'Shea and J. Hoydis, "An introduction to machine learning communications systems," *arXiv preprint: 1702.00832*, 2017.

[35] C. K. Seow and S. Y. Tan, "Non-line-of-sight localization in multipath environments,"*IEEE Trans. Mobile Comput.*, vol. 7, no. 5, pp. 647-660, May 2008.

[36] H. Miao, K. Yu and M. J. Juntti, "Positioning for NLOS propagation: Algorithm derivations and Cramer–Rao bounds," *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp. 2568-2580, Sep. 2007.

[37] K. Yu and Y. J. Guo, "Improved positioning algorithms for nonline-of-sight environments," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2342-2353, Jul. 2008.

[38] C. A. Balanis, *Antenna theory: Analysis and design*, the fourth Edition, New York, John Wiley & Sons, 2016.

[39] D. Zhu, J. Choi and R. W. Heath, "Auxiliary beam pair enabled AoD and AoA estimation in closed-loop large-scale millimeter-wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4770-4785, May 2017.