JACOBS
UNIVERSITY

Towards Reliable Low-Latency Massive Wireless Communications

by

Hiroki Iimori

a Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

Approved Dissertation Committee

Prof. Dr. Giuseppe Abreu
Jacobs University Bremen

Dr. Mathias Bode
Jacobs University Bremen

Prof. Dr.-Ing. Eduard A. Jorswieck
Technical University Braunschweig

Date of Defense: January 25th, 2022

Department of Computer Science & Electrical Engineering

Statutory Declaration

Family Name, Given/First Name	IIMORI, Hiroki
Matriculation number	20332270
What kind of thesis are you submitting: Bachelor-, Master- or PhD-Thesis	PhD-Thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

The Thesis has been written independently and has not been submitted at any other university for the conferral of a PhD degree; neither has the thesis been previously published in full.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde in der vorliegenden Form weder einer anderen Prüfungsbehörde vorgelegt noch wurde das Gesamtdokument bisher veröffentlicht.

25.01.2022, Hiroki Iimori

Date, Signature

Abstract

Owing to the ever-growing volume of mobile traffic and wireless devices as shown in network traffic reports over the last decades, the wireless mobile technology is required and expected to offer continuous improvements on the major communication performance such as data rate and connection density. Besides expanding such communication performance, various application-centric Quality of Service (QoS) requirements have been raised from a wide range of application fields, imposing on the research community the need of diverse wireless solutions to satisfy such heterogeneous QoS requirements.

In light of the above, this dissertation intends to contribute to the aforementioned trend by addressing three essentials in future wireless systems: *massive connectivity*, *low-latency*, and *reliability*, offering the corresponding algorithm design(s) and quantitative analyses to evaluate the performance.

Massive connectivity aims to increase the connection density of wireless communication systems serving a certain area. Given the fact that the number of mobile devices will likely grow continuously over the next decade subject to the limitation on the available wireless resources, the future wireless systems are expected to resort to non-orthogonal transmission, in which the number of available resources at the receiver is below that at the transmitter. One of the main challenges in this context is an efficient receiver design enabling symbol detection even under overloaded (*i.e.*, underdetermined) conditions. To this end, we propose a new symbol detection framework inspired by compressed sensing (CS), which is composed of an adaptive tight approximation of the ℓ_0-norm and a quadratic transformation via fractional programming (FP), leading to a generalization of the conventional linear estimators (*i.e.*, zero-forcing (ZF) and linear minimum mean square error (LMMSE)) in terms of adherence to the prescribed discrete constellation set.

Another challenge in uplink due to the increase in the number of mobile devices is the resource overhead required to jointly acquire channel state information (CSI) and sporadic user activity, which scales with the number of potential users, resulting in unacceptable data transmission delay (*i.e.*, latency). An emerging approach to address this issue is grant-free non-orthogonal multiple access (NOMA), in which the

CSI and user activity patterns are jointly estimated by taking advantage of the sparsity due to the sporadic access as well as non-orthogonal sequences transmitted by the users. In this context, we aim to achieve the following two different objectives. The first objective is to enhance the per-user throughput in grant-free access systems by jointly performing active user detection (AUD), channel estimation (CE), and multi-user detection (MUD) while exploiting the pseudo-orthogonality of randomly generated data sequences. The latter is to handle a peculiar phenomenon in distributed multiple-input multiple-output (MIMO) architectures, that is, spatial non-stationarity, which is the fact that the signal from each user is received only at part of the distributed antenna array, resulting in the sparsity in the antenna domain (*i.e.*, array activity), while jointly addressing AUD and CE in a grant-free fashion. We tackle the aforementioned two challenging estimation problems by means of bilinear Bayesian inference, proposing a tailored estimation algorithm for each problem.

The demand for higher data rate and the growth of the number of mobile devices jointly pose a dearth of spectrum resources in sub-6GHz frequency bands, which motivates the use of higher frequencies such as millimeter wave (mmWave) bands. Although there has been great progress in combating major challenges of mmWave systems including the severe propagation loss and efficient radio frequency (RF) design, a remaining challenge is the susceptibility of mmWave signals to random path blockage. To maintain the reliability of mmWave systems against such random blockage, we propose a novel stochastic-optimization-based framework to mitigate the QoS violation (*i.e.*, outage probability) by means of cooperative beamforming in distributed wireless architectures.

Numerical performance assessments via computer simulations are offered to evaluate the aforementioned proposed methods, illustrating their advantages against existing counterparts.

Acknowledgements

I have been lucky enough to be surrounded by people who have always supported and encouraged me to pursue my passion. This work would not have been possible, if any one of them is missing.

My sincere gratitude goes to my supervisor, Prof. Dr. Giuseppe Thadeu Freitas de Abreu, for his dedicated guidance, strong support, and freedom to pursue my curiosity over the course of the last six years including the PhD study in Jacobs University Bremen as well as my master and bachelor studies in Japan. His attitude towards both work and life has always brought to me a new perspective that helped me enjoy not just research but life.

I would also like to thank the dissertation committee members, Dr. Mathias Bode and Prof. Dr.-Ing. Eduard A. Jorswieck for agreeing to be on my dissertation committee.

Besides, I am also grateful to Dr. David González G., Prof. Dr. Koji Ishibashi, and Dr. Osvaldo Gonsa, who have mentored and supported me during the PhD study by sharing their skilled knowledge and expertise. Their insights have always been of incredible value to me. In addition, I would like to extend my thanks to the amazing collaborators and colleagues: Dr. Andre Saito Guerreiro, Prof. Dr. George C. Alexandropoulos, Hyeon Seok Rou, Kengo Ando, Dr. Omid Taghizadeh, Dr. Razvan-Andrei Stoica, Sota Uchimura, Shuto Fukue, Takanori Hara, and Prof. Dr. Takumi Takahashi, who have always shared wonderful ideas and insights, leading to better research outputs. The joint internatinal research projects with the aforementioned researchers helped me grow as a global researcher.

Also, I would like to take this opportunity to express my profound appreciation and gratitude to University of Toronto, Canada, and Prof. Dr. Wei Yu as well as his research lab members for welcoming me as a visiting researcher and offering perspicacious advice. Experiences and insights gained during the stay in Toronto were valuable to me, resulting in several research outcomes that indeed form part of this dissertation. I am indebted to Dr. Szabolcs Malomsoky and Dr. Roy Timo for providing me with industrial perspectives and friendly collaborations. I would like to thank Prof. Dr. Moe Win, who taught me attitude to research during his stay in Japan.

Last but not least, my special thanks goes to my family members, Shiro, Fumiko, and Akane as well as my grand parents for their continuous support throughout my life, and to Kanako who has made my life much warmer and more colorful over the last eight years.

To my late father,
Shiro Iimori.

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

1G	first generation
2G	second generation
3G	third generation
4G	fourth generation
5G	fifth generation
5G+	beyond fifth generation (5G)
6G	sixth generation
ADAM	adaptive moment estimation
ADC	analog-to-digital converter
ADMM	alternating direction method of multipliers
AER	activity error rate

AF	amplify-and-forward
AoD	angles of departure
AoI	age of information
AMP	approximate message passing
AP	access point
AR	augmented reality
ARP	Autoradiopuhelin
AUD	active user detection
AWGN	additive white Gaussian noise
BER	bit-error rate
BiGAMP	bilinear generalized approximate message passing
BiGaBP	bilinear Gaussian belief propagation
BP	belief propagation
BS	base station
BSGD	block-coordinate stochastic gradient descent
CDF	cumulative density function
CDMA	code division multiple access
CE	channel estimation
CF-MIMO	cell-free MIMO
CFO	carrier frequency offset
CoMP	coordinated multipoint
CPU	central processing unit
CS	compressed sensing
CSI	channel state information
CSIDCO	complex sequential iterative decorrelation via convex optimization (SIDCO)
DAC	digital-to-analog converter
DCC	dynamic cooperation clustering
DFT	Discrete Fourier Transform
DoF	degrees-of-freedom
E_b/N_0	energy per bit to noise power spectral density ratio
EE	energy efficiency
EM	expectation-maximization
eMBB	enhanced mobile broadband
EP	expectation propagation
ERM	empirical risk minimization
ERTS	enhanced reactive tabu search
FA	false alarm
FDMA	frequency division multiple access
flops	floating point operations

FP	fractional programming
FPGA	field programmable gate array
GaBP	Gaussian belief propagation
GD	gradient descent
GIGD	graph-based iterative Gaussian detector
IDD	iterative detection-and-decoding
IDLS	iterative discrete least square
i.i.d.	independent and identically distributed
IoT	Internet of Things
IRS	intelligent reflecting surfaces
ISI	inter-symbol interference
ITU-R	International Telecommunication Union Radiocommunication Sector
IW-SOAV	iterative weighted-SOAV
JACDE	joint activity, channel and data estimation
JACE	joint activity and channel estimation
JCDE	joint channel and data estimation
KKT	Karush Kuhn Tucker
KPI	key performance indicator
LASSO	least absolute shrinkage and selection operator
LDPC	low-density parity-check code
LLR	log-likelihood ratio
LMMSE	linear minimum mean square error
LoS	line-of-sight
LRMC	low-rank matrix completion
LS	least squares
LSP	log-sum-penalty
MAP	maximum a posteriori
MC	matrix completion
MCPP	Matérn-cluster point process
MD	miss-detection
MIMO	multiple-input multiple-output
ML	maximum likelihood
MMSE	minimum mean squared error
mMTC	massive machine-type communications
MMVABP	multiple measurement approximate belief propagation
MMV-AMP	multiple measurement vector approximate message passing
mmWave	millimeter wave
MNS	minimum norm solution
MRT	maximum ratio transmission

MSE	mean-square error
MSGD	mini-batch stochastic gradient descent
MUD	multi-user detection
MUI	multiuser interference
NLoS	non-line-of-sight
NMSE	normalized-MSE
NN	nuclear norm
NOMA	non-orthogonal multiple access
NP	non-deterministic polynomial-time
NR	new radio
OFDM	orthogonal frequency division multiplexing
OFDMA	orthogonal frequency division multiple access
OMA	orthogonal multiple access
OMP	orthogonal matching pursuit
OTFS	orthogonal time frequency space
PA	power amplifier
PAM	pulse amplitude modulation
PDF	probability density function
PG	proximal gradient
PIC	parallel interference cancellation
PMF	probability mass function
PPP	Poisson point process
PSK	phase shift keying
PSTN	public switched telephone network
QAM	quadrature amplitude modulation
QCQP-1	quadratically constrained quadratic program with one convex constraint
QCSIDCO	quadratic complex SIDCO (CSIDCO)
QoS	Quality of Service
QP	quadratic program
QPSK	quadrature phase shift keying
QT	quadratic transform
RAM	random-access memory
RF	radio frequency
SBR	simplicity-based recovery
SCCR	sum of concave-over-convex ratios
SCSR	sum of complex sparse regularizers
SD	sphere decoding
SDP	semidefinite programming
SDR	software defined radio

SE	spectrum efficiency
SGD	stochastic gradient descent
SIC	soft interference cancellation
SIDCO	sequential iterative decorrelation via convex optimization
SINR	signal-to-interference-plus-noise ratio
SISO	single-input single-output
SNR	signal-to-noise ratio
SOAV	sum of absolute value
SotA	state-of-the-art
SRM	sum-rate maximization
SVD	singular value decomposition
SVT	singular value thresholding
TDD	time division duplex
TDMA	time division multiple access
TD-SCDMA	time-division synchronous code division multiple access
THz	terahertz
UE	user equipment
ULA	uniform linear array
UPA	uniform planar array
URA	unsourced random access
URLLC	ultra reliable low-latency communications
VR	visibility region
WB	Welch bound
WCDMA	wideband code division multiple access
XL-MIMO	extra large MIMO
ZF	zero-forcing

Notation

Unless otherwise specified the vector notation used in the sequel assumes column orientation.

$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\boldsymbol{X}$	matrix notation
$\boldsymbol{x}$	vector notation

$(\cdot)^{\mathrm{H}}$	conjugate Hermitian operation
$(\cdot)^{\mathrm{T}}$	transposition operation
$(\cdot)^{*}$	element-wise complex conjugate operation
$\mathbf{1}_M$	all-ones vector of length M
$\mathbf{0}_M$	all-zeros vector of length M
$\mathbf{I}_M$	$M \times M$ identity matrix
$\mathbf{0}_{M\times N}$	$M \times N$ all-zeros matrix
$\boldsymbol{X}^{-1}$	inverse of the square matrix $\boldsymbol{X}$ (if it exists)
$\mathrm{diag}\,(\boldsymbol{x})$	diagonal matrix with $\boldsymbol{x}$ on the main diagonal
$\mathrm{diag}\,(\boldsymbol{X})$	column vector extracted from the main diagonal of matrix $\boldsymbol{X}$
$\mathrm{rank}\,(\boldsymbol{X})$	rank of matrix $\boldsymbol{X}$
$\mathrm{vec}\,(\boldsymbol{X})$	vectorization operation of the matrix
$\mathrm{vec}^{-1}(\boldsymbol{x})$	reciprocal operation of vectorization
$\left\|\boldsymbol{x}\right\|_p$	ℓ_p-norm of vector $\boldsymbol{x}$ with $p \geq 0$
$\left\|\boldsymbol{X}\right\|_2$	spectral norm of the matrix $\boldsymbol{X}$
$\left\|\boldsymbol{X}\right\|_{\mathrm{F}}$	Frobenius norm of the matrix $\boldsymbol{X}$
$\circ$	Hadamard entry-wise product
$\mathbb{E}_{\boldsymbol{x}}\left[\cdot\right]$	expectation operation over the random variable $\boldsymbol{x}$
$\mathfrak{R}\left[\cdot\right]$	real part entry-wise extractor of complex quantities
$\mathfrak{I}\left[\cdot\right]$	imaginary part entry-wise extractor of complex quantities

(page intentionally left blank)

Chapter 1

Introduction

1.1 General Background

The wireless technology becomes the core of networks and information infrastructures, which play the essential role of connecting people worldwide and runnning social and business activities as highlighted by unique challenges imposed by the outbreak of the COVID-19 pandemic. One of the earliest ancestors of the mobile telephone service is a car phone, in which equipments are often mounted in an automobile and was used as part of the public switched telephone network (PSTN). These services were introduced in different countries such as Finland (Autoradiopuhelin (ARP)), Germany (A-Netz) and the United States.

The concept of cellular networks was introduced by first generation (1G) in the 1980's [1], where a voice call is analogly modulated to a higher frequency by means of frequency division multiple access (FDMA) with circuit-switched control. The digitalization took place in the second generation (2G) cellular networks thanks to advances of digital circuits, which enables not only digitally-encrypted phone calls but also higher spectrum efficiency and basic data transmission such as text messages. The 2G mobile system mainly relies on time division multiple access (TDMA)-based transmission, while starting to explore the code domain (*i.e.*, code division multiple access (CDMA)). Although the latter technology was not the common choice in 2G networks, its variants (*e.g.*, wideband code division multiple access (WCDMA), CDMA2000, and time-division synchronous code division multiple access (TD-SCDMA)) had been adopted and deployed in the third generation (3G) systems from 2001, aiming at a high-speed data communication up to the effective rate of a few megabits per second [2]. In response to demands for higher data rate owing to emerging information applications, the wireless communications industry and adademic community concentrated their efforts on streching the communication capability, leading to the fourth generation (4G) cellu-

lar networks featured by a combination of orthogonal frequency division multiplexing (OFDM) and multiple-input multiple-output (MIMO) technologies.

Unlike the aforementioned precedents focusing merely on improving the communication capabilities, the fifth generation (5G) networks currently being deployed in many countries aim not only to further improve the latter but also to satisfy stringent Quality of Service (QoS) requirements arising from wide-ranging fields such as augmented reality, Internet of Things (IoT), and vehicle and financial realtime applications, which may go beyond the capability of the previous generations. In order to throw light on its definition, the International Telecommunication Union Radiocommunication Sector (ITU-R) has defined three main usage areas expected to be offered by the 5G systems [3, 4]:

- enhanced mobile broadband (eMBB) aiming at human-centric applications with requirements of high data rate access such as augmented/virtual reality and 3D gaming.

- ultra reliable low-latency communications (URLLC) addressing latency-stringent applications with reliable connection such as industrial automation and intelligent transportation.

- massive machine-type communications (mMTC) focusing on massive connectivity for a large number of devices typically with short packet communication of non delay sensitive data, such as smart cities.

From a technical perspective, in order to iron out the spectral clog in the sub-6GHz bands, the 5G wireless systems have extended its operating frequency to higher spectral bands, leading to a wider bandwidth[1] and higher throughput. At the moment, 5G waveforms leverage the same multiplexing technology as in 4G (*i.e.*, OFDM) both in downlink and uplink, which is envisioned to be exploited up to the 71 GHz bands. This extension comes along with the advanced directional signal transmission to compensate the severe attenuation of high frequency bands such as the ones mentioned above, which is supported and enabled by the massive MIMO technology. In addition to the above, another key feature of 5G new radio (NR) is the channel coding schemes: the low-density parity-check code (LDPC) coding is adopted for the data channels while the polar coding is utilized for the control channels. Also, several other features such as enhanced mobility, high accurate 5G localization, dynamic spectrum sharing, and energy efficiency continue to be discussed and incorporated, aiming at standarization in future releases such as Release 18 planned to be initiated at the beginning of 2022 and frozen by mid-2023.

[1]Release 16 supports up to 400 MHz carrier bandwidth with 120 kHz subcarrier spacing [5].

Future Directions: Drivers and Requirements

In the context briefly described above and given the fact that the 5G networks are now in deployment phase, the research community is starting to shift their attention to future wireless systems including beyond 5G (5G+) and sixth generation (6G) systems. Although it is challenging and perilous to fully determine applications and required technological enablers in such future wireless systems at the time of writing, it is highly predictable that the volume of wireless traffic and mobile devices will continuously increase in the 2020's as estimated by several existing reports [6, 7]. At the moment, a recent report [8] revealed that the total global mobile data traffic in Q1 2021 exceeded 66EB, which corresponds to the year-on-year growth rate of 46 percent compared with that of Q1 2020, confirming the accuracy of the aforementioned trafffic forecasts. This at least implies the need of further improvements of the key technical requirements in 5G systems such as peak data rate, connection density, latency, and energy efficiency. Aiming to shed light on the required quality of 6G key performance indicators (KPIs) such as those aforementioned, recent survey articles (*e.g.*, [9, 10]) have provided quantitative analyses on the latter. To name but a few, the peak data rate and user-experience rate is envisioned to reach 1 tera and giga bits per second, respectively, while that in 5G networks corresponds to 20 giga bits per second and 100 mega bits per second, respectively. As for the connection density, the authors of [9, 10] suggested tenfold increase compared with the connection density envisaged in 5G systems, namely, ten million devices within square kilometre.

In a similar way to the aforementioned extension of 5G networks, it is also expected in future wireless systems to integrate the conventional 5G usage scenarios (*i.e.*, eMBB, URLLC, and mMTC) such that the future application areas at least in the 6G era include their possible combinations such as ultra *reliable low-latency massive communications*, reliable enhanced mobile broadband, or even a combination of the three. The potential 6G use cases envisioned in [11] such as high-fidelity holographic society, digital twin, and pervasive intelligence in fact impose a joint requirement of high throughput, low-latency, high-reliablity and/or massive connectivity, which can not be fully covered by the 5G networks.

In addition to the further expansion of the 5G KPIs, data-oriented communication aspects such as security [12] and semantics [13] may be also incorporated as one of the new 6G KPIs, imposing a quantitative measurement to properly determine the required quality of such new KPIs. One of the typical strategy to measure the security, for example, is the secrecy rate that is a signal-to-noise ratio (SNR) advantage over the eavesdropper. On the other hand, how to measure the semantics of data is still controversial, although there are existing approaches to (at least partially) measure the latter such as the age of information (AoI) [14] and semantic entropy [15].

As briefly illustrated above, future wireless networks need to cover more heterogeneous requirements than those in the previous generations including 5G; thus, the research community is required to offer diverse solutions to many different usage scenarios. Consequently, technical enablers in physical-layer are also expected to be miscellaneous according to different potential applications, which should be a combination of different spectrum-, protocol-, and system-level technologies. From a view point of the carrier frequency, for instance, sub-6GHz, millimeter wave (mmWave), terahertz (THz), and visible light bands will be the spectra of choice in 6G, which should be chosen according to the bandwidth requirement. In turn, distributed passive/active MIMO setups including cell-free MIMO (CF-MIMO), extra large MIMO (XL-MIMO), and intelligent reflecting surfaces (IRS) may be considered from a system perspective, while non-terrestrial networks can also be an infrastructure-level enabler to further increase the coverage worldwide. In other words, a wide variety of solutions aiming at different possible combinations of KPIs in different scenarios need to be served by the research community.

1.2 Thesis Contributions and Outline

1.2.1 Contributions

In light of the above, this thesis contributes to the aforementioned trend by addressing part of the key KPIs in future wireless systems (*i.e.*, massive connectivity, latency, and reliablity). To elaborate, the contributions made during the course of our research are listed below:

Massive Connectivity

Due to the ever-growing number of mobile devices and the limited amount of available wireless resources, the future wireless systems highly-likely adopt non-orthogonal transmission (*e.g.*, non-orthogonal multiple access (NOMA) and overloaded MIMO), in which the number of resources at the effective transmitter exceeds that at the receiver. In this context, one of the main bottlenecks at the receiver is how to *efficiently* detect the transmitted symbols under the underdetermined condition raised by the non-orthogonality, in which the traditional well-adopted linear estimators such as linear minimum mean square error (LMMSE) and zero-forcing (ZF) methods lead to an unacceptable high error floor due to the fact that they are not aware of the discreteness of the transmitted symbols. Although the maximum likelihood (ML) and tree-search-based approaches including sphere detectors are capable of effectively finding the transmitted symbols even in such a severe condition, their scalability is questionable due to high-complexity, which contradicts the trend of massive commu-

nications. To overcome this difficulty, we propose a new detection framework that offers a reasonable compromise between the low-complexity implementation-friendly nature of linear detectors and the detection performance of the brute-force search. To elabotate further, the contributions can be compactly listed in the sequel:

- a new compressed sensing (CS)-inspired regularizer that facilitates searching overlapped transmitted symbols while adhering to the prescribed discrete constellation is proposed, where convexification is performed so that the resultant receiver becomes a generalization of the conventional linear detectors (*i.e.*, LMMSE and ZF).

- the proposed regularizer is then applied not only to a conventional ideal MIMO setup, where the channel is perfectly known and the hardware imposes no impairments, but also to an non-ideal MIMO channel with channel state information (CSI) and hardware imperfection in order to illustarte the flexibility of the proposed approach.

Latency

Another challenge in uplink stemming from the increase on the number of mobile devices is the resource overhead required for CSI acquisition and active user identification due to their sporadic traffic. Existing multiple access schemes (*i.e.*, orthogonal multiple access (OMA)) impose a prohibitive amount of overhead as the number of potential users increases, since an orthogonal pilot sequence needs to be allocated simply to a single user. A promising approach to address this issue is a combination of grant-free NOMA and CS-based detection mechanisms, where active user identification and CSI acquisition are jointly performed by taking advantage of the sparsity due to the inactive users (*i.e.*, users transmitting zeros). Following this trend, we offer the following contributions:

- we proposed a frame-theoretic pilot design so as to efficiently non-orthogonalize the pilot sequence.

- we proposed a novel grant-free access scheme based on the bilinear inference, which jointly and efficiently takes advantage of the non-orthogonal pilot and subsequent data sequences, leading to higher spectrum efficiency per user, while being compatible with either centralized or distributed MIMO setups.

- we also proposed another grant-free access method based on the bilinear inference framework with the aim of addressing spatial non-stationarity issues raised in the context of XL-MIMO systems.

Reliability

The increasing demands for higher data rate and the growth of the number of potential users jointly impose a lack of spectrum resources in sub-6GHz bands. To tackle this upcoming shortage, mmWave technology has been intensively studied in the last decade, offering excellent solutions to major challenges of mmWave systems such as the severe signal attenuation and high expenditure for mmWave hardware. A fundamental challenge that remains, however, is the susceptibility of mmWave signals to random path blockage due to the mobility of the surroundings. This leads to unpredictable channel uncertainties, affecting the performance reliablity. Our contribution to the reliablity issue in mmWave systems subject to random path blockage is as follows:

- we proposed novel outage-reduced robust beamforming algorithms based on a stochastic optimization framework so as to preserve, as much as possible, the QoS requirements even under such random path blockage.

Besides the above high-level descriptions of the contributions, more technical details of each contribution will be offered later in an itemized manner in the corresponding section. It is also worth-mentioning that most of the above contributions have been disseminated through publications in peer-reviewed IEEE journals and conferences. A list of the publications made during the course of this PhD study is shown in Appendix D.

1.2.2 Outline

This dissertation is organized as follows:

- In Chapter 1, we provide a historical background of the mobile communications and envisage the future direction, illustrating the need of diverse wireless solutions to satisfy the upcoming more heterogeneous requirements in future wireless systems. We then offer a list of high-level descriptions of the contributions made by the dissertation. Lastly, mathematical preliminaries to the techniques utilized in the subsequent sections are offered.

- In Chapter 2, we propose a new regularization approach for inverse problems with discrete inputs, aiming to generalize the conventional linear estimators with adherence to the prescribed digital (discrete) constellation. Then a wide range of applicability of the proposed regularizer is illustrated by showing the effectiveness of the latter in different signal processing applicaitons. One of the key applications offered in this section is the receiver design of NOMA systems, in which it is shown that the proposed regularizer indeed contributes to the massive connectivity requirement.

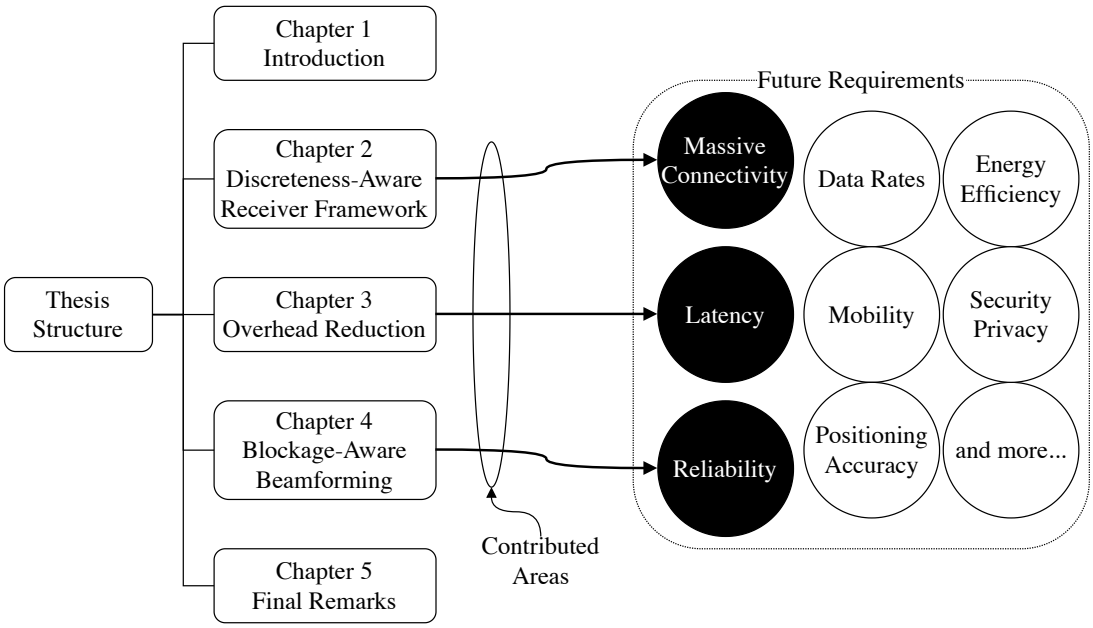


Figure 1.1: Structure and contributed areas of the thesis

- In Chapter 3, we turn our attention to latency issues in the uplink of massive devices, where the overhead for CSI acquisition and active user identification is considered one of the major challenges. Following the trend of distributing access points (APs) over a service area, we consider two emerging distributed MIMO scenarios (*i.e.*, CF-MIMO and XL-MIMO), for which we propose a bilinear inference framework to jointly address CSI acquisition and active user identification issues. In particular, the proposed algoithm for CF-MIMO systems is designed to be joint activity, channel and data estimation (JACDE) with the aim to enhance the per-user throughput while reducing the overhead. In contrast, the one for XL-MIMO is designed to be joint activity and channel estimation (JACE) with the aim to tackle its peculiarity (*i.e.*, spatial non-stationarity) while focusing on the overhead reduction.

- In Chapter 4, we tackle the reliablity issue of mmWave systems, which is posed by the susceptibility of high frequency signals to random path blockage. To this end, we propose a stochastic optimization framework to minimize the QoS violation (*i.e.*, the QoS outage probability), resulting in a blockage-robust waveform design for such systems.

- In Chapter 5, we conclude the thesis and offer a list of potential future works as meaningful extensions of the research carried out during the PhD study.

Summarizing the above, for the sake of visualization, Figure 1.1 illustrates the structure of the thesis, highlighting that the thesis addresses three major requirements in future wireless systems, namely, massive connectivity, latency, and reliability.

1.3 Preliminaries

In this section, we offer technical descriptions of essential mathematical tools, which will be leveraged throughout the rest of the thesis. The key ingredients utilized in this thesis are mainly two-fold: optimization theory and Bayesian inference, for each of which fundamentals and pragmatic techniques are described below.

1.3.1 Optimization Techniques

Fractional Programming

Convex and non-convex optimization techniques have been intensively leveraged in the wireless communications literature over the last decade [16–18], due to their flexiblity and affinity with the linearly-represented signal processing in wireless communication systems. In particular, fractional programming (FP) has been recognized as a powerful tool to address a vast array of problems in wireless systems such as resource allocation and channel estimation, which is a branch of optimization theory concerning optimization of (multiple) fractional functions.

When it comes to designing wireless systems, the signal-to-interference-plus-noise ratio (SINR) at the receiver is inextricably related with the latter as the corresponding achievable data rate is characterized by the logarithm of SINR. Since SINR is a ratio between the power of the incoming signal of interest and that of the interference-plus-noise component [19], namely,

$$\text{SINR} \triangleq \frac{\text{Power of Received Intended Signal}}{\text{Power of Interference Signals} + \text{Noise Power}}, \tag{1.1}$$

FP is a natural and suited solution to optimization problems involving SINR expressions.

Furthermore, the efficiency of data transmission in wireless communications (*i.e.*, energy efficiency (EE)) is often concerned in academic publications as well as among industries, which measures the gained utility per unit cost, that is, bits per Joule in wireless communications. In this context, the EE expression is defined as a fraction between the spectrum efficiency (SE) and the consumed energy, which is given by

$$\text{EE} \triangleq \frac{\text{Data Rate [bit/s]}}{\text{Energy Consumption [Joule/s]}} \text{ [bit/Joule]}, \tag{1.2}$$

demonstrating the convenience of FP when optimizing EE-aware wireless applications [20].

As illustrated above, it is concluded that FP is suited for a wide range of problems raised in wireless communications. In case of single-ratio scenarios, FP in general

mathematically aims to solve the optimization problem of

$$\underset{\boldsymbol{x}\in\mathbb{C}^N}{\text{maximize}} \quad \frac{A(\boldsymbol{x})}{B(\boldsymbol{x})} \tag{1.3a}$$

$$\text{subject to} \quad \boldsymbol{x} \in \mathcal{X}, \tag{1.3b}$$

where $\mathcal{X} \in \mathbb{C}^N$ is a nonempty and compact convex set, $A(\boldsymbol{x}) : \mathbb{C}^N \to \mathbb{R}_+$ is a non-negative function, and $B(\boldsymbol{x}) : \mathbb{C}^N \to \mathbb{R}_{++}$ is a strictly positive function.

It is naturally conceived from equation (1.3) that the objective ratio function can not be guaranteed to be concave even in case of a simple linear-over-linear function. In order to address this problem at low-complexity, one can leverage the Dinkelbach's algorithm, whose convergence guarantee has been shown *e.g.*, in [20, Sec. 3], which transforms equation (1.3) into the following parametrized formulation

$$\underset{\boldsymbol{x}\in\mathbb{C}^N}{\text{maximize}} \quad A(\boldsymbol{x}) - \beta B(\boldsymbol{x}) \tag{1.4a}$$

$$\text{subject to} \quad \boldsymbol{x} \in \mathcal{X}, \tag{1.4b}$$

with a newly-introduced auxiliary variable β that is iteratively updated as

$$\beta = \frac{A(\boldsymbol{x})}{B(\boldsymbol{x})}, \tag{1.5}$$

where the numerator and denominator are computed with the solution obtained at the previous iteration round.

It is worth-mentioning that if the single-ratio function given in equation (1.3) is concave-over-convex, the Dinkelbach's algorithm indeed converges to the global optimum. As stated in [20, 21], however, the Dinkelbach's algorithm can not be *directly* generalized to solve the sum-of-ratios problem given by

$$\underset{\boldsymbol{x}\in\mathbb{C}^N}{\text{maximize}} \quad \sum_i \frac{A_i(\boldsymbol{x})}{B_i(\boldsymbol{x})} \tag{1.6a}$$

$$\text{subject to} \quad \boldsymbol{x} \in \mathcal{X}, \tag{1.6b}$$

where the numerator and denominator of each ratio are indexed with $i \in \mathbb{Z}$.

It is shown in [21, Sec. II] that the function value of the transformed objective upon convergence is not necessarily the same as that of the original objective in case of multiple-ratios problems. In order to tailor the transform to meet the equivalence in the objective value while preserving the convergence guarantee and optimality conditions of the Dinkelbach's algorithm, a new FP technique, dubbed as quadratic transform

(QT), has been reported in [21], which transform equation (1.6) into

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad \sum_i 2\beta_i \sqrt{A_i(\boldsymbol{x})} - \beta_i^2 B_i(\boldsymbol{x}) \tag{1.7a}$$

$$\text{subject to} \quad \boldsymbol{x} \in \mathcal{X}, \tag{1.7b}$$

where

$$\beta_i \triangleq \frac{\sqrt{A_i(\boldsymbol{x})}}{B_i(\boldsymbol{x})}, \tag{1.8}$$

is a scaling quantity iteratively updated based on the solution obtained at the previous iteration and designed to ensure that, at that pivot point, the original objective function in equation (1.6) is equivalent to the transformed function given in equation (1.7a).

The convergence and optimality analyses of the QT have been offered in [21] for concave-over-convex scenarios. It has also been reported in *e.g.*, [20, Alg. 4] that a Dinkelbach-like algorithm can be utilized to solve sum-of-ratios problems. However the latter is not applicable to convex-quadratic-over-convex-quadratic problems as is the case with SINR-involved problems, which can be directly tackled via the QT as a sequence of convex programs with convergence guarantee to a stationary point thanks to the square root applied to the numerator.

It is also worth-mentioning that there are global optimization frameworks to seek the global optimal of sum-of-ratios problems by means of, *e.g.*, branch-and-bound search [22–24], which is however limited to a problem with small search spaces due to their prohibitive complexity. Having said that, a hybrid method that leverages well-established optimization frameworks and quantum computing to overcome the combinatorial complexity has recently been shown in [25], implying that a global search for the above sum-of-ratios problem may be feasible with reasonable run time in the near future.

Last but not least, the above FP techniques will be exploited in Section 2.

Wirtinger Derivatives

Optimization in signal processing and communications often encounters a maximization (minimization) problem of a real-valued utility (cost) objective function with complex-valued variables. For instance, a waveform design in wireless communications can be optimized by maximizing the (real-valued) system throughput with respect to complex-valued quantities such as beamforming vectors/matrices controlled at the transmitter and/or receiver. In order to solve such problems, it is often required to calculate the gradient with respect to a complex variable, for which a useful tool often utilized in engineering is the Wirtinger derivative. In this section, we intend to briefly

Table 1.1: Note of the conjugate Wirtinger derivative of key functions.

Function: $f(\boldsymbol{z})$	Conjugate Wirtinger Derivative: $\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}^*}$
$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{a}$	$\boldsymbol{0}$
$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{z}^* = \boldsymbol{z}^{\mathrm{H}}\boldsymbol{a}$	$\boldsymbol{a}$
$\boldsymbol{z}^{\mathrm{H}}\boldsymbol{z} = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}^*$	$\boldsymbol{z}$
$\boldsymbol{z}^{\mathrm{H}}\boldsymbol{A}\boldsymbol{z} = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{z}^*$	$\boldsymbol{A}\boldsymbol{z}$
$\frac{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{A}\boldsymbol{z}}{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{B}\boldsymbol{z}}$	$\frac{\boldsymbol{A}\boldsymbol{z}}{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{B}\boldsymbol{z}} - \frac{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{A}\boldsymbol{z}}{(\boldsymbol{z}^{\mathrm{H}}\boldsymbol{B}\boldsymbol{z})^2}\boldsymbol{B}\boldsymbol{z}$

describe fundamentals of the Wirtinger derivative, collecting the latter of several key functions, which will be exploited later in this thesis. Since the purpose of this section is not to offer a thorough review of the Wirtinger derivative but to provide a concise set of notes for quick reference, interested readers are kindly referred to [26, 27] for more technical details.

In what follows, we briefly introduce the need of the Wirtinger derivative in signal processing applications by showing the practical inconvenience of complex differentiability, summarizing the section by collecting a set of the conjugate Wirtinger derivative of some major functions considered in this thesis. Starting from the definition, the derivative of function $f : \mathbb{C} \to \mathbb{C}$ at a point z_0 is given by [28]

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}. \tag{1.9}$$

If the limit exists, the function f is said to be complex-differentiable at z_0. The complex-differentiable functions are also known as holomorphic functions, which must satisfy the Cauchy–Riemann equations

$$\frac{\partial \Re\{f\}}{\partial \Re\{z\}} = \frac{\partial \Im\{f\}}{\partial \Im\{z\}}, \tag{1.10a}$$

$$\frac{\partial \Re\{f\}}{\partial \Im\{z\}} = -\frac{\partial \Im\{f\}}{\partial \Re\{z\}}. \tag{1.10b}$$

Although the theory of complex differentiability and holomorphic functions has been well established in complex analysis, their applicability to practical systems is limited, since many key real-valued objective functions with complex variables fail to satify the Cauchy–Riemann equations and complex-valued objective functions make no sense in many signal processing applications including communications. To tackle this bottleneck, the Wirtinger calculus has been commonly used for such optimization

problems, which suggests to consider z and its conjugate z^* as separate variables of the function f. Thanks to this change of coordinate system, partial derivatives of the function can be written as

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial \Re\{z\}} - j \cdot \frac{\partial}{\partial \Im\{z\}}\right), \tag{1.11a}$$

$$\frac{\partial}{\partial z^*} = \frac{1}{2}\left(\frac{\partial}{\partial \Re\{z\}} + j \cdot \frac{\partial}{\partial \Im\{z\}}\right). \tag{1.11b}$$

Equation (1.11) is the well-known definition of the Wirtinger derivatives of first order. When it comes to optimization, the update rule of gradient descent algorithms to minimize a real-valued objective loss function f can be simply written as [27]

$$z^{i+1} = z^i - \alpha\frac{1}{2}\left(\frac{\partial f}{\partial \Re\{z\}} + j \cdot \frac{\partial f}{\partial \Im\{z\}}\right) = z^i - \alpha\frac{\partial f}{\partial z^*}, \tag{1.12}$$

with a step size α and i denoting the iteration index.

Interestingly, the above equation suggests that only the conjugate Wirtinger derivative is needed to update complex variables to minimize (maximize) a real-valued objective function by means of gradient-based algorithms. Although we consider a scalar case in the above paragraphs for the sake of simplicity, this principle applies to multi-dimensional variables such as complex-valued vectors and matrices. Following this important principle, we offer in Table 1.1 a set of the conjugate Wirtinger derivative of key functions, which will be leveraged later so as to design gradient-based algorithms. As for the derivative of the composition of two or more functions, the chain rule can be applied, omitting such cases in the table for brevity. The techniques described above will be leveraged mostly in Section 4.

1.3.2 Bayesian Inference

The goal of this section is to offer fundamentals of the Bayesian inference [29–31] and its key techniques so that the proposed methods can be smoothly introduced later. Bayesian inference is a method for parameter estimation based on Bayes' theorem, which is compactly represented as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{1.13}$$

where $\boldsymbol{\theta}$ is a vector representation of a set of parameter(s) of interest with N denoting the number of parameters; $\boldsymbol{y}$ denotes the observation (*e.g.*, received signals in communications); $p(\boldsymbol{y})$ is called the marginal likelihood normalizing the posterior probability; $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$; $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the probability of $\boldsymbol{y}$ for fixed $\boldsymbol{\theta}$; and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior probability that combines the prior probability and the likelihood given

observations.

In case of discrete parameter(s), we instead obtain

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}. \tag{1.14}$$

The denominator of the above equations is introduced to ensure that the posterior probability integrates to one. It can be concluded from the above equations that the posterior is proportional to the likelihood linearly scaled by the prior. With the possession of the posterior expression, the Bayesian estimate of $\boldsymbol{\theta}$ to minimize the mean-square error (MSE) is the mean of the posterior, namely,

$$\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{1.15}$$

which is also equivalent to the maximum a posteriori (MAP) estimate if the mean equals the mode of the posterior, as is the case with the Gaussian distribution.

Note that the integration above becomes the summation in case of discrete parameters. Despite the beauty of the Bayesian parameter estimation, it is often difficult to calculate the exact expression of the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ as well as the posterior in practical signal processing applications, resorting to approximating the latter to a tractable parametric distribution in such cases. This is the core idea of many Bayesian estimation methods such as message passing algorithms (*e.g.*, approximate message passing (AMP) [32] and variants of expectation propagation (EP) [33]) widely utilized for inference problems in signal processing applications. To elaborate, one may consider the posterior to be proportional to the prior times a certain surrogate function, for which different approaches have been considered. Among them, one of the popular choices in the signal processing literature is the Gaussian distribution, whose legitimacy mostly relies on the central limit theorem. In other words, we have

$$p(\boldsymbol{y}|\boldsymbol{\theta}) \approx \frac{\exp\left(-(\boldsymbol{\theta}-\boldsymbol{\mu})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)}{\pi^N|\boldsymbol{\Sigma}|}, \tag{1.16}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is the mean and covariance matrix characterized by the observed information $\boldsymbol{y}$, respectively, and how to model $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on $\boldsymbol{y}$ depends on the considered system model.

Assumig equation (1.16), the posterior can then be approximated as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{1}{C} \underbrace{\exp\left(-(\boldsymbol{\theta}-\boldsymbol{\mu})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right) p(\boldsymbol{\theta})}_{\text{Function of } \boldsymbol{\theta}}, \tag{1.17}$$

with the constant term

$$C \triangleq \int \exp\left(-(\boldsymbol{\theta}-\boldsymbol{\mu})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1.18}$$

One can observe from equations (1.15) and (1.17) that the Gaussian approximation now enables us to obtain the minimum mean squared error (MMSE) estimate of the parameter $\boldsymbol{\theta}$ for any prior distribution $p(\boldsymbol{\theta})$ under the assumption that the Gaussian approximation holds well, although burdensome numerical integration may be needed unless a closed-form expression of the integral in equation (1.15) is available. Fortunately, however, such closed-form expressions indeed exist in many practical scenarios such as channel estimation and symbol detection problems in wireless communications, which have led to a low-complexity estimation paradigm. To illustrate this, let us consider a simple but pragmatic scenario, where the parameter $\boldsymbol{\theta}$ follows the circularly symmetric multivariate complex Gaussian distribution, namely,

$$p(\boldsymbol{\theta}) \propto \frac{\exp\left(-\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}\right)}{\pi^N|\boldsymbol{\Gamma}|}, \tag{1.19}$$

which yields

$$\begin{aligned}
&p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&\approx \frac{\exp\left(-(\boldsymbol{\theta}-\boldsymbol{\mu})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)}{\pi^N|\boldsymbol{\Sigma}|} \cdot \frac{\exp\left(-\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}\right)}{\pi^N|\boldsymbol{\Gamma}|} \\
&= \frac{\exp\left(-\boldsymbol{\mu}^{\mathrm{H}}\left(\boldsymbol{\Sigma}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\mu}\right)}{\pi^N|\boldsymbol{\Gamma}+\boldsymbol{\Sigma}|} \\
&\quad \times \underbrace{\frac{\exp\left(-\left(\boldsymbol{\theta}-\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^{\mathrm{H}}\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)\left(\boldsymbol{\theta}-\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)}{\pi^N|\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)^{-1}|}}_{=\mathcal{CN}\left(\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\right)},
\end{aligned} \tag{1.20}$$

and therefore

$$C = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{\exp\left(-\boldsymbol{\mu}^{\mathrm{H}}\left(\boldsymbol{\Sigma}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\mu}\right)}{\pi^N|\boldsymbol{\Gamma}+\boldsymbol{\Sigma}|}. \tag{1.21}$$

Given the above equations, the MMSE estimate of $\boldsymbol{\theta}$ following the multivariate complex Gaussian distribution under the Gaussian approximation can be written as

$$\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} d\boldsymbol{\theta} = \left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Gamma}\left(\boldsymbol{\Sigma}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\mu}, \tag{1.22}$$

where we utilized the identity $\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}=\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}$ by the Woodbury inverse lemma.

It is worth-noting that instead of the Gaussian model, one can leverage the Gaussian mixture model in conjunction with the expectation-maximization (EM) technique as shown in [34], which brings more flexibility to deal with a wider range of inference problems. Having said that, as shown in the above example, the Gaussian approximation often facilitates deriving a closed-form estimate of parameter(s) with different prior distributions (*e.g.*, [35,36]), which leads to a foundation of low-complexity Bayesian parameter estimation frameworks including ones proposed later in this thesis. Also, part of the derivation given above will be revisited when designing the proposed Bayesian inference algorithms in Section 3.

Chapter 2

Discreteness-Aware Regularizer with Application to Symbol Detection in Massive Non-Orthogonal Systems

In this chapter, we present a novel discreteness-aware regularizer for massive overloaded wireless systems subject to a prescribed discrete modulation set as is the case with the conventional digital wireless communication systems, aiming to improve the detection performance at the receiver. The key concept of the discreteness-aware regularization approach is to explicitly incorporate the prior information of discretely modulated symbols into the conventional least squares (LS) method by means of CS-inspired techniques. Despite this simple modification, the proposed method is shown to be effective not only in underloaded scenarios, where the available resource dimension at the receiver exceeds that of the transmitter, but also in overloaded cases. We further demonstrate wide applicability of this discreteness-aware regularization approach in other signal processing domains by showing its efficacy in the low-rank matrix completion (LRMC) problem with discrete entries, which often appears in recommendation systems.

Part of this chapter is reprinted and enhanced from the following publication:

- Hiroki Iimori, Giuseppe Thadeu Freitas de Abreu, Takanori Hara, Koji Ishibashi, Razvan-Andrei Stoica, David González G, and Osvaldo Gonsa: "Robust Symbol Detection in Large-Scale Overloaded NOMA Systems," *IEEE Open J. Commun. Society*, vol. 2 pp. 512–533, Mar. 2021.
- Hiroki Iimori, Giuseppe Thadeu Freitas de Abreu, Omid Taghizadeh, and Koji Ishibashi: "Discrete-Aware Matrix Completion via Proximal Gradient," *Proc. Asilomar Conference on Signals, Systems, and Computers*, Nov. 2020.

2.1 Background and Contributions

Due to the continuing growth in the number of users and network traffic, future wireless systems will need to employ non-orthogonal transmission strategies in order to cope with the unavoidable shortage of spectral resources [37, 38]. This foreseeable future panorama may require receivers capable of handling underdetermined (overloaded) conditions in which the dimension of transmit signals is significantly larger than that of observed (received) signals [39–46][1].

The design of such (possibly massively) overloaded receivers must therefore differ fundamentally from the conventional and well-known linear ZF and LMMSE detectors, which exhibit high error floors in overloaded scenarios. In particular, unlike the classical ZF and LMMSE approaches, receivers for massively overloaded signaling must not only enforce minimal distance between reconstructed overlapped signals over the continuous multidimensional space, but also maximize the likelihood that the reconstruction satisfies the constraints imposed by the actual discrete transmit constellation(s).

An indirect mechanism to enforce such adherence to discrete constellation is parallel interference cancellation (PIC), in the sense that in PIC receivers the most-likely constellation-bound interfering signal combinations are removed from the observed signals towards detection. Several PIC receiver designs exploiting the sphere detection method have been proposed in the past [39, 44], which illustrate the feasibility of asymptotically approaching the optimal ML detection performance in overloaded systems at somewhat controlled computational complexity.

Despite the progress attained by contributions such as those mentioned above, sphere detection algorithms are known not to scale well, so that a new approach for the design of receivers for massively overloaded systems typical of ultra-dense scenarios is still a major challenge to be conquered. Aiming at addressing this challenge, lower complexity signal detectors based on a novel finite-alphabet signal regularization technique introduced in [48] have been recently proposed for non-orthogonal systems [40, 45]. In [40], for instance, an overloaded signal detector based on the Douglas-Rachford algorithm was proposed for large overloaded MIMO systems, which was shown to yield a significant bit-error rate (BER) gain over the conventional LMMSE. That technique, referred to as sum of absolute value (SOAV), was later generalized into the sum of complex sparse regularizers (SCSR) method proposed in [45], in which the alternating direction method of multipliers (ADMM) algorithm is leveraged in order to enable the detector to deal with complex-valued discrete signals.

Although the SOAV and SCSR detectors are steps in the right direction, as indi-

[1] Not to mention, this situation may include not only point-to-point communications but also multi access schemes such as uplink non-orthogonal access systems [47].

cated by the fact that both were shown to outperform previous state-of-the-art schemes including the graph-based iterative Gaussian detector (GIGD) [49], the Quad-min [50] and the enhanced reactive tabu search (ERTS) [51], both in terms of detection error and computational complexity, in those methods the ℓ_0-norm regularization function employed to capture the discreteness of input signals is replaced by an ℓ_1-norm approximation, leading to inefficiencies that can be mitigated. It is also worth-mentioning that the authors in [48,52] proposed yet another transform-based soft quantization approach, referred to as the simplicity-based recovery (SBR), to address the discreteness of signal reconstruction problems.

In light of this background, we introduce here an alternative mechanism to effectively tackle the non-convex ℓ_0-norm-based formulation of the optimal brute-force search without resorting to an ℓ_1-norm approximation, so as to yield efficient and high-performing receivers for overloaded MIMO systems, which had yet to be presented.

Contributions

In light of the above, the contributions offered in this section are listed as follows:

- A novel tightly-convexified discreteness-aware regularizer for the detection problem of discrete signals is proposed, in which the alternative ML problem via the ℓ_0-norm is tightly approximated by an asymptotically-exact expression. A recently-proposed FP technique [21] is then utilized to further transform the sum of fractions into a tractable quadratic problem, resulting in a new generalization of the conventional ZF detection with adherence to the predetermined discrete constellation set.

- Taking advantage of this new formulation of discrete symbol detetion problems, a novel iterative discrete least square detection framework for the massive non-orthogonal systems is offered, which consists of a simple iteration of a closed-form linear expression closely resembling that of the classic ZF receiver. This new detector, referred to as the iterative discrete least square (IDLS) receiver, is shown to outperform the classic MMSE as well as the recently-proposed SOAV and SCSR state-of-the-art (SotA) schemes, without need for channel statistics, unlike belief propagation methods.

- Furthermore, unlike the related literature, in which the parameterization of regularized optimization problems is separately carried out (*e.g.*, via an exhausive search [53] or machine learning aided approach [54]), a new mechanism to systematically find a sub-optimal regularization parameter of the regularized detection problem is proposed. Based on this formulation, the proposed IDLS detector is

described and shown to be given by the largest generalized eigenvalue of a matrix pencil constructed only with knowledge of the received signal, the channel estimate, and noise variance.

- Several variations of the IDLS receiver for overloaded NOMA systems are also illustrated, which enable the mitigation of impairment factors such as CSI imperfection and hardware impairments. These impairments are seemlessly integrated with the former IDLS detector by generalizing the closed-form expression at the core of the IDLS detector, leading to a Robust IDLS scheme.
- As an additonal illustration of wide applicability of the discreteness-aware regularization approaches in different signal processing domains, we also present a novel algorithm for the low-rank matrix completion problem whose entries are limited to a finite discrete alphabet set such as those in recommender systems.

2.2 System Model and Problem Formulation

Consider a possibly overloaded wireless communication system with N_t transmit and N_r receive wireless resources (*i.e.* antennas, subcarriers or time slots), such that the overloading ratio of the system can be defined as $\gamma \triangleq N_t/N_r$. Assuming perfect channel knowledge at the receiver, the received signal can then be modeled as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n},\ \boldsymbol{y} \in \mathbb{C}^{N_r \times 1}, \tag{2.1}$$

where transmit symbols are normalized to a unit average power per symbol, *i.e.* $\mathbb{E}\big[\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\big] = \mathbf{I}_{N_t}$, and such that each element of $\boldsymbol{s}$ is sampled from the same discrete and regular[2] quadrature amplitude modulation (QAM) constellation set $\mathcal{C} = \{c_1, \cdots, c_{2^b}\}$ of cardinality 2^b, where b denotes the number of bits per symbol; while $\boldsymbol{n} \in \mathbb{C}^{N_r \times 1}$ is an independent and identically distributed (i.i.d.) circularly symmetric complex additive white Gaussian noise (AWGN) vector with zero mean and covariance matrix $\sigma_{\boldsymbol{n}}^2 \mathbf{I}_{N_r}$, and $\boldsymbol{H} \in \mathbb{C}^{N_r \times N_t}$ describes the flat fading channel matrix between transmitter and receiver.

It will prove convenient hereafter to also express the complex-valued quantities in equation (2.1) in terms of their real and imaginary parts, by defining

$$\begin{aligned} \mathbf{y} &\triangleq \begin{bmatrix} \Re\{\boldsymbol{y}\} \\ \Im\{\boldsymbol{y}\} \end{bmatrix}, \qquad \mathbf{H} \triangleq \begin{bmatrix} \Re\{\boldsymbol{H}\} & -\Im\{\boldsymbol{H}\} \\ \Im\{\boldsymbol{H}\} & \Re\{\boldsymbol{H}\} \end{bmatrix}, \\ \mathbf{s} &\triangleq \begin{bmatrix} \Re\{\boldsymbol{s}\} \\ \Im\{\boldsymbol{s}\} \end{bmatrix}, \qquad \mathbf{n} \triangleq \begin{bmatrix} \Re\{\boldsymbol{n}\} \\ \Im\{\boldsymbol{n}\} \end{bmatrix}, \end{aligned} \tag{2.2}$$

[2]By *regularity*, it is meant that the sets $\Re\{\mathcal{C}\}$ and $\Im\{\mathcal{C}\}$ are identical. The assumption is without loss of generality and adopted to simplify the exposition in alignment with the constellation sets used in practical 5G systems.

such that we may write

$$\mathbf{y} = \mathbf{H}\,\mathbf{s} + \mathbf{n},\ \mathbf{y} \in \mathbb{R}^{2N_r \times 1}. \tag{2.3}$$

2.2.1 Problem Formulation

Given the above, the brute-force detection of the complex transmit signal vector $\boldsymbol{s}$ in equation (2.1) can be expressed as the following constellation-constrained ℓ_2-norm minimization problem

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 \tag{2.4a}$$

$$\text{subject to} \quad \boldsymbol{s} \in \mathcal{C}^{N_t}, \tag{2.4b}$$

or equivalently

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 \tag{2.5a}$$

$$\text{subject to} \quad \mathbf{s} \in \mathcal{P}^{2N_t}, \tag{2.5b}$$

where $\mathcal{P} \triangleq \Re\{\mathcal{C}\} = \{p_1, \cdots, p_{2^{b/2}}\}$ is a pulse amplitude modulation (PAM) constellation consisting of the $|\mathcal{P}| = 2^{b/2}$ real/imaginary parts of the symbols in $\mathcal{C}$,

It is evident that due to the disjoint constraints (2.4b) and (2.5b) the optimization problems formulated as in equations (2.4) and (2.5) are non-convex, such that their exact solution require exhaustive searches among all possible combinations of the elements of $\mathcal{C}$ and $\mathcal{P}$, respectively, resulting in a prohibitive complexity of order 2^{bN_t}.

In what follows, a continuous-space reformulation of the latter problem is obtained, which allows for convexification methods to be applied, enabling the posterior design of efficient algorithms to solve the problem at much lower complexities. To this end, we seek inspiration in the approach proposed in [48, Prop.1] and replace the constraint (2.4b) with an equivalent ℓ_0-norm expression, such that equations (2.4) and (2.5) can be respectively rewritten as

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 \tag{2.6a}$$

$$\text{subject to} \quad \sum_{i=1}^{2^b} \|\boldsymbol{s} - c_i\mathbf{1}\|_0 = N_t \cdot (2^b - 1), \tag{2.6b}$$

and

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 \tag{2.7a}$$

$$\text{subject to} \quad \sum_{i=1}^{2^{b/2}} \|\mathbf{s} - p_i\mathbf{1}\|_0 = 2N_t \cdot (2^{\frac{b}{2}} - 1). \tag{2.7b}$$

We remark that unlike constraint (2.5b), the constraint in equation (2.7b) is a continuous function of the symbol vector $\mathbf{s}$. Furthermore, it is clear that in order for a vector $\mathbf{s}$ to satisfy the equality in (2.7b), each and all of its entries must be elements of the constellation.

In other words, no relaxation penalty results from the substitution of the disjoint constraint (2.5b) by the continuous constraint (2.7b), such that an exact solution of equation (2.7) is still an ML solution of equation (2.3). As a consequence of the above, further reformulations of the problem described by equation (2.7) obtained by convex relaxations of constraint (2.7b) retain the potential to yield performance close to that of the ML solution, so long as the corresponding alternative to (2.7b) is sufficiently tight. Obviously, equivalent statements can be made for equation (2.6) with respect to constraint (2.6b), such that the ML signal detectors of equations (2.6) and (2.7) can be respectively modified into the following penalized mixed ℓ_0-ℓ_2 minimization problems

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}} \ \lambda\sum_{i=1}^{2^b}\|\boldsymbol{s}-c_i\mathbf{1}\|_0+\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2, \tag{2.8a}$$

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \ \lambda\sum_{i=1}^{2^{b/2}}\|\mathbf{s}-p_i\mathbf{1}\|_0+\|\mathbf{y}-\mathbf{H}\mathbf{s}\|_2^2, \tag{2.8b}$$

where $\lambda \geq 0$ is a weighting parameter to be determined later.

2.2.2 Comparative Discussion of Recent SotA Approaches

The latter formulations elucidate that the SCSR scheme of [45] and the SOAV MIMO detector proposed in [40] are convexified alternatives to equations (2.8a) and (2.8b), respectively, with the rather classical replacement of the ℓ_0-norm by its convex hull ℓ_1-norm. To elaborate, the SCSR and SOAV machineries essentially aim at addressing the following mixed-norm convex optimization problems, respectively, [40, 45]:

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}} \ \lambda\sum_{i=1}^{2^b}\|\boldsymbol{s}-c_i\mathbf{1}\|_1+\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2, \tag{2.9a}$$

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \ \lambda\sum_{i=1}^{2^{b/2}}\|\mathbf{s}-p_i\mathbf{1}\|_1+\|\mathbf{y}-\mathbf{H}\mathbf{s}\|_2^2, \tag{2.9b}$$

from which one may notice that both methods result in the same recovery performance provided that the balancing parameter λ is properly chosen for each distinct method and a symmetric QAM modulation is selected.

Therefore, the aforementioned SCSR and SOAV formulation reveals the following findings. Firstly, the replacement of the ℓ_0-norm originally appearing in equations

(2.6) to (2.8) by the ℓ_1-norm employed in equation (2.9) can be further improved by tightening the relaxation gap [55].

Secondly, the penalization parameter λ introduced in the reformulation of equations (2.6) and (2.7) into equations (2.8a) and (2.8b) is left to be conquered. And thirdly, various issues of practical relevance such as the influence of noise, imperfect CSI, and hardware imperfection are left unaddressed.

In what follows, we present a new framework to address the intractable ℓ_0-norm in equation (2.8), which is based on an asymptotically-tight approximation of the latter in combination with FP techniques for convexification. This leads to a powerful framework that is subsequently extended later to incorporate robustness to practical impairment factors such as CSI imperfection and hardware impairments. Better still, the resulting framework yields receivers whose algorithms resume to iterations of simple closed-form expressions, which greatly resemble, and thus in fact generalize, the classic ZF and LMMSE detectors.

2.3 Proposed IDLS Framework

We seek to improve over the state of the art on large-scale overloaded multidimensional signal detection schemes. To this end, in this section we propose a new framework for the symbol detection in overloaded NOMA systems.

2.3.1 Fundamentals and Reformulation

Given the equivalence between the complex- and real-valued formulations of equations (2.8a) and (2.8b), we shall, for simplicity of exposition, focus hereafter on the complex-valued variation of equation (2.8a), without loss of generality. Our objective is therefore to obtain a tight relaxation of the ML-derived detection problem described by equation (2.8a), which, unlike that of equation (2.9a), does *not* resort to the convex ℓ_1-norm, while still circumventing the non-convexity of the ℓ_0-norm in a manner that allows for a low-complexity solution in the form of iterations of a simple closed-form expression.

To that end, consider the following asymptotically tight and smooth approximation of the ℓ_0-norm

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{L} \frac{|x_i|^2}{|x_i|^2+\alpha} = L - \sum_{i=1}^{L} \frac{\alpha}{|x_i|^2+\alpha}, \tag{2.10}$$

where $\mathbf{x}$ denotes an arbitrary sparse vector of length L, with $0 < \alpha \ll 1$, such that for $\alpha \to 0$ the approximation becomes exact, as illustrated in Figure 2.1.

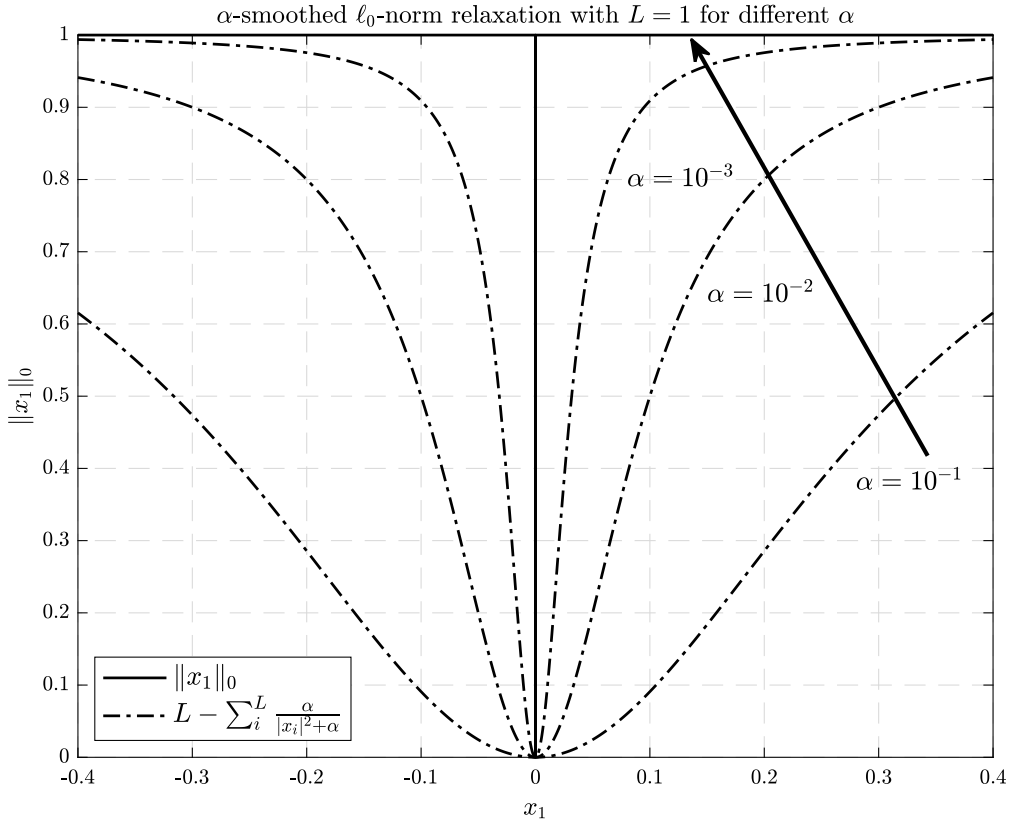


Figure 2.1: Illustration of the ℓ_0-norm approximation via equation (2.10) for different values of α with $L = 1$ for the sake of visualization. It is visible that the smooth approximation asymptotically approaches the ℓ_0-norm as $\alpha \to 0$.

Substituting equation (2.10) into equation (2.8a) yields

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}} \; -\lambda \sum_{i=1}^{2^b} \sum_{j=1}^{N_t} \frac{\alpha}{|s_j - c_i|^2 + \alpha} + \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2. \tag{2.11}$$

Notice that the objective in equation (2.11) is, unlike that of equation (2.8a), a smooth and differentiable function which, despite non-convexity with respect to $\boldsymbol{s}$, is characterized by a sum of concave-over-convex ratios (SCCR). An effective technique to address this non-convexity, referred to as the QT, has been recently proposed in [21]. As the QT has been described in Section 1, we omit to provide technical background of the latter to avoid redundancy,

Applying the QT to equation (2.10) yields

$$\|\mathbf{x}\|_0 \approx L - \Big(\sum_{i=1}^{L} 2\beta_i \sqrt{\alpha} - \beta_i^2 (|x_i|^2 + \alpha) \Big) \tag{2.12a}$$

$$= \sum_{i=1}^{L} \beta_i^2 |x_i|^2 + \underbrace{L - \Big(\sum_{i=1}^{L} 2\beta_i \sqrt{\alpha} + \alpha \Big)}_{\text{independent of } \mathbf{x}}, \tag{2.12b}$$

where we remark that the latter terms independent of $\mathbf{x}$ in equation (2.12b) can be discarded in the context of a minimization problem on the variable $\mathbf{x}$.

Substituting equation (2.12), with the constant terms discarded, into equation (2.11) yields

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}}\ \lambda\sum_{i=1}^{2^b}\sum_{j=1}^{N_t}\beta_{i,j}^2|s_j-c_i|^2+\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2, \tag{2.13}$$

with

$$\beta_{i,j}=\frac{\sqrt{\alpha}}{|s_j-c_i|^2+\alpha},\ \forall\, i\in\{1,\cdots,2^b\}, j\in\{1,\cdots,N_t\}. \tag{2.14}$$

Next, define the quantities

$$\boldsymbol{b}\triangleq\sum_{i=1}^{2^b}c_i\left[\beta_{i,1}^2,\beta_{i,2}^2,\ldots,\beta_{i,N_t}^2\right]^{\mathrm{T}}, \tag{2.15a}$$

$$\boldsymbol{B}\triangleq\sum_{i=1}^{2^b}\mathrm{diag}(\beta_{i,1}^2,\beta_{i,2}^2,\ldots,\beta_{i,N_t}^2)\succ 0, \tag{2.15b}$$

such that equation (2.13) can be written as

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}}\ \lambda(\boldsymbol{s}^{\mathrm{H}}\boldsymbol{B}\boldsymbol{s}-2\Re\{\boldsymbol{b}^{\mathrm{H}}\boldsymbol{s}\})+\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2, \tag{2.16}$$

which again can be expressed more compactly by expanding the latter quadratic term, namely

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}}\ \boldsymbol{s}^{\mathrm{H}}(\lambda\boldsymbol{B}+\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})\boldsymbol{s}-2\Re\{(\lambda\boldsymbol{b}^{\mathrm{H}}+\boldsymbol{y}^{\mathrm{H}}\boldsymbol{H})\boldsymbol{s}\}. \tag{2.17}$$

Remark 2.3.1. *A key aspect of this new formulation is that the ℓ_0-norm term in equation* (2.8a) *is approximated with arbitrary tightness by choosing α sufficiently small, which therefore differs fundamentally from SotA methods.*

2.3.2 Iterative Discrete Least Square Solution

For future convenience, let us define the function

$$q(\boldsymbol{s})\triangleq\boldsymbol{s}^{\mathrm{H}}(\lambda\boldsymbol{B}+\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})\boldsymbol{s}-2\Re\{(\lambda\boldsymbol{b}^{\mathrm{H}}+\boldsymbol{y}^{\mathrm{H}}\boldsymbol{H})\boldsymbol{s}\}, \tag{2.18}$$

which is in fact the objective function in the optimization problem described by equation (2.17).

Given that $q(\boldsymbol{s})$ is quadratic on $\boldsymbol{s}$, the latter problem can be readily solved in closed-form by setting its Wirtinger derivative [56] with respect to $\boldsymbol{s}$ equal to 0, that is,

$$\frac{\partial q(\boldsymbol{s})}{\partial\boldsymbol{s}^*}=(\lambda\boldsymbol{B}+\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})\boldsymbol{s}-\left(\lambda\boldsymbol{b}+\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}\right)=0, \tag{2.19}$$

which readily yields the solution

$$\boldsymbol{s} = \big(\lambda \boldsymbol{B} + \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\big)^{-1}\big(\lambda \boldsymbol{b} + \boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}\big). \tag{2.20a}$$

Obviously, a real-domain equivalent of equation (2.20) – which will prove convenient later – can also be obtained following the same steps as above, yielding

$$\mathbf{s} = \big(\lambda \mathbf{B} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\big)^{-1}\big(\lambda \mathbf{b} + \mathbf{H}^{\mathrm{T}}\mathbf{y}\big), \tag{2.20b}$$

where we remind that $\mathbf{y}$, $\mathbf{H}$ and $\mathbf{s}$ are as previously defined in equation (2.2), and the majorization pivot quantities $\beta_{i,j}$, $\mathbf{b}$ and $\mathbf{B}$ are respectively redefined as

$$\beta_{i,j} \triangleq \frac{\sqrt{\alpha}}{(s_j - p_i)^2 + \alpha} \text{ with } \begin{cases} i \in \{1, \cdots, 2^{b/2}\} \\ j \in \{1, \cdots, 2N_t\} \end{cases}, \tag{2.21}$$

and

$$\mathbf{b} \triangleq \sum_{i=1}^{2^{\frac{b}{2}}} p_i \left[\beta_{i,1}^2, \beta_{i,2}^2, \ldots, \beta_{i,2N_t}^2\right]^{\mathrm{T}}, \tag{2.22a}$$

$$\mathbf{B} \triangleq \sum_{i=1}^{2^{\frac{b}{2}}} \mathrm{diag}(\beta_{i,1}^2, \beta_{i,2}^2, \ldots, \beta_{i,2N_t}^2) \succ 0, \tag{2.22b}$$

with $p_i \in \mathcal{P} \triangleq \Re\{\mathcal{C}\}$.

We emphasize the remarkably simple structure of the receiver described by equation (2.20), which is characterized by the mere iteration – over which $\boldsymbol{b}$, $\boldsymbol{B}$, and consequently $\boldsymbol{s}$ are updated – of an expression that is linear on the input $\boldsymbol{y}$, and which relies on the inversion[3] of a matrix guaranteed to be invertible due to the positive definiteness of the term $\lambda\boldsymbol{B}$.

Remark 2.3.2. *Notice that equation* (2.20) *is akin to the conventional linear ZF receiver, except for the penalization factor* λ *and the dependence on the iteratively-computed regularization terms* $\boldsymbol{b}$ *and* $\boldsymbol{B}$*, which together enforce the constellation compliance of the solution. In particular, with* $\lambda = 0$*, equation* (2.20) *reduces to a conventional ZF receiver, with* $(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{H}}$ *yielding the pseudo-inverse of the channel* $\boldsymbol{H}$*. In other words,*

$$\lim_{\lambda \to 0^+} \big(\lambda \boldsymbol{B} + \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\big)^{-1}\big(\lambda \boldsymbol{b} + \boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}\big) = (\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}. \tag{2.23}$$

Furthermore, the solution given in equation (2.20) *converges to a stationary point of the regularized ML problem relaxed by equation* (2.10) *due to [21, Th. 1].*

[3]Since the matrix $\lambda\boldsymbol{B} + \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}$ is only updated by the term $\lambda\boldsymbol{B}$ at each iteration, its inverse can be accelerated employing tracking methods, *e.g.* [57].

Unlike the conventional ZF receiver, however, which is known to perform poorly if the channel matrix $\boldsymbol{H}$ is rank deficient – as is the case of overloaded systems or spatially-correlated channels – the iterative discrete least square linear detector here derived and summarized by equation (2.20) is the solution of the optimization problem described by equation (2.17), which in turn was obtained by a tight relaxation of the brute-force problem of equation (2.8a). It is thus expected and confirmed via computer simulations that the iterative linear detector here presented delivers better performance with improved robustness to overloading. Finally, since $(\lambda\boldsymbol{B}+\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}) \succ 0$, the inversion may be performed.

All in all, it can be said that the receiver given by equation (2.20) is a generalization of the classical LS receiver, which adheres to the discrete constellation in a continuous, asymptotically exact and converging manner. Alluding to this fact, we refer to our scheme as the IDLS framework.

2.3.3 Auto-Parameterization via Generalized Eigen Decomposition

Regularized optimization methods have grown in popularity in recent years [53] due to their ability to solve complex problems at relatively low computational cost. In these methods, penalized terms are added to an objective function in order to balance multiple desired features in the solution. Well-known examples are CS methods [55] and the classic least absolute shrinkage and selection operator (LASSO) estimator, in which sparse solutions are desired and ensured by the ℓ_0- or the ℓ_1-norm, and matrix completion methods [58] in which low rank solutions are desired and ensured by the nuclear norm.

An important reason for the performance and robustness of such methods is that the choice of regularization parameter is not too sensitive, such that it can be dealt with offline [59] via dedicated techniques. To cite a couple of examples, an approach to optimize regularization parameters based on deep unfolding was proposed in [54], and in [60] another method is proposed based on duality theory.

Despite the existence of these and several other techniques to optimize regularization parameters, for the sake of completeness we contribute also to this issue by proposing a new algorithm to optimize the penalization parameter λ employed in the IDLS detection framework presented above. To this end, first the real-valued equivalent of the fundamental ML-derived optimization problem described in equation (2.16), from which our framework has been obtained, which is given by

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \ \mathbf{s}^{\mathrm{T}}\mathbf{B}\mathbf{s} - 2\mathbf{b}^{\mathrm{T}}\mathbf{s} + \frac{1}{\lambda}\|\mathbf{y}-\mathbf{H}\mathbf{s}\|_2^2, \tag{2.24}$$

where the penalization parameter has been moved to the ℓ_2-norm term of the objective (only for future convenience and without prejudice to the formulation itself).

Now, notice that equation (2.24) is merely a regularized mixed norm variation of the original ML problem described by equation (2.7), and observe that in order to satisfy the equality constraint of that formulation, the term $\sum_{i=1}^{2^{b/2}} \|\mathbf{s} - x_i\mathbf{1}\|_0$ in equation (2.7b) – which corresponds to the term $\mathbf{s}^{\mathrm{T}}\mathbf{B}\mathbf{s} - 2\mathbf{b}^{\mathrm{T}}\mathbf{s}$ in equation (2.24) – needs to be globally minimized.

It follows from the above that the regularized ℓ_2-norm term in equation (2.24) can be placed as a constraint, with no penalty to the optimality of the formulation. In other words, the ML-derived formulation of equation (2.24) – and by extension to the original ML formulation of equation (2.7) – are equivalent to the following real-valued quadratically constrained quadratic program with one convex constraint (QCQP-1) formulation

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \quad \mathbf{s}^{\mathrm{T}}\mathbf{B}\,\mathbf{s} - 2\,\mathbf{b}^{\mathrm{T}}\mathbf{s} \tag{2.25a}$$

$$\text{subject to} \quad \mathbf{s}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\,\mathbf{s} - 2\,\mathbf{y}^{\mathrm{T}}\mathbf{H}\,\mathbf{s} + \mathbf{y}^{\mathrm{T}}\mathbf{y} - \delta \leq 0, \tag{2.25b}$$

where the constraint (2.25b) obeys Slater's condition.

We remark that the bounding parameter δ (a.k.a. "search ball radius") establishes the tightness within which the squared distance $\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2$ is made to adhere, and is typically determined by the noise power [61], since noise variance is standardly available in practical systems [62].

With that in mind, for the sake of simplicity, the quadratic function in (2.25b) is defined as

$$k(\mathbf{s}) \triangleq \mathbf{s}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\,\mathbf{s} - 2\,\mathbf{y}^{\mathrm{T}}\mathbf{H}\,\mathbf{s} + \mathbf{y}^{\mathrm{T}}\mathbf{y} - \delta. \tag{2.26}$$

All that is left for us to do then is to obtain an efficient method to solve equation (2.25), which among many alternatives can be achieved by applying the result presented in [63, Th.3.3]. Brought to the context hereby, that result states that if there exists a minimizer $\bar{\mathbf{s}}$ of equation (2.25a) satisfying the constraint (2.25b), then $\bar{\mathbf{s}}$ is the global solution to equation (2.25) if and only if (iff) there exists a parameter $\mu^{\mathrm{opt}} \geq 0$ such that the following Karush Kuhn Tucker (KKT) conditions are satisfied

$$(\mathbf{B} + \mu^{\mathrm{opt}}\mathbf{H}^{\mathrm{T}}\mathbf{H})\,\bar{\mathbf{s}} = (\mathbf{b} + \mu^{\mathrm{opt}}\mathbf{H}^{\mathrm{T}}\mathbf{y}), \tag{2.27a}$$

$$k(\bar{\mathbf{s}}) \leq 0, \tag{2.27b}$$

$$\mu^{\mathrm{opt}}k(\bar{\mathbf{s}}) = 0. \tag{2.27c}$$

In recognition to the outstanding work presented in [63], we refer to this formulation of the QCQP-1 problem of equation (2.25) as Moré's Theorem, which in fact admits two distinct cases. The first is when $\mu^{\mathrm{opt}} = 0$, in which case equation (2.27a) reduces to equation (2.25a), with unique global minimum at $\bar{\mathbf{s}} = \mathbf{B}^{-1}\mathbf{b}$, which is obviously

a "solution" of no relevance since it is independent of the input. The second and only relevant case is when $\mu^{\text{opt}} > 0$, in which case equations (2.27b) and (2.27c) both coincide and reduce to $k(\bar{\mathbf{s}}) = 0$, such that Moré's Theorem then yields

$$\bar{\mathbf{s}} = (\mathbf{B} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{H})^{-1}(\mathbf{b} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{y}), \tag{2.28a}$$

$$\bar{\mathbf{s}}^{\text{T}}(\mathbf{H}^{\text{T}}\mathbf{H}\,\bar{\mathbf{s}} - \mathbf{H}^{\text{T}}\,\mathbf{y}) - \mathbf{y}^{\text{T}}\mathbf{H}\,\bar{\mathbf{s}} + \mathbf{y}^{\text{T}}\mathbf{y} - \delta = 0, \tag{2.28b}$$

where we have in equation (2.28a) inverted equation (2.27a), and in equation (2.28b) expressed $k(\bar{\mathbf{s}}) = 0$ explicitly, with the term $2\,\mathbf{y}^{\text{T}}\mathbf{H}\,\bar{\mathbf{s}}$ expanded and slightly rearranged.

A comparison of equations (2.20) and (2.28a) reveals that, except for the complex and real domains, respectively, both are identical if $\mu^{\text{opt}} = 1/\lambda$. There is, however, a crucial difference between both equations, namely, that in equation (2.28a), the parameter μ^{opt} results not from a regularization – as indeed the QCQP-1 formulation of equation (2.25) is not regularized – but rather from the KKT conditions required to solve equation (2.25), such μ^{opt} is an integral part of such solution.

In addition, notice that equation (2.28a) is tied to the accompanying equation (2.28b), in the sense that *both* must be simultaneously satisfied in order for the solution to hold. In other words, the solution of the system of equations (2.28) yields within it the optimum KKT parameter μ^{opt}, which in turn determines the optimum regularization parameter $\lambda^{\text{opt}} = 1/\mu^{\text{opt}}$ required in equation (2.20).

To accomplish this task, we first introduce the scaling quantity ρ and the auxiliary vector $\bar{\mathbf{s}} \triangleq \frac{\mathbf{x}_1}{\rho}$, such that equations (2.28) can be rewritten as

$$\mathbf{x}_1^{\text{T}} = \rho\,(\mathbf{b} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{y})^{\text{T}}(\mathbf{B} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{H})^{-1}, \tag{2.29a}$$

$$(\mathbf{y}^{\text{T}}\mathbf{y} - \delta)\rho - \mathbf{y}^{\text{T}}\mathbf{H}\,\mathbf{x}_1 + \frac{1}{\rho}\mathbf{x}_1^{\text{T}}(\mathbf{H}^{\text{T}}\mathbf{H}\,\mathbf{x}_1 - \rho\mathbf{H}^{\text{T}}\mathbf{y}) = 0, \tag{2.29b}$$

where, for future convenience, we exploited the symmetry in $(\mathbf{B} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{H})$ to transpose equation (2.28a) into equation (2.29a), and rearranged equation (2.28b) into equation (2.29b).

Next, using equation (2.29a) in place of $\mathbf{x}_1^{\text{T}}$ in the last term of equation (2.29b), and rearranging the terms yields

$$\underbrace{(\mathbf{y}^{\text{T}}\mathbf{y} - \delta)}_{\triangleq q_{11}}\,\rho\,\underbrace{-\mathbf{y}^{\text{T}}\mathbf{H}}_{\triangleq \mathbf{q}_{12}}\,\mathbf{x}_1 + \underbrace{\mathbf{b}^{\text{T}}}_{\triangleq \mathbf{q}_{13}}\mathbf{x}_2 = \mu^{\text{opt}}\,\underbrace{(-\mathbf{y}^{\text{T}}\mathbf{H})}_{\triangleq \mathbf{p}_{13}}\mathbf{x}_2, \tag{2.30}$$

where again for future convenience we have implicitly defined the quantities q_{11}, $\mathbf{q}_{12}$, $\mathbf{q}_{13}$, $\mathbf{p}_{13}$, and

$$\mathbf{x}_2 \triangleq (\mathbf{B} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{H})^{-1}(\mathbf{H}^{\text{T}}\mathbf{H}\,\mathbf{x}_1 - \rho\mathbf{H}^{\text{T}}\mathbf{y}), \tag{2.31}$$

which in turn can be rearranged as

$$\underbrace{-\mathbf{H}^{\mathrm{T}}\mathbf{y}}_{=\mathbf{q}_{12}^{\mathrm{T}}}\rho+\underbrace{\mathbf{H}^{\mathrm{T}}\mathbf{H}}_{\triangleq\mathbf{q}_{22}}\mathbf{x}_1+\underbrace{(-\mathbf{B})}_{\triangleq\mathbf{q}_{23}}\mathbf{x}_2=\mu^{\mathrm{opt}}\underbrace{\mathbf{H}^{\mathrm{T}}\mathbf{H}}_{\triangleq\mathbf{p}_{23}}\mathbf{x}_2, \tag{2.32}$$

where we have highlighted the reappearance of the previously defined quantity $\mathbf{q}_{12}$ and implicitly defined the further quantities $\mathbf{q}_{22}$, $\mathbf{q}_{13}$ and $\mathbf{p}_{23}$ for future convenience.

Finally, notice that equation (2.29a) can also be rewritten as

$$\underbrace{\mathbf{b}}_{=\mathbf{q}_{13}^{\mathrm{T}}}\rho+\underbrace{(-\mathbf{B})}_{=\mathbf{q}_{23}^{\mathrm{T}}}\mathbf{x}_1=\mu^{\mathrm{opt}}\underbrace{(-\mathbf{H}^{\mathrm{T}}\mathbf{y})}_{=\mathbf{p}_{13}^{\mathrm{T}}}\rho+\mu^{\mathrm{opt}}\underbrace{\mathbf{H}^{\mathrm{T}}\mathbf{H}}_{=\mathbf{p}_{23}^{\mathrm{T}}}\mathbf{x}_1, \tag{2.33}$$

where we once more highlighted the reappearance of the quantities $\mathbf{q}_{13}$, $\mathbf{q}_{23}$, $\mathbf{p}_{13}$ and $\mathbf{p}_{23}$.

Now define the vector $\mathbf{x}\triangleq[\rho,\mathbf{x}_1^{\mathrm{T}},\mathbf{x}_2^{\mathrm{T}}]^{\mathrm{T}}$ and notice that the collection of equations (2.30), (2.32) and (2.33), in that order, can be written compactly as

$$\begin{bmatrix} q_{11} & \mathbf{q}_{12} & \mathbf{q}_{13} \\ \mathbf{q}_{12}^{\mathrm{T}} & \mathbf{q}_{22} & \mathbf{q}_{23} \\ \mathbf{q}_{13}^{\mathrm{T}} & \mathbf{q}_{23}^{\mathrm{T}} & \mathbf{0}_{2N_t} \end{bmatrix}\cdot\begin{bmatrix} \rho \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}=\mu^{\mathrm{opt}}\begin{bmatrix} 0 & \mathbf{0}_{1\times 2N_t} & \mathbf{p}_{13} \\ \mathbf{0}_{2N_t\times 1} & \mathbf{0}_{2N_t} & \mathbf{p}_{23} \\ \mathbf{p}_{13}^{\mathrm{T}} & \mathbf{p}_{23}^{\mathrm{T}} & \mathbf{0}_{2N_t} \end{bmatrix}\cdot\begin{bmatrix} \rho \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{2.34}$$

or simply

$$\mathbf{Q}\,\mathbf{x}=\mu^{\mathrm{opt}}\mathbf{P}\,\mathbf{x}, \tag{2.35}$$

with

$$\mathbf{Q}\triangleq\begin{bmatrix} \mathbf{y}^{\mathrm{T}}\mathbf{y}-\delta & -\mathbf{y}^{\mathrm{T}}\mathbf{H} & \mathbf{b}^{\mathrm{T}} \\ -\mathbf{H}^{\mathrm{T}}\mathbf{y} & \mathbf{H}^{\mathrm{T}}\mathbf{H} & -\mathbf{B} \\ \mathbf{b} & -\mathbf{B} & \mathbf{0}_{2N_t} \end{bmatrix}, \tag{2.36a}$$

$$\mathbf{P}\triangleq\begin{bmatrix} 0 & \mathbf{0}_{1\times 2N_t} & -\mathbf{y}^{\mathrm{T}}\mathbf{H} \\ \mathbf{0}_{2N_t\times 1} & \mathbf{0}_{2N_t} & \mathbf{H}^{\mathrm{T}}\mathbf{H} \\ -\mathbf{H}^{\mathrm{T}}\mathbf{y} & \mathbf{H}^{\mathrm{T}}\mathbf{H} & \mathbf{0}_{2N_t} \end{bmatrix}. \tag{2.36b}$$

One can readily recognize that equation (2.35) defines a generalized eigenvalue problem [64] over the pencil defined by the pair of matrices $(\mathbf{Q},\mathbf{P})$. In other words, the solution of the system of equations in (2.28), and therefore of the QCQP-1 problem described by equation (2.25), is among the generalized eigenvalues of the pencil $(\mathbf{Q},\mathbf{P})$.

Problems described by a quadratic program with a single quadratic constraint, such that the one dealt with here, were studied thoroughly in [65]. It was shown thereby, in particular in [65, Lem.3 and Th.4], that in fact the solution of the QCQP-1 extracted from equation (2.34) is given by its *smallest* generalized eigenpair.

It was also shown thereby, however, that such a solution is also equivalent to the

Algorithm 1 IDLS Detector

Inputs: Received signal $\mathbf{y}$, channel matrix $\mathbf{H}$ and noise power $\sigma_{\mathbf{n}}^2$
Outputs: Estimate $\hat{\boldsymbol{s}}$

1: Set iteration counter $k = 0$.
2: Set initial solution to $\mathbf{s}^{(k)} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}$.
3: **repeat**
4: Increase iteration counter $k = k + 1$.
5: Update $\beta_{i,j} \forall\, i, j$, as in equation (2.21).
6: Construct $\mathbf{b}$ and $\mathbf{B}$ from equations (2.22a) and (2.22b).
7: Construct $\mathbf{Q}$ and $\mathbf{P}$ from equations (2.36a) and (2.36b).
8: Obtain $\lambda^{\mathrm{opt}(k)}$ as in equation (2.38).
9: Update $\mathbf{s}^{(k)}$ as in equation (2.20b).
10: **until** convergence or maximum iterations reached
11: $\hat{\boldsymbol{s}} \leftarrow \mathbf{s}^{(k)}$

largest finite real generalized eigenvalue of the Möbius-transform equivalent of equation (2.34), namely

$$\mathbf{P}\,\mathbf{x} = \lambda^{\mathrm{opt}}\mathbf{Q}\,\mathbf{x}, \tag{2.37}$$

where $\lambda^{\mathrm{opt}} = 1/\mu^{\mathrm{opt}}$.

We remark that the computation of the dominant generalized eigenvalue of a pencil of symmetric matrices can be done accurately, stably and at relatively low complexity, since many classic [64] and modern [66–68] algorithms exist to that end. In fact, since the matrix $\mathbf{P}$ is constant over the iterations of the IDLS detector, while only the last block column and row of the matrix $\mathbf{Q}$ is updated in that process, the recursive adaptation of λ^{opt} from a previous to a next iteration is also possible via Jacobian techniques such as *e.g.* [68].

All in all, we have therefore arrived at the conclusion that the regularization parameter λ^{opt} required to evaluate equation (2.20b) can be computed by the *largest* finite generalized eigenvalue of the pencil $(\mathbf{P}, \mathbf{Q})$, that is

$$\lambda^{\mathrm{opt}} = \mathrm{maxeig}(\mathbf{P}, \mathbf{Q}). \tag{2.38}$$

With that, for the convenience of the reader, we summarize the overall IDLS scheme in the form of a pseudo-code in Algorithm 1 adopting the real-valued variation for mathematical consistency, since the proofs of [63, Th.3.3] and [65, Lem.3 and Th.4] were offered only for real matrices.

2.3.4 Performance Assessment

In this subsection we offer a computer simulation-based performance evaluation of the IDLS scheme described above. In order to focus on the gains achieved by the new method over SotA, we conduct the assessment here under ideal receiver conditions, namely, under the assumption that no distortions due to hardware impairments exists, and that perfect CSI is available at the receiver, leaving those factors to be addressed later in the subsequent section, when our proposed framework will be also revisited to yield IDLS variations robust to such distortions. Aside from these receiver-related limitations, however, system level conditions such as fading, channel correlation and overloading, which typically contribute to the performance deterioration of multiuser symbol detection, are considered. In particular, the following system parameters are considered during the assessments.

- *Uncorrelated Rayleigh Fading Model*: This is the most commonly used channel model to evaluate ideal detection performance in the wireless literature. Each element of the channel matrix $\boldsymbol{H}$ is assumed to be an i.i.d. circularly symmetric complex Gaussian random variable with zero mean and variance 1, *i.e.*, $h_{m,n} \sim \mathcal{CN}(0,1)$ where $h_{m,n}$ denotes the element in the m-th row and n-th column of $\boldsymbol{H}$. This model captures, among others, the channel conditions of the uplink of a cell-free MIMO system serving distributed single-antenna users with either OMA or NOMA protocol [69], in the fully loaded and overloaded cases, respectively.

- *Exponentially-correlated Jakes Fading Model*: In this case, the spatial correlation between antenna elements is captured via the classical Jakes correlation model, in which the block-fading matrix $\boldsymbol{H} \in \mathbb{C}^{N_r \times N_t}$ is described via the following doubly-correlated channel matrix:

$$\boldsymbol{H} = \boldsymbol{\Phi}_r^{\frac{1}{2}} \boldsymbol{H}_{\text{i.i.d.}} \boldsymbol{\Phi}_t^{\frac{1}{2}}, \tag{2.39}$$

  where $\boldsymbol{H}_{\text{i.i.d.}}$ denotes an uncorrelated i.i.d. zero mean, unit variance complex Gaussian matrix, and with the (k,ℓ)-th element of the spatial correlation signatures $\boldsymbol{\Phi}_r$ and $\boldsymbol{\Phi}_t$ given by $J_0\big(\frac{2\pi f_c d|k-\ell|}{c}\big)$, where $f_c = 5$ [GHz] depicts the carrier frequency, d is defined as the corresponding half antenna spacing, c denotes the speed of light, and $J_0(\cdot)$ is the zeroth order Bessel function of the first kind [70].

- *Other System and Algorithmic Parameters*: As for the remaining system configuration, the following is considered. The ℓ_0-norm approximation tightness parameter is set to $\alpha = 0.1$. Various loading conditions are simulated, ranging from fully loaded systems with $N_t = N_r = 100$, to moderately overloaded systems are subjected to either a 25% of excess load, with $N_t/N_r = 1.25$, to severely overloaded systems with a 50% of excess load, with $N_t/N_r = 1.5$. Results are

shown as a function of the energy per bit to noise power spectral density ratio (E_b/N_0), such that the corresponding AWGN noise variance $\sigma_{\boldsymbol{n}}^2$ in a system with N_t transmit antennas is given by $\sigma_{\boldsymbol{n}}^2 = N_t/(b \cdot 10^{\frac{E_b/N_0[\text{dB}]}{10}})$. Finally, it is assumed that the elements of the symbol vector $\boldsymbol{s}$ are sampled with uniform probability from a Gray-coded quadrature phase shift keying (QPSK) modulation, such that $b = 2$.

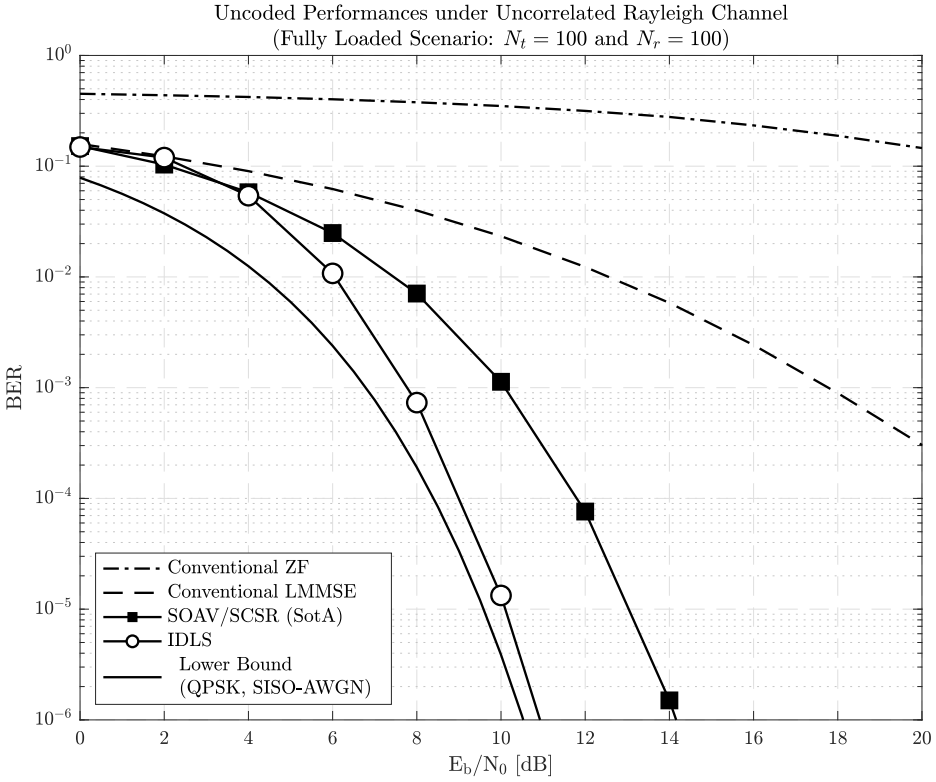
Results

The set of simulated results are displayed in Figure 2.2, 2.3, and 2.4, which shows plots of the BER performance achieved by the IDLS multiuser symbol detection method described in the preceding subsections, compared against those of the conventional LMMSE, low-complexity SOAV and SCSR SotA receivers, as well as the single-input single-output (SISO) AWGN lower bound, under various SNR, loading and channel correlation conditions.

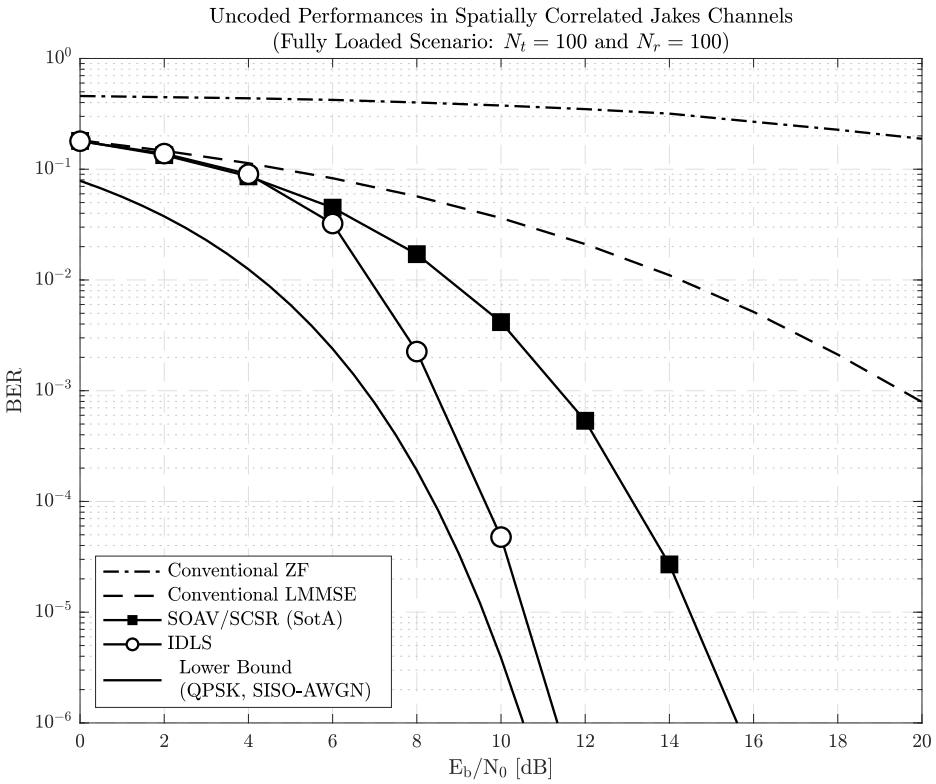
The results altogether clearly demonstrate not only that the new IDLS detector offers substantial gains over the SotA and classic alternatives, but also that the proposed method exhibits a remarkable robustness to overloading and channel correlation conditions, which are known to be a fundamental cause of performance degradation in multiuser systems. In particular, it is visible that in all cases the inclination of the BER curves corresponding to the IDLS method is similar to that achieved by in a SISO with the same spectral efficiency.

We also call attention to the fact that all simulation results displayed are for systems of relatively large scales, which are therefore not tractable to other types of ML-approaching receiver architectures based on techniques such as sphere detection [39]. In other words, the results serve the additional purpose of effectively illustrating that the remarkable performances observed are achieved at a sufficiently low complexity.

In order to further highlight this high-performance-at-low-complexity feature of the IDLS framework, plots depicting its convergence and asymptotic behaviors, as a function of the number of iterations and system size, respectively, are shown in Figure 2.5. The results indicate that less than 25 iterations are sufficient for IDLS to converge in almost all the conditions considered, with 37 iterations sufficing for all the cases. Interestingly (but non-surprisingly), it is also found that the number of iterations required until convergence is not strongly affected by the dimension of the system or loading conditions, as shown in the figure. As for asymptotic behavior, it can be seen from the figure that indeed the BER performance of the IDLS detector enjoys the array or dimension gain as expected. This is due to the fact that owing to the ℓ_0-norm formulation embedded into IDLS, a large system condition facilitates the detection performance of IDLS as is the case with existing compressive sensing algorithms.

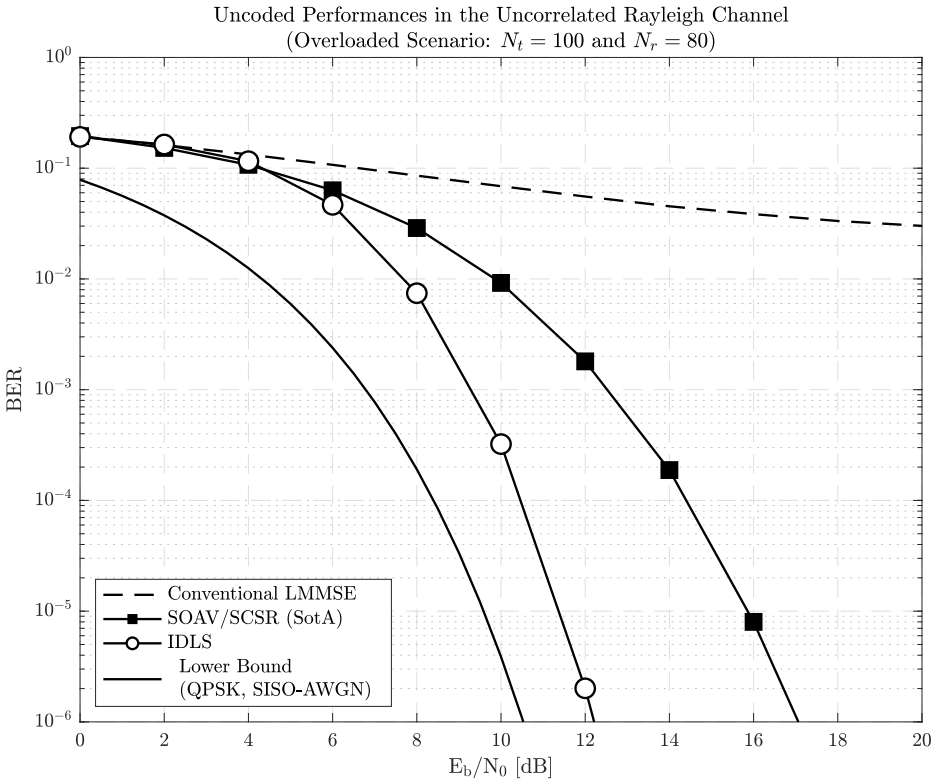


(a) Fully Loaded Uncorrelated.

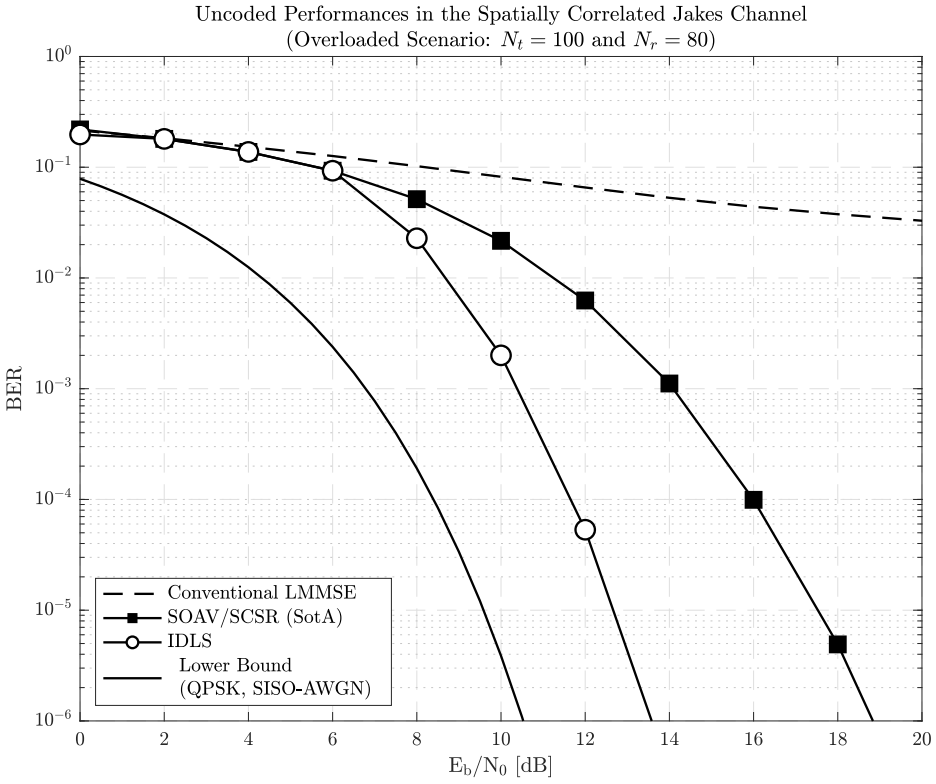


(b) Fully Loaded Jakes Correlated.

Figure 2.2: BER performance with fully-loaded conditions.

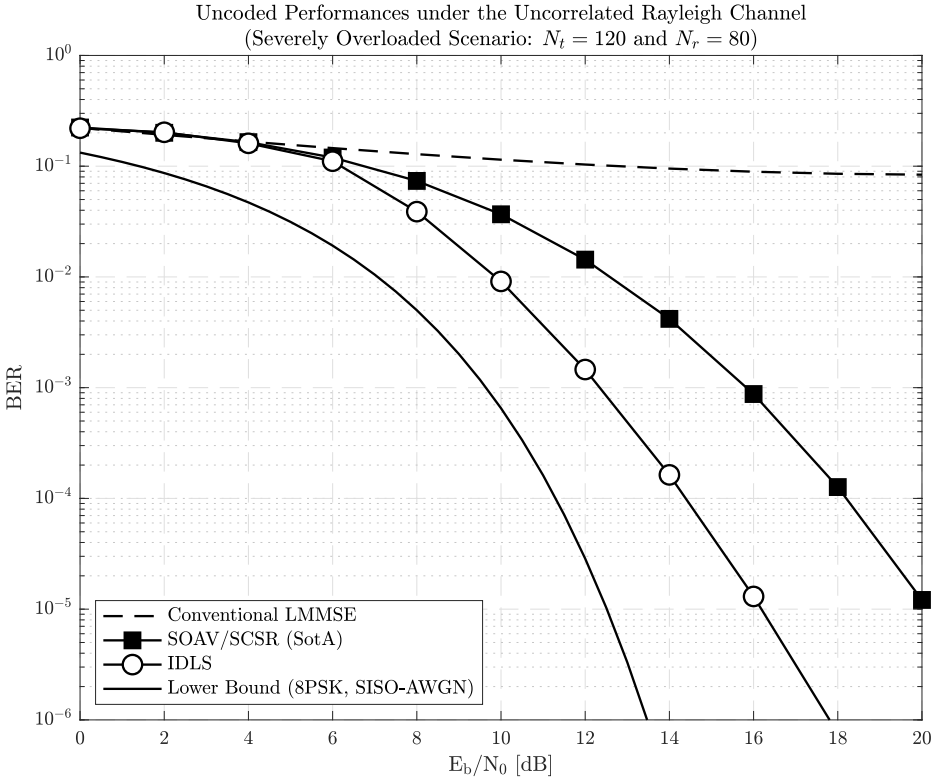


(a) Moderately Overloaded Uncorrelated.

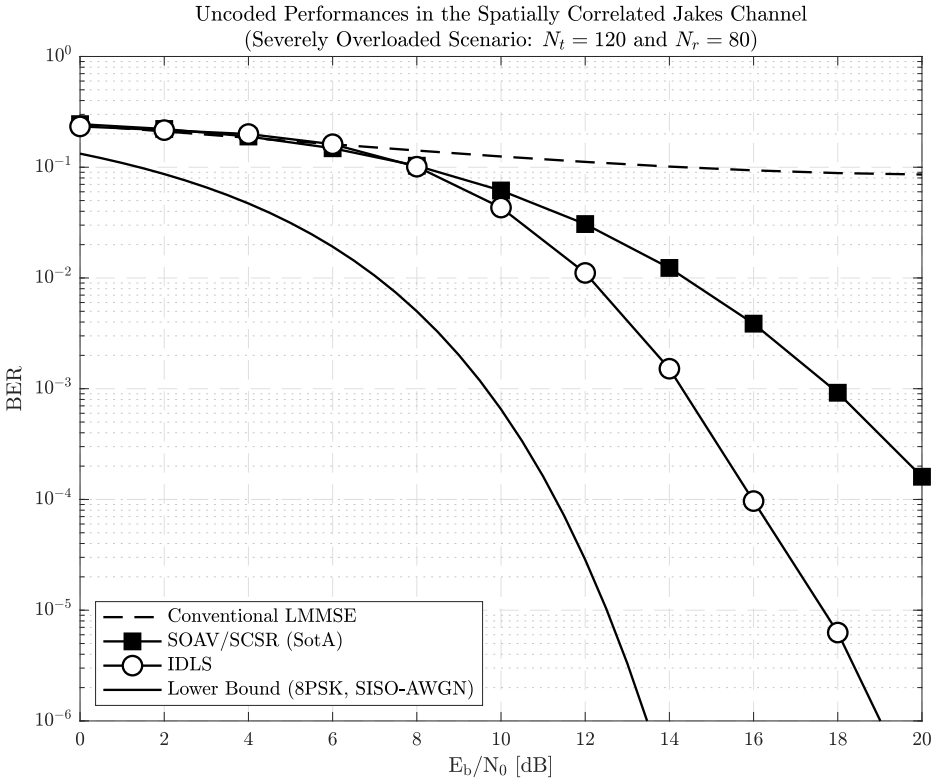


(b) Moderately Overloaded Jakes Correlated.

Figure 2.3: BER performance with moderately overloaded conditions.



(a) Severely Overloaded Uncorrelated.



(b) Severely Overloaded Jakes Correlated.

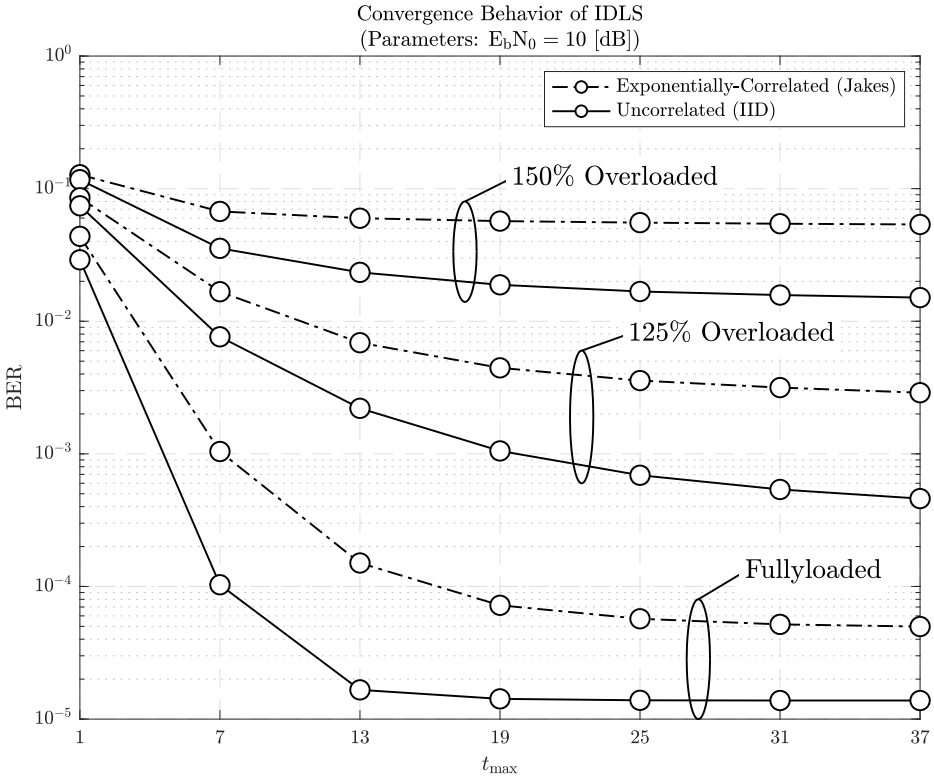
Figure 2.4: BER performance with severely overloaded conditions.

This remarkable resilience of the IDLS detector against inter-symbol interference (ISI) and multiuser interference (MUI), demonstrated when operating over uncoded symbols, which shows no signs of error floors even under severe overloading conditions and subjected to the further rank deficiencies resulting from channel correlation, motivates us to probe the performance of the method when aided by channel coding. To that end, we evaluate in Figure 2.6 the BER performance achieved by the IDLS detector, the conventional LMMSE receiver, and the recently-proposed iterative detection-and-decoding (IDD)-based receiver [46] in a system with overloading ratio $\gamma = 1.5$, this time equipping all receivers with the standardized DVB-S2 LDPC of rate 1/2, decoded with 50 sum-product iterations.

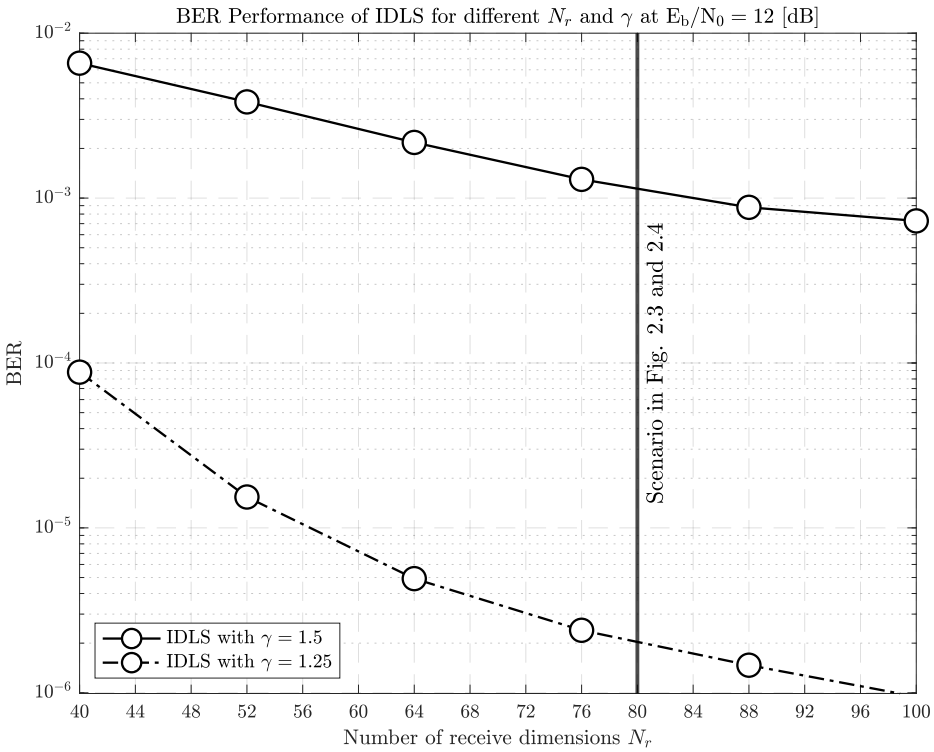
Since the focus of this article is to design a robust detection method for NOMA systems, we show the result of IDLS without the IDD procedure (*i.e.*, no loop between the detector and decoder), because iterations over detector and decoder are not target of our design. In order to illustrate the possible integration of our method with such IDD approaches, however, the BER of IDLS with multiple IDD iterations is calculated by taking advantage of outputs of IDLS as the initializer to the IDD approach of [46], while assuming that the total number of IDD iterations is the same.

Figure 2.6 indicates that, despite the severe overloading conditions, all the methods aided by LDPC code exhibit the waterfall BER curves typical of receivers free of ISI or MUI. Comparing the performance in case of no IDD iteration, the proposed IDLS demonstrates its advantage against the other two methods, while the ones with multiple IDD iterations are significantly improved as the IDD iterations increase. Please note that in the figure, "#" indicates the number of IDD iterations. Taking into account the fact that [46] adopts a variant of the LMMSE detector, this figure directs us to a possible IDD extension of the IDLS framework. This is due to the fact that as shown in Remark 2.3.2, the proposed IDLS framework is a generalization of the conventional ZF detectors for compliance with the prescribed discrete constellation set, which is however left for future work.

It is this overall combination of high performance and low-complexity offered by the IDLS scheme that motivates us, in the subsequent section, to further develop its robustness against various other factors of practical relevance such as noisy conditions, CSI imperfection and hardware impairments, with great modularity, consequently extending the IDLS method into a more general framework for the symbol detection of overloaded systems.



(a) Convergence of IDLS.



(b) Asymptotic behavior of IDLS.

Figure 2.5: Convergence and asymptotic behaviors of the IDLS detector.

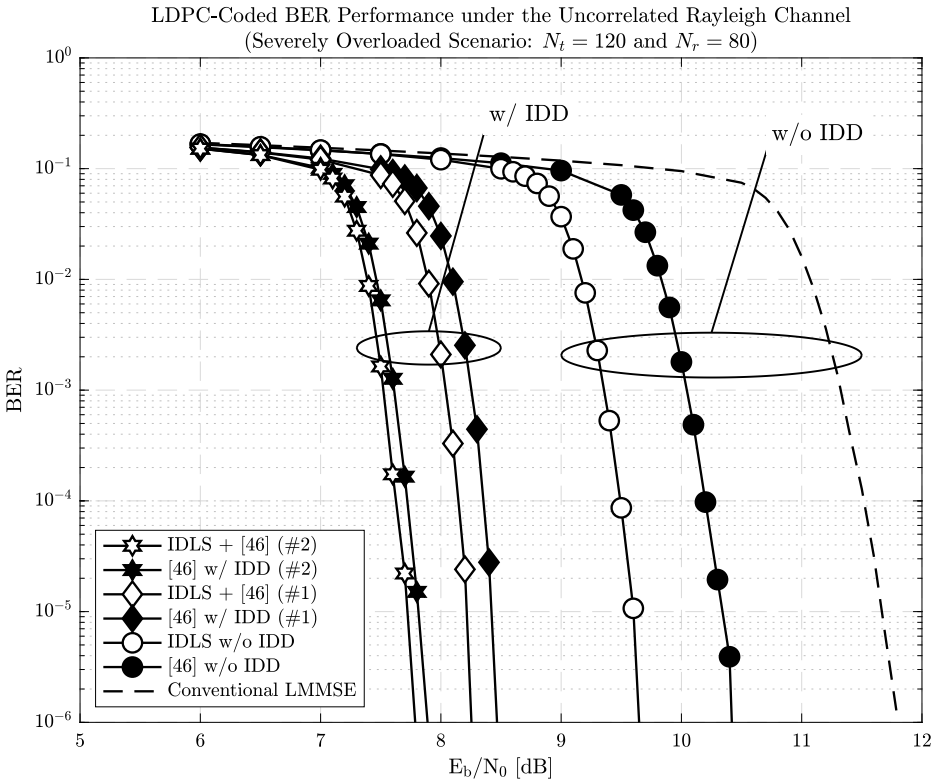


Figure 2.6: LDPC-coded BER performance with perfect CSI.

2.4 Extension for Robust IDLS Detector

In the preceding section, we constructed a novel ML-derived iterative discrete least squares detector which, taking advantage of a tight convexified approximation of the ℓ_0-norm, enables an efficient detection of symbol vectors subject to rank-deficient channels resulting from overloading or correlation. The new scheme, which we refer to as the IDLS detector, was found to yield BER performances that are significantly superior to those of the classic LMMSE receiver and of more recent SotA methods.

The IDLS detector was, however, derived and evaluated under the ideal assumption of perfect CSI knowledge and impairment-free hardware, while in practice CSI errors and hardware imperfection are possible additional causes of performance degradation. And although CSI and hardware imperfections can be mitigated by increasing the transmit power and quantity of pilot symbols, and the quality of radio frequency (RF) components, respectively, these measures come along with undesirable consequences such as the increase in energy consumption, communication delay and engineering costs.

We therefore proceed to show in this section that the IDLS detector can be extended so as to incorporate robustness to the aforementioned factors. In the process, it will become visible that, thanks to the flexibility of its formulation and the simplic-

ity of its solution, the IDLS scheme is indeed very convenient in enabling these and possibly further extensions, implicating therefore that the method can be seen as a generic framework for the design of robust, effective and low-complexity detectors for overloaded systems.

2.4.1 Mitigating Noisy Conditions

In many cases, the effects of imperfections in CSI or hardware are perceived in the form of an increase in the noise level experienced at the receiver, without the possibility for the receiver to estimate proportionality parameters that quantify the percentage of such noise increase relative to thermal and background RF noise.

It makes sense, therefore, to start our effort of extending the IDLS detector by considering a mechanism to combat a generic and uncategorized noise level, under the knowledge of only the aggregate noise power $\sigma_{\mathbf{n}}^2$. To that end, recall the conventional linear MMSE detector described by the equation

$$\boldsymbol{s}^{\mathrm{MMSE}} = (\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r})^{-1}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}. \tag{2.40}$$

Recall also the Woodbury inverse lemma [71, Sec.2.7.3], which applied to the context hereby establishes the equivalence

$$\boldsymbol{H}^{\mathrm{H}}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} + \sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r})^{-1} = (\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r})^{-1}\boldsymbol{H}^{\mathrm{H}}. \tag{2.41}$$

Next, notice that by force of the Woodbury inverse lemma, equation (2.40) can in fact be recognized as the solution of the following regularized least square problem

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\operatorname{argmin}} \ \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \sigma_{\boldsymbol{n}}^2\|\boldsymbol{s}\|_2^2. \tag{2.42}$$

In light of the discussions above, as well as the derivations and results in Section 2.3, it can now readily be understood that the well-known vulnerability of the classic LMMSE to overloading stems from the lack of a constraint that enforces compliance of the solution to the prescribed symbol constellation $\mathcal{C}$, which can then be corrected by reformulating the latter as

$$\underset{\boldsymbol{s}\in\mathbb{C}^{N_t}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \sigma_{\boldsymbol{n}}^2\|\boldsymbol{s}\|_2^2, \tag{2.43a}$$

$$\text{subject to} \quad \sum_{i=1}^{2^b} \|\boldsymbol{s} - c_i\mathbf{1}\|_0 = N_t \cdot (2^b - 1), \tag{2.43b}$$

which in turn can be transformed into a regularized LS variation, namely

$$\min_{\boldsymbol{s}\in\mathbb{C}^{N_t}} \ \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \sigma_{\boldsymbol{n}}^2\|\boldsymbol{s}\|_2^2 + \lambda\sum_{i=1}^{2^b} \|\boldsymbol{s} - c_i\mathbf{1}\|_0. \tag{2.44}$$

Succinctly, introducing the ℓ_0-norm approximation into equation (2.44) and subsequently applying the QT, we readily obtain

$$\min_{\boldsymbol{x}\in\mathbb{C}^{N_t}} \ \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2+\sigma_{\boldsymbol{n}}^2\|\boldsymbol{s}\|_2^2+\lambda\sum_{i=1}^{2^b}\sum_{j=1}^{N_t}\beta_{i,j}^2|s_j-c_i|^2, \tag{2.45}$$

with $\beta_{i,j}$ given, as previously, by equation (2.14), and which can be rewritten as

$$\min_{\boldsymbol{s}\in\mathbb{C}^{N_t}} \ \boldsymbol{s}^{\mathrm{H}}\Big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+\sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r}+\lambda\boldsymbol{B}\Big)\boldsymbol{s}-2\mathrm{Re}\Big\{\big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}+\lambda\boldsymbol{b}\big)^{\mathrm{H}}\boldsymbol{s}\Big\}, \tag{2.46}$$

with $\boldsymbol{b}$ and $\boldsymbol{B}$ as defined in equations (2.15a) and (2.15b), respectively, and which in turn admits the iterative closed-form solution

$$\boldsymbol{s}=\Big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+\sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r}+\lambda\boldsymbol{B}\Big)^{-1}\big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}+\lambda\boldsymbol{b}\big), \tag{2.47a}$$

or equivalent in the real domain

$$\mathbf{s}=\big(\mathbf{H}^{\mathrm{T}}\mathbf{H}+\sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r}+\lambda\mathbf{B}\big)^{-1}\big(\mathbf{H}^{\mathrm{T}}\mathbf{y}+\lambda\mathbf{b}\big), \tag{2.47b}$$

in which case $\mathbf{s}$ and $\mathbf{H}$ are as in equation (2.2), while $\mathbf{b}$ and $\mathbf{B}$ are defined in equation (2.22).

Remark 2.4.1. *Note that if the noise variance is not available at the receiver (i.e., $\sigma_{\boldsymbol{n}}^2=0$), equation* (2.47) *reduces to the previously derived IDLS detector described by equation* (2.20)*, while setting $\lambda=0$ leads to the conventional linear MMSE detector of equation* (2.40)*, namely,*

$$\lim_{\lambda\to 0^+}\Big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+\sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r}+\lambda\boldsymbol{B}\Big)^{-1}\big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}+\lambda\boldsymbol{b}\big)=\Big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+\sigma_{\boldsymbol{n}}^2\boldsymbol{I}_{N_r}\Big)^{-1}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{y}. \tag{2.48}$$

In other words, it can be concluded that equation (2.47) *is a generalization of both the IDLS and the conventional linear MMSE receivers in terms of noise- and constellation-awareness, respectively.*

As for the regularization parameter λ, once again the same procedure described in Subsection 2.3.3 can be reproduced. In particular, the real-valued equivalent of equation (2.46) can once again be cast into the QCQP-1

$$\underset{\mathbf{s}\in\mathbb{R}^{2N_t}}{\text{minimize}} \ \ \mathbf{s}^{\mathrm{T}}\mathbf{B}\,\mathbf{s}-2\,\mathbf{b}^{\mathrm{T}}\mathbf{s} \tag{2.49a}$$

$$\text{subject to} \ \ \mathbf{s}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H}+\sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r})\,\mathbf{s}-2\,\mathbf{y}^{\mathrm{T}}\mathbf{H}\,\mathbf{s}+\mathbf{y}^{\mathrm{T}}\mathbf{y}-\delta\leq 0, \tag{2.49b}$$

from which the combination of KKT conditions with Moré's Theorem [63] yields

$$\bar{\mathbf{s}} = (\mathbf{B} + \mu^{\text{opt}}(\mathbf{H}^{\text{T}}\mathbf{H}+\sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r}))^{-1}(\mathbf{b} + \mu^{\text{opt}}\mathbf{H}^{\text{T}}\mathbf{y}), \tag{2.50a}$$

$$\bar{\mathbf{s}}^{\text{T}}(\underbrace{(\mathbf{H}^{\text{T}}\mathbf{H}+\sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r})}_{=\mathbf{q}_{22}}\bar{\mathbf{s}} - \mathbf{H}^{\text{T}}\mathbf{y}) - \mathbf{y}^{\text{T}}\mathbf{H}\,\bar{\mathbf{s}} + \mathbf{y}^{\text{T}}\mathbf{y} - \delta = 0. \tag{2.50b}$$

For brevity, we omit the derivation details which are similar to those of Subsection 2.3.3, but solving the system of equations (2.50) in the same manner described in Subsection 2.3.3 and applying the results of [65, Lem.3 and Th.4] and the Möbius-transform, one finally arrives once again at the regularization parameter given in equation (2.38), only with the entry identified in equation (2.50b) updated in the matrices $\mathbf{Q}$ and $\mathbf{P}$, which yields

$$\mathbf{Q} \triangleq \begin{bmatrix} \mathbf{y}^{\text{T}}\mathbf{y} - \delta & -\mathbf{y}^{\text{T}}\mathbf{H} & \mathbf{b}^{\text{T}} \\ -\mathbf{H}^{\text{T}}\mathbf{y} & \mathbf{H}^{\text{T}}\mathbf{H} + \sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r} & -\mathbf{B} \\ \mathbf{b} & -\mathbf{B} & \mathbf{0}_{2N_t} \end{bmatrix}, \tag{2.51a}$$

$$\mathbf{P} \triangleq \begin{bmatrix} 0 & \mathbf{0}_{1\times 2N_t} & -\mathbf{y}^{\text{T}}\mathbf{H} \\ \mathbf{0}_{2N_t\times 1} & \mathbf{0}_{2N_t} & \mathbf{H}^{\text{T}}\mathbf{H} + \sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r} \\ -\mathbf{H}^{\text{T}}\mathbf{y} & \mathbf{H}^{\text{T}}\mathbf{H} + \sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r} & \mathbf{0}_{2N_t} \end{bmatrix}. \tag{2.51b}$$

From the above, it is evident that replacing (2.20b) by equation (2.47b), as well as equations (2.36a) and (2.36b) by equations (2.51a) and (2.51b) yields a noise-robust extension of the IDLS detector, which maintains the same overall structure and thus the same complexity, while offering improved BER performance at low SNRs. We also remark, however, that at high SNRs or if the noise power is unknown and thus set to zero, the term $\sigma_{\mathbf{n}}^2\mathbf{I}_{2N_r}$ in equations (2.47b) and (2.51) disappear, reducing them to equations (2.20b) and (2.36), respectively, and consequently the detector itself back to the original IDLS described in Section 2.3.

In case $\sigma_{\mathbf{n}}^2$ is unknown or negligibly small, a value of δ can be chosen in $\mathbf{Q}$, for instance, based on the minimum Euclidean distance between points in the constellation $\mathcal{P}$. In other words, taking advantage of lattice reduction approaches *e.g.,* [61], a robust choice of δ can be made as a function of the dominant eigenvalue of the channel matrix.

With the modularity of the IDLS framework well identified, we proceed to further extend it to mitigate imperfect CSI and hardware impairments as previously announced.

2.4.2 Mitigating Imperfect CSI and Hardware Impairments

For the sake of brevity, in this subsection we omit repetitive details and succinctly describe only the necessary modifications in the channel and system models, as well as in the fundamental equations of the IDLS framework, required to identify how the latter can be extended to also mitigate imperfect CSI and hardware impairments.

First, consider the well-known Gauss-Markov uncertainty model [72] commonly used to incorporate the impact of CSI errors in channel estimates, which is described by

$$\boldsymbol{H} = \sqrt{1-\tau^2}\hat{\boldsymbol{H}} + \tau\boldsymbol{E}, \tag{2.52}$$

where $\hat{\boldsymbol{H}}$ denotes the estimate of the true channel matrix $\boldsymbol{H}$ (*i.e.*, imperfect observation of $\boldsymbol{H}$), $\boldsymbol{E}$ corresponds to the error matrix whose distribution is associated with that of $\boldsymbol{H}$, and $\tau \in [0,1]$ denotes the Gauss-Markov uncertainty parameter that characterizes the CSI estimation inaccuracy.

For the sake of completeness, we combine the CSI error model of (2.52) with the channel correlation model described in equation (2.39) to obtain [72]

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{\Phi}_r^{\frac{1}{2}} \left(\sqrt{1-\tau^2}\boldsymbol{H}_{\text{i.i.d.}} + \tau\boldsymbol{E}_{\text{i.i.d.}}\right) \boldsymbol{\Phi}_t^{\frac{1}{2}} \\ &= \sqrt{1-\tau^2}\underbrace{\boldsymbol{\Phi}_r^{\frac{1}{2}}\boldsymbol{H}_{\text{i.i.d.}}\boldsymbol{\Phi}_t^{\frac{1}{2}}}_{\triangleq\hat{\boldsymbol{H}}\text{ (known)}} + \tau\underbrace{\boldsymbol{\Phi}_r^{\frac{1}{2}}\boldsymbol{E}_{\text{i.i.d.}}\boldsymbol{\Phi}_t^{\frac{1}{2}}}_{\triangleq\boldsymbol{E}\text{ (unknown)}},\end{aligned} \tag{2.53}$$

where $\boldsymbol{E}_{\text{i.i.d.}}$ follows the circular symmetric complex Gaussian distribution with zero mean and unit variance while assuming that perfect (or considerably accurate) knowledge of the spatial correlation matrices $\boldsymbol{\Phi}_r$ and $\boldsymbol{\Phi}_t$ is available at the receiver, since such correlations vary much slower than the instantaneous channel, and therefore can be estimated accurately even though the channel estimates $\hat{\boldsymbol{H}}$ themselves are imperfect [72].

Notice that equations (2.52) and (2.53) indeed extend the perfect CSI model employed in the preceding sections, such that setting $\tau = 0$, one returns simply to $\boldsymbol{H} = \hat{\boldsymbol{H}}$. Next, we turn our attention to hardware impairments, which typically refer to practical imperfections in power amplifiers (PAs), digital-to-analog converters (DACs), analog-to-digital converters (ADCs), I/Q mixers or other RF chain components, and whose effect is to cause distortions in transmit signals that were shown in *e.g.* [73] to be well-modeled by i.i.d additive zero-mean Gaussian random variables, with variance proportional to the power of the undistorted signal. To elaborate, in the presence of hardware impairment, an intended transmit symbol vector $\boldsymbol{s} \in \mathcal{C}^{N_t \times 1}$ is distorted into the transmit signal $\boldsymbol{x} \in \mathbb{C}^{N_t \times 1}$ described by [74, 75]

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{w}, \tag{2.54}$$

where $\boldsymbol{w}$ denotes an additive hardware distortion vector modeled as $\boldsymbol{w} \sim \mathcal{CN}\left(\mathbf{0}, \eta \cdot \mathrm{diag}\left(\mathbf{C}_{\boldsymbol{s}}\right)\right)$ with η denoting the RF distortion level parameter characterized by the quality of the RF chain components and $\mathbf{C}_{\boldsymbol{s}} \triangleq \mathbb{E}\left[\boldsymbol{s}^{\mathrm{H}}\boldsymbol{s}\right]$.

Assuming that the elements of the intended symbol vector $\boldsymbol{s}$ are independent from each other and have unit power (*i.e.,* $\mathbf{C}_{\boldsymbol{s}} = \mathbf{I}$), the received signal corresponding to the transmit signal in equation (2.54) is given by

$$\begin{aligned}\boldsymbol{y} =& \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \\ =& \underbrace{\sqrt{1-\tau^2}\hat{\boldsymbol{H}}\boldsymbol{s}}_{\text{Intended}} + \overbrace{\tau\boldsymbol{E}\boldsymbol{s} + \underbrace{\tau\boldsymbol{E}\boldsymbol{w}}_{\text{Hardware impairment}}}^{\text{CSI imperfection}} + \sqrt{1-\tau^2}\hat{\boldsymbol{H}}\boldsymbol{w} + \boldsymbol{n} \in \mathbb{C}^{N_r \times 1}.\end{aligned} \tag{2.55}$$

It will prove convenient to normalize the received signal by the scalar $\sqrt{1-\tau^2}$, and rearrange the terms so as to yield

$$\bar{\boldsymbol{y}} = \hat{\boldsymbol{H}}\boldsymbol{s} + \underbrace{\hat{\boldsymbol{H}}\boldsymbol{w} + \frac{\tau\boldsymbol{E}\boldsymbol{s} + \tau\boldsymbol{E}\boldsymbol{w} + \boldsymbol{n}}{\sqrt{1-\tau^2}}}_{\triangleq \tilde{\boldsymbol{n}}}, \tag{2.56}$$

where we have implicitly defined the total effective noise $\tilde{\boldsymbol{n}}$.

At this point it is worth mentioning that in the absence of knowledge of the Gauss-Markov uncertainty parameter τ and the RF distortion level η, a receiver would simply perceive the total effective noise $\tilde{\boldsymbol{n}}$ as a higher noise level with power $\sigma^2_{\tilde{\boldsymbol{n}}}$, such that the robust IDLS scheme of Subsection 2.4.1 could be in fact employed directly, as long as $\sigma^2_{\tilde{\boldsymbol{n}}}$ could be estimated, under the assumption of whiteness of $\tilde{\boldsymbol{n}}$.

On the other hand, if τ and η can themselves be estimated, and under the knowledge of the spatial correlation matrices $\boldsymbol{\Phi}_r$ and $\boldsymbol{\Phi}_t$, a more effective mitigation of the CSI and hardware imperfection can be achieved under the IDLS framework, by incorporating the resulting knowledge.

First, as shown in Appendix A, the covariance matrix of the total effective noise $\tilde{\boldsymbol{n}}$ is given by

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}} = \underbrace{\eta\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^{\mathrm{H}} + \frac{\tau^2}{1-\tau^2}(1+\eta)\mathrm{Tr}\left(\boldsymbol{\Phi}_t\right)\boldsymbol{\Phi}_r}_{\triangleq \boldsymbol{\Sigma}^{\mathrm{C}}_{\tilde{\boldsymbol{n}}}} + \underbrace{\frac{\sigma^2_{\boldsymbol{n}}}{1-\tau^2}\mathbf{I}_{N_r}}_{\triangleq \boldsymbol{\Sigma}^{\mathrm{U}}_{\tilde{\boldsymbol{n}}}}, \tag{2.57}$$

where, for future convenience, we have decomposed $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}$ into a sum of the matrices $\boldsymbol{\Sigma}^{\mathrm{C}}_{\tilde{\boldsymbol{n}}}$ and $\boldsymbol{\Sigma}^{\mathrm{U}}_{\tilde{\boldsymbol{n}}}$, the first corresponding to quantities that are subjected to correlation, and the second corresponding to terms that are not.

Next, we observe that the presence of the correlated covariance component $\boldsymbol{\Sigma}^{\mathrm{C}}_{\tilde{\boldsymbol{n}}}$ in $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}$ implicates that, in general,

$$\hat{\boldsymbol{H}}^{\mathrm{H}}(\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^{\mathrm{H}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}})^{-1} \neq (\hat{\boldsymbol{H}}^{\mathrm{H}}\hat{\boldsymbol{H}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}})^{-1}\hat{\boldsymbol{H}}^{\mathrm{H}}, \tag{2.58}$$

such that the total effective noise $\tilde{\boldsymbol{n}}$ due to CSI imperfection and hardware impairment does not enjoy the homoscedasticity assumed in the previous formulations of IDLS described in preceding sections.

To circumvent this new challenge, we consider an extension of the IDLS framework under the prism of the generalized total least square regression problem [76], which we here modify to include our constellation-aware ℓ_0-norm regularizer, yielding

$$\min_{\boldsymbol{s}\in\mathbb{C}^{N_t}} \; (\bar{\boldsymbol{y}} - \hat{\boldsymbol{H}}\boldsymbol{s})^{\mathrm{H}}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}^{\mathrm{C}} + \boldsymbol{I}_{N_r})^{-1}(\bar{\boldsymbol{y}} - \hat{\boldsymbol{H}}\boldsymbol{s}) + \frac{\sigma_{\boldsymbol{n}}^2}{1-\tau^2}\|\boldsymbol{s}\|_2^2 + \lambda \sum_{i=1}^{2^b} \|\boldsymbol{s} - c_i\mathbf{1}\|_0, \tag{2.59}$$

which seeks to minimize the Mahalanobis distance between the output and input vectors while enforcing the constellation-compliance of the solution.

Proceeding succinctly hereafter, introducing the ℓ_0-norm approximation of equation (2.12) and applying the QT into equation (2.59), we obtain

$$\min_{\boldsymbol{s}\in\mathbb{C}^{N_t}} \; (\bar{\boldsymbol{y}} - \hat{\boldsymbol{H}}\boldsymbol{s})^{\mathrm{H}}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}^{\mathrm{C}} + \boldsymbol{I}_{N_r})^{-1}(\bar{\boldsymbol{y}} - \hat{\boldsymbol{H}}\boldsymbol{s}) + \frac{\sigma_{\boldsymbol{n}}^2}{1-\tau^2}\|\boldsymbol{s}\|_2^2 + \lambda(\boldsymbol{s}^{\mathrm{H}}\boldsymbol{B}\boldsymbol{s} - 2\Re\{\boldsymbol{s}^{\mathrm{H}}\boldsymbol{b}\}), \tag{2.60a}$$

or alternatively, in the real domain,

$$\min_{\mathbf{s}\in\mathbb{C}^{N_t}} \; (\bar{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{s})^{\mathrm{T}}(\boldsymbol{\Sigma}_{\tilde{\mathbf{n}}}^{\mathrm{C}} + \mathbf{I}_{2N_r})^{-1}(\bar{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{s}) + \frac{\sigma_{\mathbf{n}}^2}{1-\tau^2}\|\mathbf{s}\|_2^2 + \lambda(\mathbf{s}^{\mathrm{T}}\mathbf{B}\mathbf{s} - 2\mathbf{s}^{\mathrm{T}}\mathbf{b}), \tag{2.60b}$$

which in turn admit the respective solutions

$$\boldsymbol{s} = \Big(\hat{\boldsymbol{H}}^{\mathrm{H}}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}^{\mathrm{C}} + \boldsymbol{I}_{N_r})^{-1}\hat{\boldsymbol{H}} + \tfrac{\sigma_{\boldsymbol{n}}^2}{1-\tau^2}\boldsymbol{I}_{N_t} + \lambda\boldsymbol{B}\Big)^{-1} \times \Big(\hat{\boldsymbol{H}}^{\mathrm{H}}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}^{\mathrm{C}} + \boldsymbol{I}_{N_r})^{-1}\bar{\boldsymbol{y}} + \lambda\boldsymbol{b}\Big), \tag{2.61a}$$

and

$$\mathbf{s} = \Big(\hat{\mathbf{H}}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\tilde{\mathbf{n}}}^{\mathrm{C}} + \mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}} + \tfrac{\sigma_{\mathbf{n}}^2}{1-\tau^2}\mathbf{I}_{2N_t} + \lambda\mathbf{B}\Big)^{-1} \times \big(\hat{\mathbf{H}}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\tilde{\mathbf{n}}}^{\mathrm{C}} + \mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}} + \lambda\mathbf{b}\big), \tag{2.61b}$$

where, as previously, all the straight up bold letters are the real-domain equivalent of their bold italic counterparts in the complex domain.

Finally, in order to optimize the regularization parameter, we once again transform the formulation given in equation (2.60b) into a QCQP-1, build the corresponding KKT conditions, and apply Moré's Theorem [63] to obtain

$$\bar{\mathbf{s}} = \Big(\mathbf{B} + \mu^{\text{opt}}\big(\hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}} + \mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}} + \tfrac{\sigma^2_{\mathbf{n}}}{1-\tau^2}\mathbf{I}_{2N_t}\big)\Big)^{-1} \times\Big(\mathbf{b} + \mu^{\text{opt}}\hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}} + \mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}}\Big), \tag{2.62a}$$

$$\bar{\mathbf{s}}^{\text{T}}\Big(\overbrace{(\hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}}+\tfrac{\sigma^2_{\mathbf{n}}}{1-\tau^2}\mathbf{I}_{2N_t})}^{=\mathbf{q}_{22}}\bar{\mathbf{s}} - \hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}}\Big) \underbrace{-\bar{\mathbf{y}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}}}_{=\mathbf{q}_{12}}\bar{\mathbf{s}} + \underbrace{\bar{\mathbf{y}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}} + \mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}} - \delta}_{=q_{11}} = 0, \tag{2.62b}$$

such that the matrices $\mathbf{Q}$ and $\mathbf{P}$ to be used in equation (2.38) in order to calculate λ^{opt} are updated as follows.

$$\mathbf{Q} \triangleq \begin{bmatrix} \bar{\mathbf{y}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}} + \mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}} - \delta & -\bar{\mathbf{y}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}} & \mathbf{b}^{\text{T}} \\ -\hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}} & \hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}}+\tfrac{\sigma^2_{\mathbf{n}}}{1-\tau^2}\mathbf{I}_{2N_t} & -\mathbf{B} \\ \mathbf{b} & -\mathbf{B} & \mathbf{0}_{2N_t} \end{bmatrix}, \tag{2.63a}$$

$$\mathbf{P} \triangleq \begin{bmatrix} 0 & \mathbf{0}_{1\times 2N_t} & -\bar{\mathbf{y}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}} \\ \mathbf{0}_{2N_t\times 1} & \mathbf{0}_{2N_t} & \hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}}+\tfrac{\sigma^2_{\mathbf{n}}}{1-\tau^2}\mathbf{I}_{2N_t} \\ -\hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}} & \hat{\mathbf{H}}^{\text{T}}(\boldsymbol{\Sigma}^{\text{C}}_{\tilde{\mathbf{n}}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}}+\tfrac{\sigma^2_{\mathbf{n}}}{1-\tau^2}\mathbf{I}_{2N_t} & \mathbf{0}_{2N_t} \end{bmatrix}. \tag{2.63b}$$

Remark 2.4.2. *We remark that in case perfect CSI is available at the receiver (i.e., when $\tau = 0$), and the transmitter is free of distortions due to hardware impairments (i.e., when $\eta = 0$), equations* (2.59) *through* (2.63) *reduce to the corresponding equations (i.e., equations* (2.47) *and* (2.50)*) in Subsection 2.4.1. Likewise, setting the noise variance to zero further reverts the detector described in this subsection back to the fundamental IDLS design introduced in Section 2.3, as shown in Remark 2.4.1. In other words, it is evident that the robust IDLS variation introduced here is in fact an extension of the original IDLS method presented earlier, with embedded robustness to noise, CSI imperfection and hardware impairments, for which it shall be referred to as the Robust IDLS detector.*

Also, equation (2.61) converges to a stationarity point similarly to equation (2.20).

For the convenience of the reader, a pseudo-code summarizing the Robust IDLS detector is offered above in Algorithm 2. It can be readily recognized that indeed Algorithm 2 is generalization of Algorithm 1.

2.4.3 Performance Assessment: Robust IDLS

In this section, we offer performance evaluation of the Robust IDLS detector for different CSI error and hardware impairment setups in order to illustrate the effectiveness

Algorithm 2 Robust IDLS Detector

Inputs: Received signal $\mathbf{y}$, channel matrix $\mathbf{H}$ and noise power $\sigma_{\mathbf{n}}^2$, Gauss-Markov uncertainty parameter τ and hardware impairment parameter η.
Outputs: Estimate $\hat{\boldsymbol{s}}$

1: Set iteration counter $k = 0$.
2: Normalize received signal as $\bar{\mathbf{y}} \triangleq \frac{\mathbf{y}}{\sqrt{1-\tau^2}}$.
3: Set initial solution to $\mathbf{s}^{(k)}=\Big(\hat{\mathbf{H}}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\tilde{\mathbf{n}}}^{\mathrm{C}}+\mathbf{I}_{2N_r})^{-1}\hat{\mathbf{H}}+\frac{\sigma_{\mathbf{n}}^2}{1-\tau^2}\mathbf{I}_{2N_t}\Big)^{-1}\big(\hat{\mathbf{H}}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\tilde{\mathbf{n}}}^{\mathrm{C}}+\mathbf{I}_{2N_r})^{-1}\bar{\mathbf{y}}\big)$.
4: **repeat**
5: Increase iteration counter $k = k + 1$.
6: Update $\beta_{i,j} \forall\, i, j$, as in equation (2.21).
7: Construct $\mathbf{b}$ and $\mathbf{B}$ from equations (2.22a) and (2.22b).
8: Construct $\mathbf{Q}$ and $\mathbf{P}$ from equations (2.63a) and (2.63b).
9: Obtain $\lambda^{\mathrm{opt}(k)}$ as in equation (2.38).
10: Update $\mathbf{s}^{(k)}$ as in equation (2.61b).
11: **until** convergence or maximum iterations reached
12: $\hat{\boldsymbol{s}} \leftarrow \mathbf{s}^{(k)}$

and flexibility of the proposed framework.

Before we proceed to the BER performance assessment of the Robust IDLS detector compared to the state of the art, let us offer a few comments on the complexity of the framework in the following subsection.

2.4.4 Complexity Analysis

For starters, we emphasize that although the computation of the regularization parameter $\lambda^{\mathrm{opt}(k)}$ obtained from the robust IDLS was included among the repetitive steps executed in Algorithms 1 and 2, for the sake of completeness, in practice such step does not need to be repeated at every iteration of the IDLS detector. Indeed, notice that only the first row and column of the matrix $\mathbf{Q}$ is signal-dependent while the matrix $\mathbf{P}$ is constant for a given channel realization. Consequently, it is both mathematically justified and empirically observed that the sequences of regularization parameters $\{\lambda^{\mathrm{opt}(0)}, \lambda^{\mathrm{opt}(1)}, \cdots, \lambda^{\mathrm{opt}(k_{\max})}\}$ obtained from Algorithm 1 and 2 do not vary significantly with channel realizations and signal transmissions, for fixed system size and SNR.

That, allied with the fact that in SotA methods the optimization of λ is typically performed via exhaustive search, justifies us not to consider the optimization of λ as a core contributor to the complexity of the IDLS framework. Consequently, the most expensive operation in the IDLS framework is the matrix inversion required to

Table 2.1: Table of Complexity Order.

Method	IDLS and Robust IDLS	SOAV [40] (SotA)	SCSR [45] (SotA)	[46] (SotA)
Complexity Order	$\mathcal{O}(N_t^3)$	$\mathcal{O}(N_t^3)$	$\mathcal{O}(N_t^3)$	$\mathcal{O}(N_t^3)$

evaluate the closed-form iterative expression in equations (2.20), (2.47) and (2.61), which requires $2N_t^3/3 + \mathcal{O}(N_t^2)$ floating point operations (flops) to complete, if carried out via a conventional matrix inversion algorithm such as the naive Gauss-Jordan elimination method [64].

However, since the quadratic coefficient matrices in equations (2.20), (2.47) and (2.61) are Hermitian positive definite, the latter complexity can be halved by taking advantage of the Cholesky factorization, which transforms the latter equations into a linear system involving triangular matrices, resulting in an order of $N_t^3/3 + \mathcal{O}(N_t^2)$ flops [77].

We have numerically found via Monte-Carlo simulations, for instance, that the Cholesky factorization approach suffices to efficiently solve the inversions under moderate problem setups (*i.e.*, $N_t = N_r \approx 100$) within an average time of hundreds of microseconds when using 64-bit MATLAB 2019a in a computer with an Intel Core i9 processor with a clock speed of 3.6GHz and 32GB of random-access memory (RAM). Thus, the operation in practice can be computed in the order of nanoseconds on field programmable gate arrays (FPGAs), which complies with latency requirements posed by various global standards including 5G NR. Note that efficient factorization-based inverse solvers are publicly available as native functions in numerical computing languages, *e.g.*, `mldivide` in MATLAB and `cho_solve` in Python.

Remark 2.4.3. *The complexity order of all the algorithms, both proposed (i.e., IDLS and Robust IDLS) and SotA (i.e., SOAV, SCSR and [46]) is proportional to $\mathcal{O}(N_t^3)$, which incidentally is also the worst-case complexity of the MMSE detector that is currently employed in practice [62].*

We remark, however, that further complexity reduction techniques developed for massive MIMO systems can be employed (e.g., [78]), leading to $\mathcal{O}(N_t^2)$ at best. It can be found via Monte-Carlo simulations, for instance, that the Cholesky-based factorization can offer a reasonable runtime performance in order to solve equations (2.20)*,* (2.47)*, and* (2.61) *in case of moderately large systems, although this approach might also suffer when the system dimensions become extremely large.*

For such systems of larger size (*e.g.*, $N_t \gg 500$), numerous iterative algorithms including variates of the conjugate gradient methods, accelerated first/second-order gradient methods, and minimum residual methods exist [64], which can also be uti-

lized to solve the linear equations (2.19), (2.47), and (2.61). Furthermore, quantum algorithms for linear equations [79] will be commercially available in the future, which may further accelerate the speed.

All in all, we summarize in Table 2.1 the complexity order of the proposed IDLS and its precedents (*i.e.*, SOAV and SCSR).

2.4.5 BER Performance: IDLS versus SotA Alternatives

Next, we turn our attention to the evaluation of the BER performance of Algorithm 2 in comparison to the conventional LMMSE and recent SotA alternatives, in mitigating the effect of imperfections such as CSI errors and hardware impairments, inevitable in real-life systems. For the sake of a more effective comparison, in particular in terms of capturing the gains achieved as a result of the IDLS approach itself rather then the system model alone, and given that the derivation of the total effective noise covariance matrix described in Appendix A and summarized in equation (2.57) is adjacent to the IDLS detector, which therefore can also be utilized outside of the context of the IDLS framework, we have fed the conventional LMMSE estimator with $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}}$ as described by equation (2.57).

In other words, in the figures to follow, curves attributed to the "conventional" LMMSE corresponds to the results obtained with the receiver

$$\boldsymbol{s}^{\mathrm{MMSE}} = \hat{\boldsymbol{H}}^{\mathrm{H}}(\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^{\mathrm{H}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}})^{-1}\bar{\boldsymbol{y}}. \tag{2.64}$$

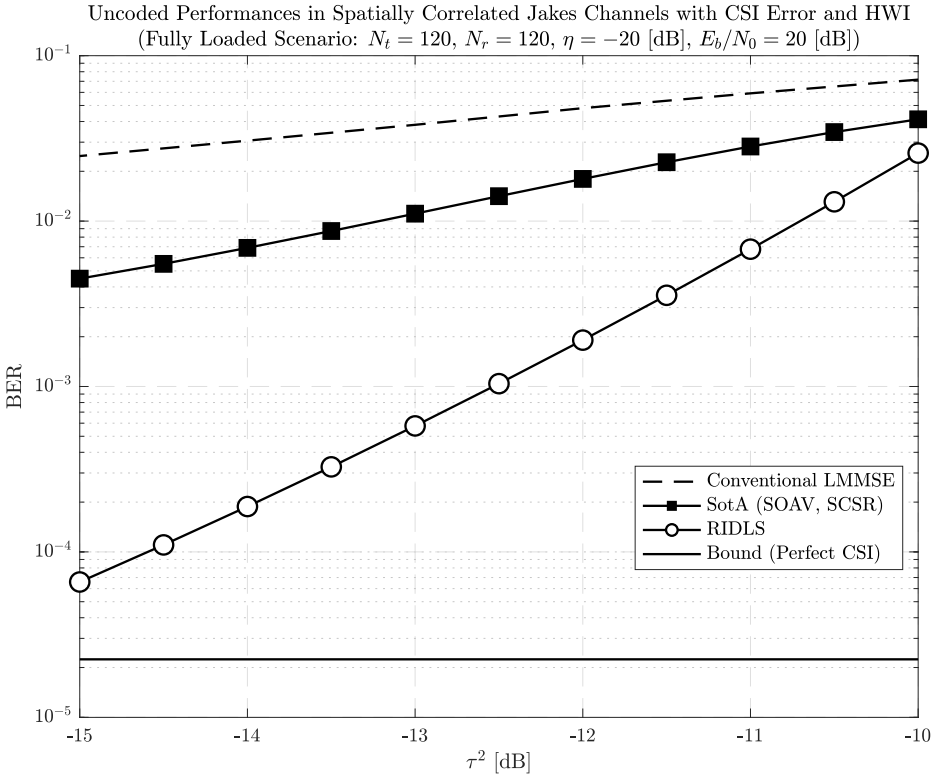
Following [74] and [75], it is assumed that the CSI imperfection level τ^2 varies from -15 [dB] to -10 [dB], while the hardware imperfection parameter η takes from the interval $[-20, -10]$ [dB].

Fully loaded scenarios will be set with $N_t = N_r = 120$, whereas overloaded scenarios will be set with $N_t = 120$ and $N_r = 96$, yielding an overloading ratio of $\gamma = 1.25$ so that the results of Section 2.3 may serve as a reference.

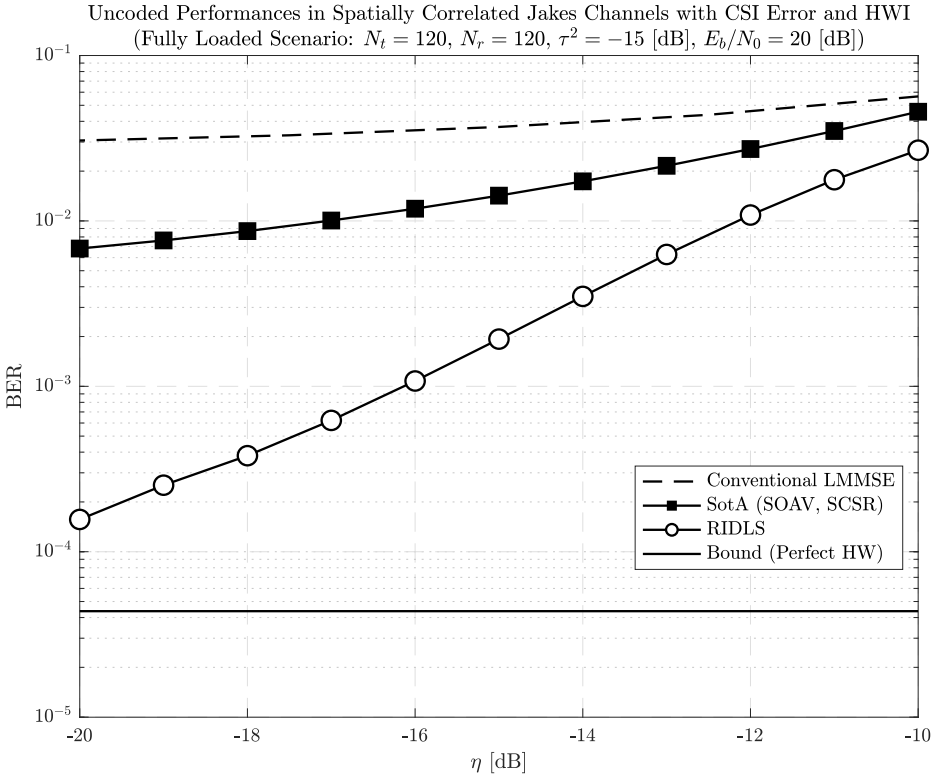
Our comparisons start with Figure 2.7, which shows the gains in BER performance achieved by the robust IDLS detector over SotA alternatives in a fully-loaded scenario (*i.e.,* $N_t = N_r = 120$) with different CSI error and hardware imperfection condition, respectively.

It is found that the conventional LMMSE estimator is significantly outperformed not only by the newly-proposed robust IDLS detector, but also by the SOAV and SCSR schemes, which follows from the constellation-awareness that these techniques have in common, and which demonstrates its robustness of the approach against non-ideal conditions such as CSI and hardware imperfections.

It can also be seen, however, that the robust IDLS consistently outperforms all

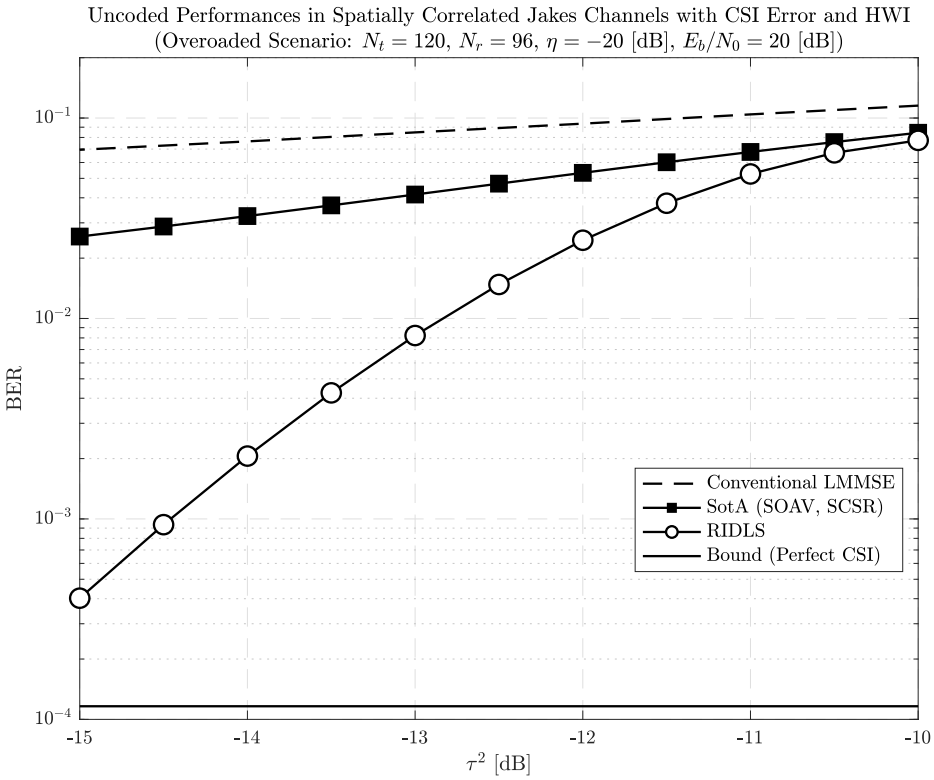


(a) Impact of CSI Error with Fixed Hardware Impairment Level.

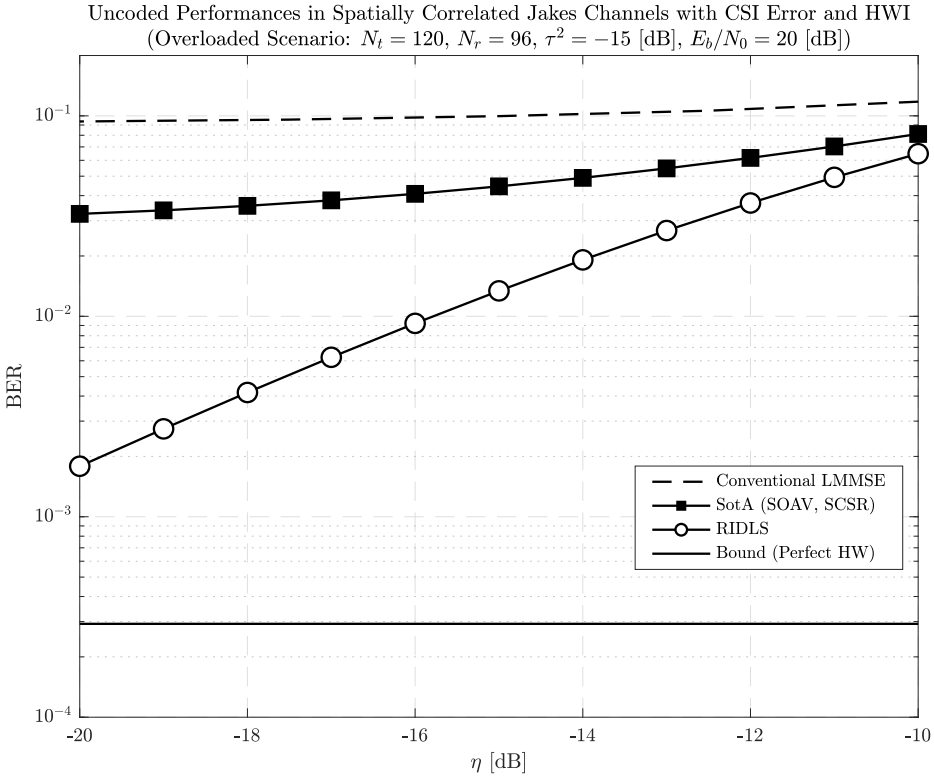


(b) Impact of Hardware Impairment with Fixed CSI Error Level.

Figure 2.7: Uncoded BER performance of Robust IDLS detector compared to SotA alternatives, under fully-loaded ($\gamma = 1$) conditions in spatially correlated channels and subjected to different CSI error and hardware impairment levels.



(a) Impact of CSI Error with Fixed Hardware Impairment Level.



(b) Impact of Hardware Impairment with Fixed CSI Error Level.

Figure 2.8: Uncoded BER performance of Robust IDLS detector compared to SotA alternatives, under overloaded ($\gamma = 1.25$) conditions in spatially correlated channels and subjected to different CSI error and hardware impairment levels.

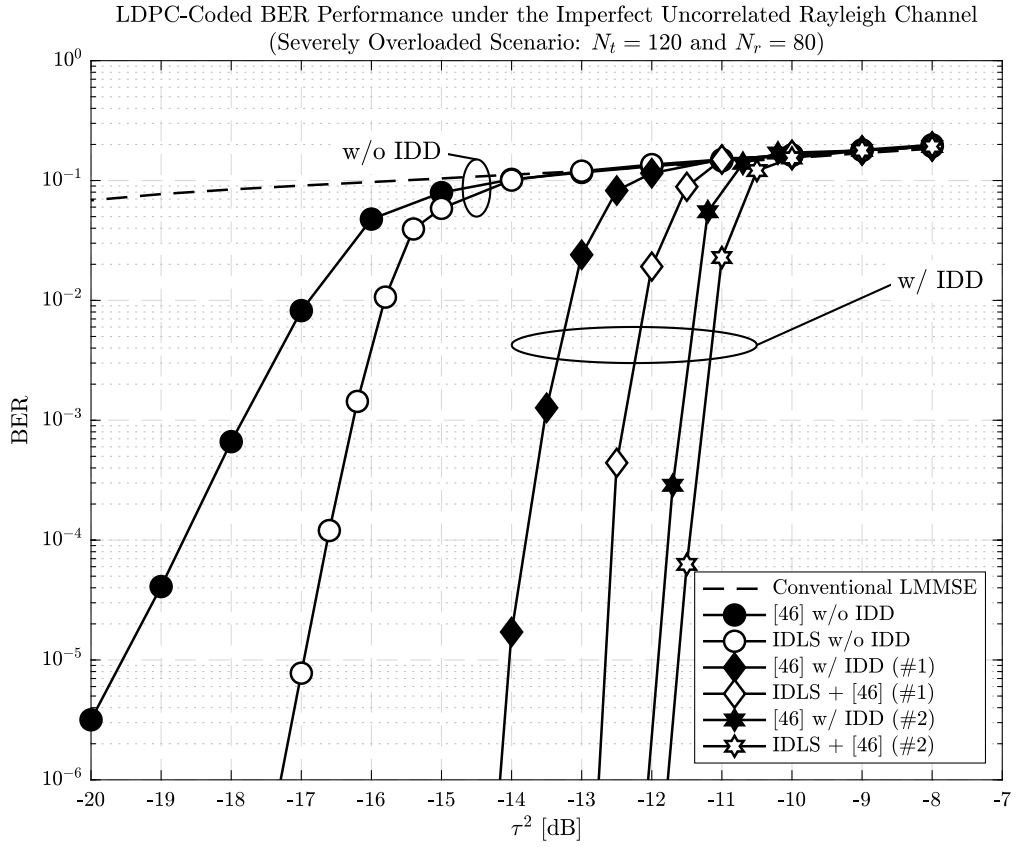


Figure 2.9: Coded BER performance with imperfect CSI.

alternatives including SOAV and SCSR across the entire range of CSI error and hardware imperfection levels, with an especially large gain over the latter over a wide range of values of τ and η.

Next, we examine the impact of overloading on the aforementioned gains in order to reveal the performance in a NOMA setup. To this end, Figure 2.8 compares the uncoded BERs achieved by a system with $N_t = 120$ and $N_r = 96$, yielding an overloading ratio of $\gamma = 1.25$. It can be seen that while the harsher overloading has an overall impact on the performances of all schemes compared, as a consequence of the higher levels of ISI, also exacerbated by channel and noise correlation resulting from the CSI errors and hardware distortions, the rebust IDLS detector continues to provide superior performance with substantial gains over all alternatives.

These gains are a consequence of the incorporation of imperfection-aware extensions via the generalized LS method into the IDLS framework. As confirmed by the figures, it is shown that the proposed framework is generic and compatible with various system setups in a plug-and-play fashion.

Last but not least, the BER performance of the robust IDLS, [46] and the conventional LMMSE under the assumption that the receiver suffers from imperfect CSI

errors is shown in Figure 2.9, where the coding setup follows the same as that of Figure 2.6. The figure shows the performance gain obtained by utilizing the Robust IDLS as an initializer to the IDD method of [46], compared to the technique proposed thereby as is. It can be seen that the Robust IDLS-initialized IDD results in an acceleration of the method of [46] towards capacity-achieving performance.

The results are encouraging as they indicate that an optimized integration of the proposed Robust IDLS detector with decoding is likely to yield further advancement of the state-of-the-art, which is left for a future work.

2.5 Further Application: Low-Rank Matrix Completion

The proposed discreteness-aware regularizer is shown to be effective in NOMA systems. In order to illustrate a wide range of its applicability and flexibility to different signal processing applications, in this section we seek to develop an LRMC algorithm based on the concept of the discreteness-aware regularization approach with the application to a matrix completion problem subject to discreteness constraints. Before moving on to descriptions of the developed algorithm, let us start with the technical background of the related works.

2.5.1 Background and Prior Works

With fair-winds of big data and IoT, modern signal and information processing applications such as information filtering systems, networking, machine learning, and wireless communications often face a structured LRMC problem, which intends to infer a low-rank matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ given a partially observed incomplete matrix $\mathbf{O} \in \mathbb{R}^{m\times n}$ [58, 80, 81]. The matrix completion (MC) optimization problem can be written as the following rank minimization problem:

$$\underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{argmin}} \quad \text{rank}\,(X) \tag{2.65a}$$

$$\text{s.t.} \quad P_{\mathbb{\Omega}}\,(\mathbf{X}) = P_{\mathbb{\Omega}}\,(\mathbf{O})\,, \tag{2.65b}$$

where $\text{rank}\,(\cdot)$ denotes the rank of a given input matrix and $P_{\mathbb{\Omega}}\,(\cdot)$ indicates the mask operator (*i.e.*, projection) defined as

$$[P_{\mathbb{\Omega}}\,(\mathbf{A})]_{ij} = \begin{cases} [\mathbf{A}]_{ij} & \text{if } (i,j) \in \mathbb{\Omega} \\ 0 & \text{otherwise} \end{cases}, \tag{2.66}$$

with $[\cdot]_{ij}$ being the (i,j)-th element of a given matrix and $\mathbb{\Omega}$ denoting the observed index set.

Although the global solution of equation (2.65) corresponds to a matrix that has the lowest rank and matches observations corresponding to indexes belonging to the indicator set $\mathbb{\Omega}$, naively solving the above rank minimization problem is known to be non-deterministic polynomial-time (NP)-hard due to the non-convexity of the rank operator $\mathrm{rank}\,(\cdot)$. Similar to the idea that the ℓ_0-norm function can be replaced by its convex surrogate ℓ_1-norm in CS-related problems, the above rank minimization problem can be relaxed by introducing the nuclear norm (NN) $\|\mathbf{A}\|_*$ (*i.e.*, the sum of the singular values of $\mathbf{A}$) [58], namely,

$$\underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{argmin}} \quad \|\mathbf{X}\|_* \tag{2.67a}$$

$$\text{s.t.} \quad P_{\mathbb{\Omega}}\left(\mathbf{X}\right) = P_{\mathbb{\Omega}}\left(\mathbf{O}\right), \tag{2.67b}$$

where NN is known to be the tightest convex surrogate of the rank operator [82].

Although equation (2.67) can be cast as a semidefinite programming (SDP) problem [83], many different solvers for the latter have been proposed for further complexity reduction and noisy scenarios, which can be categorized as a solution to either the problem:

$$\underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{argmin}} \quad \|\mathbf{X}\|_* \tag{2.68a}$$

$$\text{s.t.} \quad \underbrace{\frac{1}{2}\|P_{\mathbb{\Omega}}\left(\mathbf{X}-\mathbf{O}\right)\|_F^2}_{\triangleq f(\mathbf{X})} \leq \varepsilon, \tag{2.68b}$$

or its regularized form

$$\underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{argmin}} \quad f(\mathbf{X}) + \lambda\|\mathbf{X}\|_*, \tag{2.69}$$

or with the rank information

$$\underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{argmin}} \quad f(\mathbf{X}) \tag{2.70a}$$

$$\text{s.t.} \quad \mathrm{rank}\,(X) \leq s, \tag{2.70b}$$

where $f(\cdot)$ is implicitly defined for notational convenience.

Soft-Impute and its accelerated variates are ones of the state-of-the-art algorithms for large-scale LRMC problems, which aim at solving an optimization problem similar to equation (2.69) and therefore to equation (2.68). To elaborate, Soft-Impute consists of the following recursion

$$\mathbf{X}_{\mathrm{t}} = \mathrm{SVT}_{\lambda}(\mathbf{X}_{\mathrm{t}-1} + P_{\mathbb{\Omega}}\left(\mathbf{O}-\mathbf{X}_{\mathrm{t}-1}\right)), \tag{2.71}$$

where we utilized the fact that $f(\mathbf{X})$ is a convex function with 1-Lipschitz constant,

t denotes the iteration index and the singular value thresholding (SVT) function is given by [84, Theorem 2.1] as

$$\mathrm{SVT}_{\lambda}(\mathbf{A}) \triangleq \mathbf{U}\left(\mathbf{\Sigma} - \lambda\mathbf{I}\right)_{+}\mathbf{V}^{\mathrm{T}}, \tag{2.72}$$

with $\mathbf{A} \triangleq \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ and $(\cdot)_{+}$ being the positive part of the input.

It has recently been shown that Soft-Impute can be categorized as a proximal gradient (PG) algorithm [85], and therefore, the well-known Nesterov-type momentum acceleration technique can be employed without loss of convergence guarantee [86,87], leading to

$$\mathbf{X}_{\mathrm{t}} = \mathrm{SVT}_{\lambda}(\mathbf{Y}_{\mathrm{t}} + P_{\mathbb{\Omega}}\left(\mathbf{O} - \mathbf{Y}_{\mathrm{t}}\right)), \tag{2.73}$$

with $\mathbf{Y}_{\mathrm{t}} \triangleq (1+\beta_{\mathrm{t}})\mathbf{X}_{\mathrm{t}-1} + \beta_{\mathrm{t}}\mathbf{X}_{\mathrm{t}-2}$ where β_{t} is the momentum weight.

2.5.2 Discreteness-Aware LRMC

As recently pointed out in [81], most of the LRMC techniques including ones mentioned above assume that entries of the targeted low-rank matrix are randomly generated (*i.e.*, continuous random variables) in spite of the fact that many real data matrices including recommendation systems are composed of a finite set of discrete numbers (*e.g.*, $1, 2, \ldots, 5$), indicating potential to improve the recovery performance of the existing state-of-the-art algorithms. To this end, in this section we bring the notion of discreteness-awareness by means of regularization to this context, proposing a novel discrete-aware MC algorithm as a sequence of developments [85,88] stemming from Soft-Impute [89]. Notice that the proposed regularizer can be employed in various other MC optimization frameworks, leaving such further extensions to future open problems.

It is also worth noting that one may confuse the word "discrete-aware" with the existing similar research items [90, 91], which exploit binary hashing codes for terminal user devices to reduce the storage volume and time complexity, and therefore are differentiated from the herein proposed method in the problem setup and optimization approach. Also, the proposed method is different from [92–94] in terms of the optimization approach.

Assuming that entries of the matrix to be recovered belong to a certain finite discrete alphabet set $\mathcal{A} \triangleq \{a_1, a_2, \cdots\}$ (*e.g.*, integers in case of recommendation systems), we intend to tackle a variety of the following regularized minimization problem

$$\underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\operatorname{argmin}} \; f(\mathbf{X}) + \lambda g(\mathbf{X}) + \xi r(\mathbf{X}|p), \tag{2.74}$$

where $g(\mathbf{X})$ denotes a non-smooth (possibly non-convex) low-rank regularizer [95], $\xi \geq 0$, and

$$r(\mathbf{X}|p) \triangleq \sum_{k=1}^{|\mathcal{A}|} \|\mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{X}) - a_k\mathbf{1}\|_p \tag{2.75}$$

where $r(\mathbf{X}|p)$ is the discrete-space regularizer[4] with $0 \leq p$, $\mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{X})$ denotes vectorization of entries of $\mathbf{X}$ corresponding to a given index set $\mathbb{\Omega}^c$, and $\mathbb{\Omega}^c$ being the complementary set of $\mathbb{\Omega}$.

Although non-convex scenarios where either $g(\mathbf{X})$, $r(\mathbf{X}|p)$ or both are non-convex regularizer(s) can be considered, we hereafter focus on the convex scenario (*i.e.*, $g(\mathbf{X}) = \|\mathbf{X}\|_*$ and $r(\mathbf{X}|1) = \sum_{k=1}^{|\mathcal{A}|} \|\mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{X}) - a_k\mathbf{1}\|_1$) for the sake of simplicity. The accelerated PG algorithm for a discrete-aware convex variate of Soft-Impute as shown in equation (2.74), which hold the convergence rate $\mathcal{O}\big(\frac{1}{t^2}\big)$, can be summarized as the following recursion:

$$\mathbf{Y}_{\mathrm{t}} = (1+\beta_{\mathrm{t}})\mathbf{X}_{\mathrm{t}-1} + \beta_{\mathrm{t}}\mathbf{X}_{\mathrm{t}-2} \tag{2.76a}$$

$$\mathbf{Z}_{\mathrm{t}} = \mathrm{prox}_{\xi r}(\mathbf{Y}_{\mathrm{t}}) \tag{2.76b}$$

$$\mathbf{X}_{\mathrm{t}} = \mathrm{SVT}_{\lambda}(P_{\mathbb{\Omega}^c}(\mathbf{Z}_{\mathrm{t}}) + P_{\mathbb{\Omega}}(\mathbf{O})) \tag{2.76c}$$

where $\mathrm{prox}_{\xi r}(\mathbf{Y}_{\mathrm{t}})$ is the proximal operator given by

$$\mathrm{prox}_{\xi r}(\mathbf{Y}_{\mathrm{t}}) \triangleq \underset{\mathbf{U}}{\mathrm{argmin}} \;\; r(\mathbf{U}|1) + \frac{1}{2\xi}\|\mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{U} - \mathbf{Y}_{\mathrm{t}})\|_2^2 . \tag{2.77}$$

Taking into account the fact that the proximal operator of a sum of convex regularizers can be computed from a sequence of individual proximal operators [86], we readily obtain

$$\mathrm{prox}_{\xi r}(\mathbf{Y}_{\mathrm{t}}) = \mathrm{prox}_{\xi r_1}\Big(\mathrm{prox}_{\xi r_2}\Big(\cdots \mathrm{prox}_{\xi r_{|\mathcal{A}|}}(\mathbf{Y}_{\mathrm{t}})\Big)\Big), \tag{2.78}$$

where $r_k(\mathbf{Y}_{\mathrm{t}}) \triangleq \|\mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{Y}_{\mathrm{t}}) - a_k\mathbf{1}\|_1$ for $k \in \{1, 2, \ldots, |\mathcal{A}|\}$.

To this end, each proximal operator can be written as

$$\mathrm{prox}_{\xi r_k}(\mathbf{Y}_{\mathrm{t}}) \triangleq \underset{\mathbf{U}}{\mathrm{argmin}} \;\; \|\mathbf{u} - a_k\mathbf{1}\|_1 + \frac{1}{2\xi}\|\mathbf{u} - \mathbf{y}_{\mathrm{t}}\|_2^2 , \tag{2.79}$$

with $\mathbf{u} \triangleq \mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{U})$ and $\mathbf{y}_t \triangleq \mathrm{vec}_{\mathbb{\Omega}^c}(\mathbf{Y}_t)$, which can be compactly written element-by-element as

$$\underset{\bar{u}_\ell}{\mathrm{argmin}} \;\; |\bar{u}_\ell| + \frac{1}{2\xi}(\bar{u}_\ell - \bar{y}_{\mathrm{t},\ell})^2, \tag{2.80}$$

where $\bar{u}_\ell \triangleq [\mathbf{u}]_\ell - a_k$, $\bar{y}_{\mathrm{t},\ell} \triangleq [\mathbf{y}_t]_\ell - a_k$, $\bar{\mathbf{u}} \triangleq [\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_{|\mathbb{\Omega}^c|}]^{\mathrm{T}}$, $\bar{\mathbf{y}}_{\mathrm{t}} \triangleq [\bar{y}_{\mathrm{t},1}, \bar{y}_{\mathrm{t},2}, \ldots, \bar{y}_{\mathrm{t},|\mathbb{\Omega}^c|}]^{\mathrm{T}}$

[4] Although it has been shown that the base of the norm function is set to be $p = 0$ or $p = 1$ to enhance the discreteness of the inputs, the base p can be any positive number in principle.

and $\ell \in \{\mathbb{Z}|1 \leq \ell \leq |\Omega^c|\}$.

One readily notice that equation (2.80) has a closed form solution (*i.e.*, soft-thresholding function) given by

$$\bar{\mathbf{u}} = \text{sign}\left(\bar{\mathbf{y}}_\text{t}\right) \odot \left(|\bar{\mathbf{y}}_\text{t}| - \xi \mathbf{1}\right)_+, \tag{2.81}$$

where $\odot$ is the Hadamard product and sign $(\cdot)$ denotes the (element-wise) sign function.

Notice that in equation (2.81), $|\bar{\mathbf{y}}_\text{t}|$ performs the element-wise absolute operation. Finally we recover $\mathbf{u}$ by

$$\mathbf{u} = \bar{\mathbf{u}} + a_k \mathbf{1}, \tag{2.82}$$

and $\mathbf{U}$ by mapping $\mathbf{u}$ onto the unobserved indexes, namely,

$$\mathbf{U} = \text{vec}_{\Omega^c}^{-1}(\mathbf{u}), \tag{2.83}$$

where $\text{vec}_{\Omega^c}^{-1}(\cdot)$ denotes the inverse function of $\text{vec}_{\Omega^c}(\cdot)$.

2.5.3 Numerical Evaluation

In this section, we perform numerical experiments on discrete-valued real-world data sets to evaluate the proposed discrete-aware MC algorithm. To this end, we adopt the MovieLens-100k data set[5] for recommender systems, one of the popular data sets utilized in MC literature for performance evaluations, which is composed of integer ratings (from 1 to 5) associated with many different user-movie pairs and possesses a low-rank nature due to the inter-user correlation in preferred movies. To evaluate the robustness jointly with the recovery performance, we vary the observed ratio from 20% to 60%, while normalized-MSE (NMSE) is utilized as the performance metric, which is given by

$$\text{NMSE} \triangleq \frac{\|P_{\Omega^c}\left(\mathbf{X} - \mathbf{O}\right)\|_F^2}{\|P_{\Omega^c}\left(\mathbf{O}\right)\|_F^2}. \tag{2.84}$$

Besides the Soft-Impute algorithm [89], we compare our proposed algorithm with other state-of-the-art methods such as AIS-Impute [85], an accelerated variate of Soft-Impute, niAPG [88], a non-convex variate of Soft-Impute with the log-sum-penalty (LSP) non-convex regularizer.

The NMSE performance results comparing our proposed discrete-aware variates of Soft-Impute and the aforementioned state-of-the-art LRMC algorithms as a function of ratio of observed ratings are shown in Figure 2.10, where the red lines correspond to our proposed methods and black or gray lines are associated with the state-of-the-arts. For the sake of clarity, the NMSE performance gaps due to discreteness-awareness is

[5] https://grouplens.org/datasets/movielens/

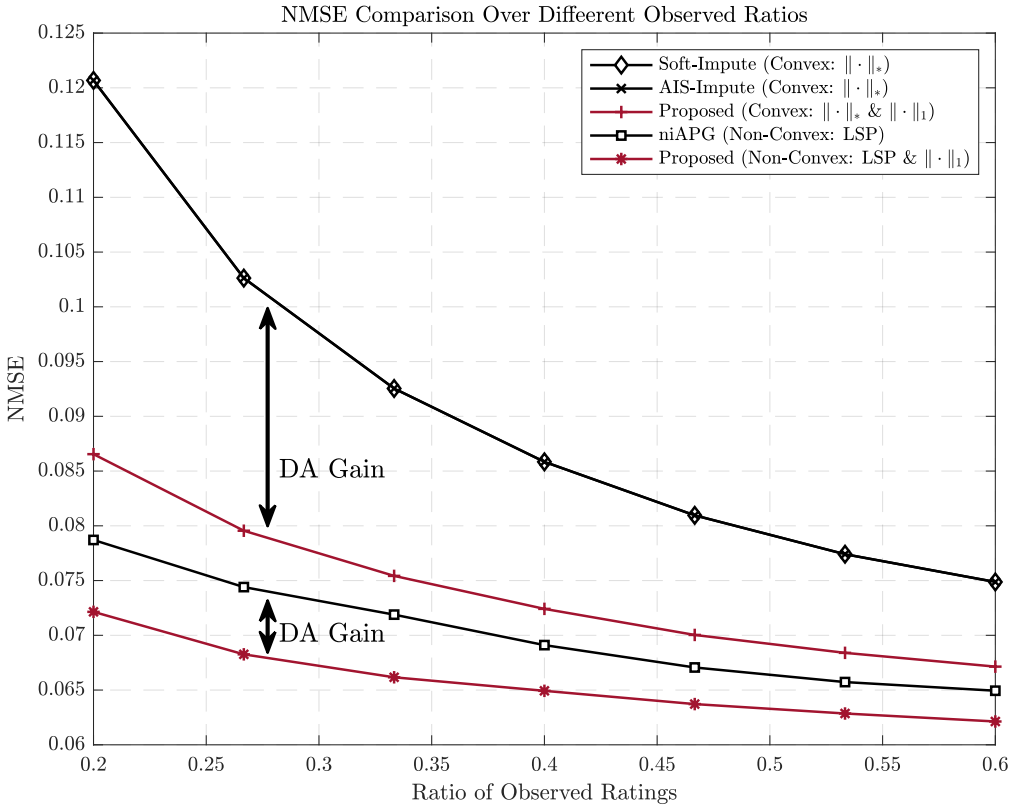


Figure 2.10: NMSE performance evaluations of the proposed discrete-aware MC algorithms (red) and other state-of-the-art methods (gray and black) with respect to different observation ratios. ©2020 IEEE

highlighted by annotation arrows. It can be observed from the figure that most of the algorithms are able to successfully achieve less than 0.1 in terms of NMSE for a wide range of observed ratios, albeit non-convex algorithms with LSP can reduce the performance degradation in a severe scenario, where only a few number of entries of the matrix can be observed. More interestingly, even in case of convex algorithms, the discreteness-awareness considerably decreases increment of the NMSE curve at the low observed ratio range, which indicates the robustness of the proposed discrete-aware regularizer.

In Figure 2.11, the NMSE convergence behavior of the algorithms with respect to the number of algorithmic iterations is presented, where we can perceive that most of the algorithms converge within 100 iterations in case with the convex NN regularizer and 180 iterations in case with the non-convex LSP regularizer, respectively. Furthermore, the figure illustrates the accelerated convergence of the proposed algorithm with the convex NN regularizer. According to this observation, it may be concluded that the discreteness-awareness can improve the NMSE performance.

Besides the above, we remark that the additional complexity due to the discreteness-aware regularizer in equation (2.75) with $p = 1$ is linear with respect to the cardinality

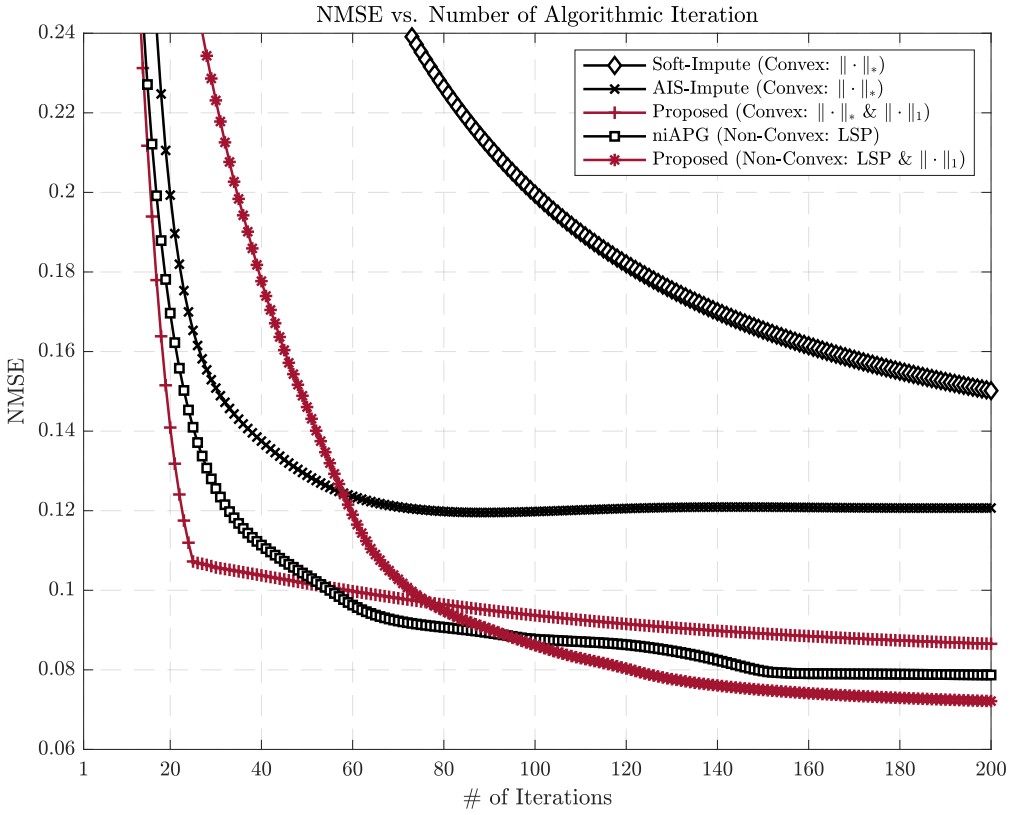


Figure 2.11: NMSE performance behavior on the MovieLens-100k data set as a function of algorithmic iterations with 20% observation of the total non-zero entries. ©2020 IEEE

of the unknown index set (*i.e.*, $|\Omega^c|$) as one may readily observe from the element-by-element operation in equations (2.81)–(2.83). Therefore, one may conclude that the most expensive part of the algorithm in terms of complexity is the same as that of the state-of-the-art methods, *i.e.*, SVT, indicating that the proposed algorithm maintains the same complexity order.

In case of $p = 0$, however, the regularizer may affect the convergence or the complexity of the proposed PG algorithm due to many different reasons such as expansiveness of $r_0(\mathbf{X})$ [96] or successive convex approximation to relax the ℓ_0-norm function. Taking into account the aforementioned issues, it is an open problem to develop a fully non-convex algorithm (*i.e.*, a non-convex low-rank regularizer and a discrete-aware regularizer with $p < 1$) and analyse its convergence property.

In light of all the above, we conclude from the numerical performance evaluations that our proposed discreteness-aware MC algorithm may further accelerate the convergence and improve the completion performance in case of adopting convex functions for both regularizers (*i.e.*, $g(\cdot) = \|\cdot\|_*$ and $r_1(\cdot)$), while enjoying the uniqueness of the solution due to the convexity of equation (2.74). In case of non-convex low-rank regularizer (*i.e.*, LSP) while maintaining convex discreteness-aware regularizer (*i.e.*, $r_1(\cdot)$),

it has been shown that at the expense of slower convergence, the NMSE performance can be enhanced as shown in Figure 2.11.

2.6 Conclusion

In this chapter, we proposed a flexible framework for the symbol detection problem in overloaded communication systems (*e.g.*, NOMA and overloaded MIMO) without assuming particular channel statistics. The key concept of the proposed framework is the compliance of its solution to the prescribed symbol constellation $\mathcal{C}$. Technically speaking, the proposed detection framework is based on a novel adaptable quadratic discreteness-aware regularizer, in which the alternative ML detection problem via the ℓ_0-norm is first approximated by an asymptotically-exact non-convex expression and later convexified by a FP technique. The proposed IDLS detector was then shown to be a generalization of the well-known ZF and LMMSE methods with adherence to the predetermined discrete constellation set such as phase shift keying (PSK) and QAM modulation. Thanks to the flexibility of the proposed framework in dealing with possible non-ideal conditions, the proposed IDLS detector has been further extended to Robust IDLS method so as to mitigate the harmful effect of CSI imperfection and hardware impairments while obeying the discreteness constraint of the symbols.

Also, it is worth-mentioning that since on principle the proposed IDLS receiver generalizes the classic ZF and LMMSE algorithms for an estimation problem subject to discrete sets, that part of the contribution is likely to also find applications in a variety of problems beyond communications. As an illustration of the aforementioned claim, we further developed an LRMC algorithm leveraging the concept of the discreteness-aware regulization for a matrix completion problem subject to discrete entries. The software simulation demonstrates the effectiveness of the incorporation of the discreteness-awareness concept in MC problems. Note that a further extension of the developed LRMC algorithm with the non-convex regularizer proposed in this section is left for future work.

Furthermore, the work on this subject considered in this section can be extended to other communication systems such as mmWave and full-duplex systems. As mentioned earlier, an IDD extension of the IDLS framework may be considered as a future work. Furthermore, as the proposed method can be seen as an M-estimator, its detection performance might also be precisely analyzed via the convex Gaussian min- max theorem (CGMT) under the assumption of i.i.d. Gaussian channels.

Chapter 3

Grant-Free Schemes for Low-Latency Massive Access in Distributed MIMO Systems

In this chapter, we intend to tackle the latency issue in distributed antenna wireless systems. In particular, we present two new grant-free access schemes towards low-latency massive communications in two distinct emerging distributed MIMO architectures, *i.e.*, CF-MIMO and XL-MIMO. The first algorithm developed for CF-MIMO architectures aims to improve the *per-user throughout* in grant-free systems while efficiently reducing the length of pilot sequences, which is a bottleneck of most of the existing related methods. To this end, we tailor the recently-introduced bilinear Bayesian inference framework to conquer this challenging task, leading to a novel joint estimation algorithm simultaneously addressing active user identification, channel estimation, and multiuser data detection. While the second algorithm developed for XL-MIMO also aims at the same overhead reduction task, the latter is designed to address the peculiarility of XL-MIMO systems rather than the per-user throughput bottleneck. In other words, the second algorithm addresses a joint detection problem of active user identification, channel estimation, and *active sub-array identification* (a.k.a. spational non-stationarity), which is enabled by properly customizing the bilinear inference framework. The effectiveness of the two algorithms is also demonstrated by quantitative simulation-based analyses.

Part of this chapter is reprinted and enhanced from the following publication:

- Hiroki Iimori, Takumi Takahashi, Koji Ishibashi, Giuseppe Thadeu Freitas de Abreu, and Wei Yu: "Grant-Free Access via Bilinear Inference for Cell-Free MIMO with Low-Coherence Pilots," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7694–7710, Nov. 2021.
- Hiroki Iimori, Takumi Takahashi, Koji Ishibashi, Giuseppe Thadeu Freitas de Abreu, David González G., and Osvaldo Gonsa: "Joint Activity and Channel Estimation for Extra-Large MIMO Systems," to appear in *IEEE Trans. Wireless Commun.*, 2022.

3.1 Background and Contributions

Multiple-antenna architectures, in particular massive MIMO and its extensions, will continue to be one of essential technologies in 5G and future 6G networks, in order to satisfy the ever-growing demand for higher data rates and user capacities as well as the heterogeneous requirements raised by mMTC, eMBB, URLLC and their various combinations. What makes massive MIMO technology a simultaneous enabler of high throughput communications and massive connectivity is the significant amount of the spatial degrees-of-freedoms (DoFs) it provides, which can be exploited to solve inherent problems such as multi-user detection (MUD), channel estimation (CE), and active user detection (AUD) in uplink scenarios, among others [97, 98].

Compared to conventional coherent MIMO communication mechanism, where AUD and CE are sequentially performed based on predetermined reference signals (*e.g.*, pilot sequences) followed by MUD relying on the estimated CSI, a main challenge of massive uplink access is the communication overhead required for CSI acquisition, which scales with the number of potential uplink users in the system due to the need of orthogonal pilot sequences so as to maintain accurate CSI knowledge. It is also worth mentioning that utilizing non-orthogonal pilot sequences for channel estimation, while contributing to reducing the overhead, leads to severe MUD performance deterioration due to the rank-deficient (*i.e.*, underdetermined) conditions typically faced, even under the assumption that perfect AUD is available at the receiver. In addition, in massive MIMO settings, excessive piloting might exceed channel coherence time, particular in the case of fast fading environments, which makes non-orthogonal pilots necessity.

A promising emerging approach to tackle this issue is joint channel and data estimation (JCDE) which takes advantage of estimated data symbols as soft pilot sequences, while exploiting their statistical quasi-orthogonality to improve system performance and efficiency. In particular, the bilinear generalized approximate message passing (BiGAMP) scheme proposed in [99] has been considered a key ingredient to solve such a detection problem in wireless systems. In that scheme, Onsager correction is employed to decouple the self-feedback of messages across iterations as is the case with the AMP, leading to stable convergence behavior as shown, for instance, in [100, 101].

It has, however, been recently shown in [102] that the estimation performance of BiGAMP severely deteriorates when non-orthogonal pilot sequences are exploited, even if adaptive damping is employed, because the derivation of BiGAMP relies heavily on the assumption of very large systems, although shortening the pilot sequence is the very aim of the method itself.

In order to circumvent this issue, the authors in [103] proposed a novel bilinear message passing algorithm, referred to as bilinear Gaussian belief propagation (BiGaBP), with the aim of generalizing BiGAMP on the basis of belief propagation (BP) [104]

for robust recovery subject to non-orthogonal piloting. Despite the aforementioned progresses, many existing works including the ones mentioned above, focus only on joint CE and MUD while assuming that perfect AUD is available at the receiver.

One of the solutions to the AUD problem is grant-free random access [105, 106], which has been intensively investigated in the last few years and can be categorized as a variation of JACE or (if symbol detection is also integrated) of JACDE, with Bayesian receiver design components. In the context of Bayesian approaches, Bayesian JACE can be seen as a non-orthogonal pilot-based random access protocol in which active users simultaneously transmit their unique spreading signatures to their base stations (BSs), and the BS employs a massage passing algorithm – *e.g.* multiple measurement vector approximate message passing (MMV-AMP) – as receiver, with the aim of detecting user activity patterns and their corresponding channel responses, while taking advantage of the time-sparsity resulting from the activity patterns. As for Bayesian JACDE schemes, most of the existing works on that approach, *e.g.* [107–111], are extensions of the aforementioned Bayesian JACE in which spreading data sequences generated by multiplying data symbols with their unique spreading signatures are transmitted[1], while leveraging a similar receiver design as that of Bayesian JACE methods.

There is also another approach to AUD that takes advantage of the sample covariance matrix constructed from the large number of antennas at the receiver. This covariance-based method has also attracted attention due to its applicability to unsourced random access (URA), where JACDE can be achieved by letting active users transmit a codeword sequence selected from a common predetermined codebook over a given time slot. To elaborate, it has been shown in [112] that the covariance-based approach is able to accommodate a larger number of active users, while using limited per-user wireless resources due to the nature of index-type modulation based on spread codewords, which is therefore suited to super low-rate mMTC scenarios.

Compounding on the above, a fundamental challenge of grant-free approaches from a system (infrastructure) viewpoint is the spatial correlation of the co-located massive MIMO channel, which has been argued in [113] to be a limiting factor of centralized massive MIMO, although most of the existing work in the area, including *e.g.* [102, 103,107,108,114–116], make use of the assumption that channels are subjected to ideal (uncorrelated) Rayleigh fading.

Aiming at addressing these challenges, new concepts of distributed massive MIMO have been recently proposed, among which are CF-MIMO [117] and XL-MIMO systems [118]. In a cell-free massive MIMO system, which can be seen as an instance of spatially distributed MIMO concept, a large number of APs geographically scattered over a certain service area and connected via fronthaul links to a common central processing

[1] please refer to the system model given in *e.g.* [107–111] for more technical details

unit (CPU), simultaneously serve multiple user equipments (UEs). Thanks to its spatial diversity, CF-MIMO are inherently robust to spatial correlation, but pay to that end the price of requiring long cabling infrastructure, high-capacity fronthaul links, compression techniques, or all of them [119]. The original idea of CF-MIMO shown in [117] has been further investigated and extended in the literature. Comprehensive analyses of CF-MIMO networks under different degrees of cooperation among the APs have been studied in [69], while the authors in [120] addressed the scalability issue of CF-MIMO by proposing the dynamic cooperation clustering (DCC) method, which provides a reasonable compromise between fully-distributed and fully-centralized scenarios. In addition to the above, a CF-MIMO network with multiple CPUs has been considered in the literature, leaving an open research problem with how to distribute and control signal processing tasks among multiple CPUs under limited backhaul and fronthaul links.

In contrast, XL-MIMO systems can be said to follow the strategy of forming a "MIMO continuum", in which a vastly large number of antennas are directly integrated into the ambient, by embedding them on the walls and ceilings of buildings, stadiums, train stations and airports. In this context, several works have been proposed. To mention a few, a user-grouping based approximated ZF precoding design was proposed for XL-MIMO beamforming in [121], which was shown to offer reasonable performance-complexity tradeoff. An EP-based multiuser data detection mechanism for XL-MIMO systems was proposed in [122], while an array selection method for higher energy efficient communications in XL-MIMO settings was proposed in [123]. From an access viewpoint, a grant-based random access strategy for XL-MIMO systems was considered in [124] assuming perfect knowledge of active user indices. A theoretical interpretation of achievable throughput in uplink XL-MIMO systems was offered in [125].

Since XL-MIMO systems rely on the use of large-aperture sub-arrays employed on a wide-ranging surface, XL-MIMO systems have to cope with its peculiarilities including spatial non-stationarity [126], *i.e.*, the fact that the signal from each user is apparent only to distributed portions of the XL-MIMO antenna array, referred to as its visibility region (VR). In other words, only a portion of the total antennas can observe the signal from each user. Several studies have been presented to address this spatial non-stationarity in XL-MIMO from different aspects. The CE problem in XL-MIMO subjected to non-stationarity was considered in [127, 128], while implicitly assuming a grant-based access protocol.

All in all, despite the importance of the antenna distribution, the grant-free access in a distributed MIMO architecture has less been addressed.

Contributions

Given the above, the contributions of this chapter are offered to address the latency issue in the aforementioned distributed MIMO architectures. In particular, they are summarized as follows.

- We employ a signal model incorporating a frame-theoretic non-orthogonal pilot design, whose mutual coherence approach Welch's theoretical coherence lower bound even for relatively *short pilot sequences*. This is in contrast to much of existing literature, which relies on Gaussian pilot designs, exhibiting poor mutual coherence properties, except in the asymptotic (large-system) regime.

In the Context of CF-MIMO

- We demonstrate the feasibility of grant-free JACDE *without spreading data sequences* drawn from a predetermined constellation as is the case for the conventional coherent MIMO systems. This is in contrast to most of related literature addressing AUD, CE, and MUD jointly in a grant-free fashion [107–110], in which the spreading of data sequences by pilot sequences is required, sacrificing spectral efficiency, limiting data rate per user, and increasing sensitivity to fading.
- We extend previous works such as [114,129,130] such that the proposed algorithm is directly applicable to a *cell-free architecture*, without sacrificing suitability also to centralized MIMO systems. We remark that a potential advantage of the cell-free architecture is that it helps resolve spatial correlation problems of massive MIMO. To the best of our knowledge, no grant-free design for cell-free massive MIMO scheme without pilot-based data spreading has been proposed yet.
- In order to enable the above, a *novel JACDE algorithm, dubbed activity-aware BiGaBP*, is presented here, in which bilinear inference, message passing rules, Gaussian approximation, and a new belief scaling technique that forges resilience of the derived messages, are combined.

In the Context of XL-MIMO

- Non-stationarity, user activity patterns, and channel fading jointly imposes a new estimation problem of random variables following a nested Bernoulli-Gaussian distribution, which is captured in the system model section of the corresponding section. To the best of our knowledge, this formulation appears for the first time in the literature.
- Owing to the nested nature of the variables of interest, the JACE problem is decoupled into a tractable bilinear inference problem.

- In order to solve the reformulated bilinear estimation problem, a novel message passing rule has been derived, in which estimates are obtained in closed-form. Based on the derived message passing rules, an iterative JACE algorithm is proposed for grant-free XL-MIMO systems subject to non-stationarity.
- In order to capture the cluster-like nature of sub-array activities, unlike the existing literature, we introduce a Matérn-cluster point process (MCPP)-based sub-array activity model, based on which the estimation performance of different approaches is compared.

3.2 Non-Orthogonal Pilot Design via Frame-Theory

In this section we review a frame-theoretic approach to effectively design a structured pilot matrix for non-orthogonal transmission [44, 131–133], aiming at efficiently reducing the pilot length while preserving the linear independence between vectors in the pilot matrix as much as possible. To this end, assuming a non-orthogonal representation of the pilot matrix, let us first define the mutual coherence as a measure of the similarity between non-orthogonal pilot sequences.

Definition 3.2.1 (Mutual Coherence)**.** *Let $\boldsymbol{F} \triangleq [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L] \in \mathbb{H}^{J\times L}$ be a* frame *matrix over a Hilbert space $\mathbb{H}^{J\times L}$, comprising of frame vectors $\boldsymbol{f}_\ell \in \mathbb{H}^{J\times 1}$, with $\ell \in \{1, \ldots, L\}$ and $J < L$. The mutual coherence of $\boldsymbol{F}$ is given by*

$$\mu(\boldsymbol{F}) \triangleq \max_{\ell\neq\ell'} \frac{|\langle \boldsymbol{f}_\ell, \boldsymbol{f}_{\ell'}\rangle|}{\|\boldsymbol{f}_\ell\|_2 \|\boldsymbol{f}_{\ell'}\|_2}, \forall\{\ell, \ell'\} \in \{1, 2 \ldots, L\}, \tag{3.1}$$

which in the case of an equal-norm frame, reduces to

$$\mu(\boldsymbol{F}) \triangleq \max_{\ell\neq\ell'} |\langle \boldsymbol{f}_\ell, \boldsymbol{f}_{\ell'}\rangle|, \forall\{\ell, \ell'\} \in \{1, 2 \ldots, L\}. \tag{3.2}$$

One readily notices from the above that the mutual coherence of a frame matrix $\boldsymbol{F}$ is equivalent to the maximum absolute value of the non-diagonal elements of the corresponding gram matrix $\boldsymbol{G} \triangleq \boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}$, which for $L \leq J^2$ is known to be lower-bounded by the Welch bound[2] [134]

$$\mu(\boldsymbol{F}) \geq \sqrt{\frac{L-J}{J(L-1)}}. \tag{3.3}$$

Besides low mutual coherence, for the sake of fairness – not in terms of throughputs but in terms of resource utilization [44] – that pilot sequences employed by different

[2]One may consider an extremely severe scenario $L > J^2$, although this is beyond the scope of this dissertation. In this case, the Welch bound is no longer a proper benchmark, implying that one needs another alternative that can be drawn from *e.g.*, [135].

users have the same energy, which in the context of frame designs translates to the following desired property.

Definition 3.2.2 (Tightness)**.** *By means of the Rayleigh-Ritz Theorem, a frame matrix* $\boldsymbol{F}$ *possesses the following inequalities.*

$$\alpha\|\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{F}^{\mathrm{H}}\boldsymbol{a}\|_2^2 \leq \beta\|\boldsymbol{a}\|_2^2, \forall \boldsymbol{a} \in \mathbb{H}^J, \tag{3.4}$$

where $0 < \alpha \leq \beta < \infty$ *and* $\boldsymbol{F}$ *is called* tight *if and only if (iff)* $\alpha = \beta$*.*

In light of the above, our goal is to design low-coherence tight frames as pilot sequences, which can be utilized even under severe non-orthogonal scenarios. It has been recently shown in [136] that a group-theoretic tight frame construction approach achieves near-Welch-bound performance, but unfortunately such cyclic-group approach is applicable only to particular cases in which the number of frame vectors (*i.e.*, L) is a prime number, while in the context hereby frames with arbitrary L and J are required.

We therefore consider instead a convex optimization based construction method recently proposed in [131, 132] and further developed in [44, 133]. Such a low-coherence unit-norm tight frame with arbitrary dimensions can be obtained by taking advantage of an iterative method, referred to as sequential iterative decorrelation via convex optimization (SIDCO) [131], whose extension to the complex space – dubbed as complex SIDCO (CSIDCO) [132] – has been studied, where the strategy to minimize the mutual coherence is to solve *iteratively* (*i.e.*, for different τ) the following convex optimization problem:

$$\min_{\substack{\boldsymbol{f}_\ell \in \mathbb{C}^{J\times 1} \\ \forall \ell \in \{1,2,\ldots,L\}}} \quad \|\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell\|_\infty \tag{3.5a}$$

$$\text{s.t.} \quad \|\boldsymbol{f}_\ell - \tilde{\boldsymbol{f}}_\ell\|_2^2 \leq T_\ell^{(\tau)}, \tag{3.5b}$$

where $\tilde{\boldsymbol{F}}_\ell \in \mathbb{C}^{J\times L-1}$ is obtained by prunning the ℓ-th column of $\tilde{\boldsymbol{F}}$, τ indicates the iteration index, and the search region is limited to a multidimensional Euclidean ball of radius

$$T_\ell^{(\tau)} = 1 - \max_{\ell'|\ell'\neq\ell} \frac{|\langle \boldsymbol{f}_\ell^{(\tau-1)}, \boldsymbol{f}_{\ell'}^{(\tau-1)} \rangle|^2}{\|\boldsymbol{f}_\ell^{(\tau-1)}\|_2^2 \|\boldsymbol{f}_{\ell'}^{(\tau-1)}\|_2^2}. \tag{3.6}$$

Although the CSIDCO reformulation given in equation (3.5) already follows the disciplined convex programming conventions, such that this problem can be easily solved via interior point methods through CVX available in high-level numerical computing programming languages such as MATLAB and Python, the abstraction penalty of such high-level programming languages are too high for real-world communication systems. Aiming at real-time processing of convex optimization problems, the authors in [137] have developed an automatic low-level code generator for conic programming

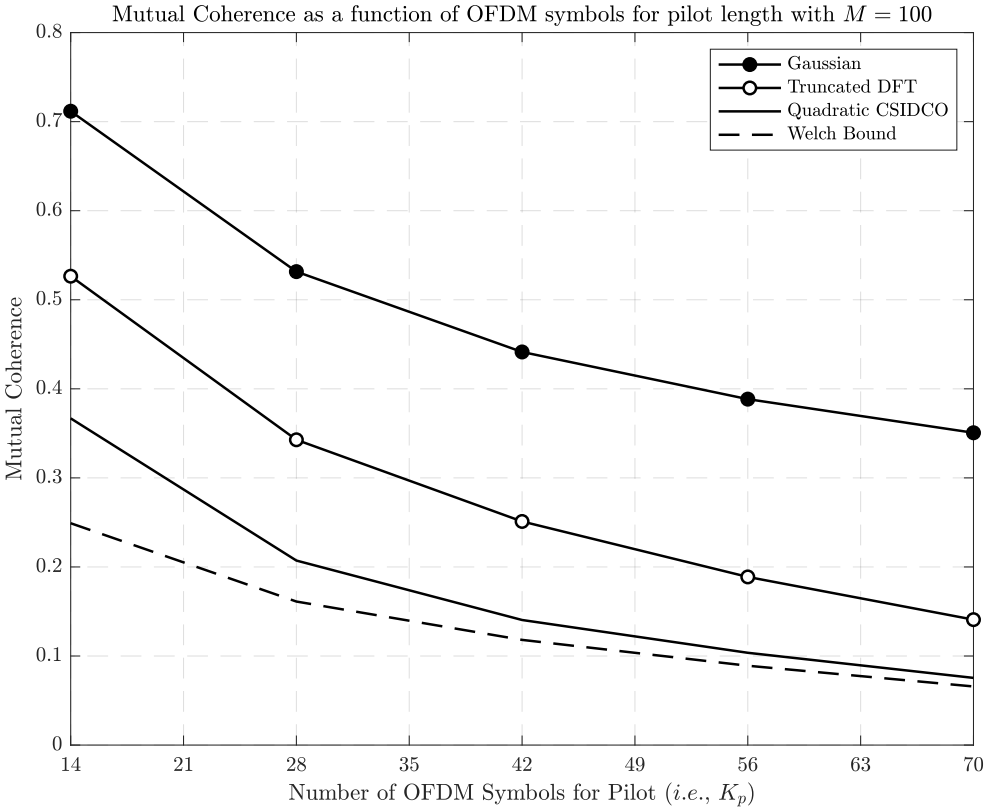


Figure 3.1: Mutual coherence comparison as a function of OFDM symbols utilized for pilot lengths K_p with $M = 100$ uplink users. As benchmarks, we adopt the popular random Gaussian pilot sequence considered in, *e.g.*, [47,115], and a randomly truncated discrete Fourier transform matrix. The average performance is shown for the Gaussian and truncated DFT approaches.

problems, which solves convex problems with moderate size on the order of microseconds or milliseconds, although it is limited to quadratic program (QP)-representable convex problems. In light of the above, for the sake of completeness, equation (3.5) was transformed in [44, 133] into the following quadratic program.

Theorem 3.2.1 (Quadratic CSIDCO)**.** *Introducing* $\boldsymbol{x}_\ell \triangleq [\Re\{\boldsymbol{f}_\ell\}; \Im\{\boldsymbol{f}_\ell\}; t_{\ell,R}; t_{\ell,I}] \in \mathbb{R}^{2J+2\times 1}$ *with slack variables* $t_{\ell,R} \in \mathbb{R}_+$ *and* $t_{\ell,I} \in \mathbb{R}_+$ *for all* ℓ*, the CSIDCO formulation given in equation* (3.5) *for a unit-norm low-coherence frame construction can be rewritten as*

$$\begin{aligned}
\min_{\substack{\boldsymbol{x}_\ell \\ \forall \ell \in \{1,2,\ldots,L\}}} \quad & \boldsymbol{x}_\ell^{\mathrm{T}} \boldsymbol{\Phi} \boldsymbol{x}_\ell && \text{(3.7a)} \\
\text{s.t.} \quad & \boldsymbol{A}_{\ell,R,1} \boldsymbol{x}_\ell \leq \boldsymbol{0}, && \text{(3.7b)} \\
& \boldsymbol{A}_{\ell,R,2} \boldsymbol{x}_\ell \leq \boldsymbol{0}, && \text{(3.7c)} \\
& \boldsymbol{A}_{\ell,I,1} \boldsymbol{x}_\ell \leq \boldsymbol{0}, && \text{(3.7d)} \\
& \boldsymbol{A}_{\ell,I,2} \boldsymbol{x}_\ell \leq \boldsymbol{0} && \text{(3.7e)} \\
& \boldsymbol{x}_\ell^{\mathrm{T}} \boldsymbol{\Xi} \boldsymbol{x}_\ell - 2\boldsymbol{b}_\ell^{\mathrm{T}} \boldsymbol{x}_\ell + 1 - T_\ell \leq 0, && \text{(3.7f)}
\end{aligned}$$

where $\boldsymbol{\Phi} \triangleq \left[\begin{smallmatrix} \mathbf{0}_{2J} & \mathbf{0}_{2J\times 2} \\ \mathbf{0}_{2\times 2J} & \mathbf{I}_2 \end{smallmatrix}\right]$, $\boldsymbol{\Xi} \triangleq \left[\begin{smallmatrix} \mathbf{I}_{2J} & \mathbf{0}_{2J\times 2} \\ \mathbf{0}_{2\times 2J} & \mathbf{0}_{2\times 2} \end{smallmatrix}\right]$, $\boldsymbol{b}_\ell \triangleq \left[\tilde{\boldsymbol{f}}_\ell^{\mathrm{T}} \;\; 0 \;\; 0\right]^{\mathrm{T}}$, *the other coefficient matrices are listed in Appendix B, and the iteration index* τ *is omitted for brevity.*

Proof See Appendix B □

One may argue that the frame matrix constructed via the quadratic CSIDCO method described above is not strictly tight due to the fact that tightness is not enforced during the optimization process. However, the tightening approach proposed in [138] based on the polar decomposition can be applied to the output of the quadratic CSIDCO. Consequently, one can obtain an arbitrarily-sized low-coherent equal-norm tight frame sufficiently close to the ideal equiangular tight frames which are not suited to practice as they exist only for particular dimensions. Owing to this flexibility and sufficiently high-performance, this approach is therefore adopted as pilot sequences considered in this section.

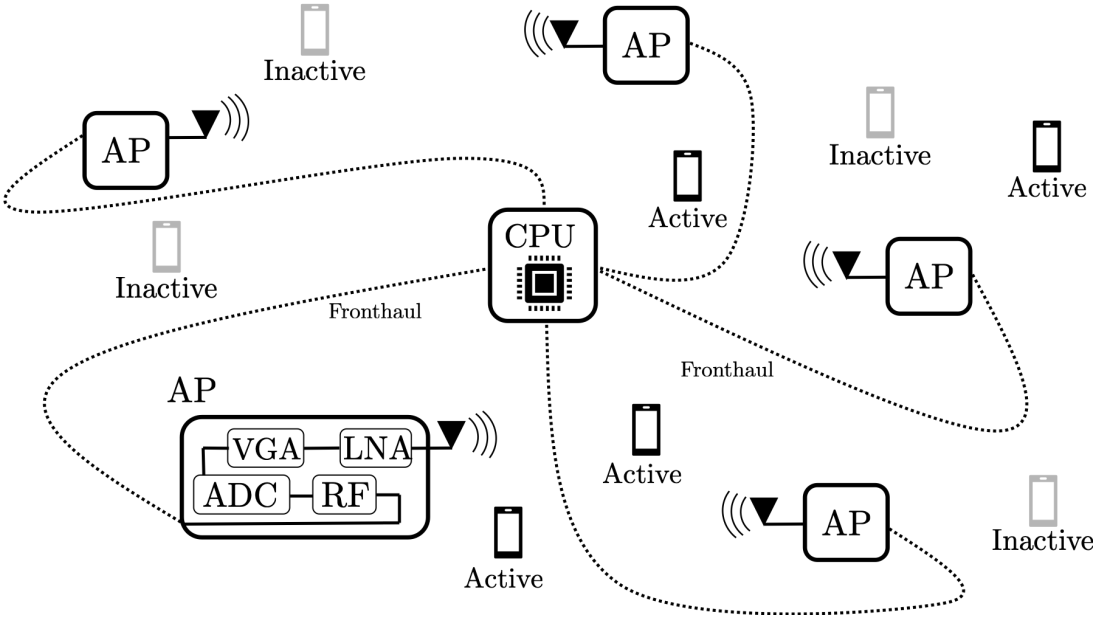
It is also worth-noting that an alternative to design a low-coherence pilot matrix is to leverage Grassmanian packing techniques [135, 139–141]. For instance, a pilot matrix design based on a particular structure such as one proposed in [135] may bring beneficial aspects in practice due to its low-complexity nature. However, optimizing such a pilot design in terms of performance and/or complexity is beyond the scope of this thesis, since the focus of this section is not to propose a new pilot design but to bring to light that such alternative exists and is suitable for pilot matrix design in grant-free access schemes.

To illustrate the performance of the quadratic CSIDCO method described above especially in 5G NR setups, mutual coherence comparisons of the method against popular pilot construction approaches (*i.e.*, the random Gaussian and truncated discrete Fourier transform matrices) as a function of the number of OFDM symbols leveraged for pilot lengths are shown in Figure 3.1, which demonstrates that in fact the quadratic CSIDCO approaches the Welch bound while reducing the correlation between pilot sequences in comparison with the other two approaches, implying capability of sufficiently mitigating the inter-user interference even in highly non-orthogonal scenarios and efficiently decreasing the number of pilot lengths simultaneously.

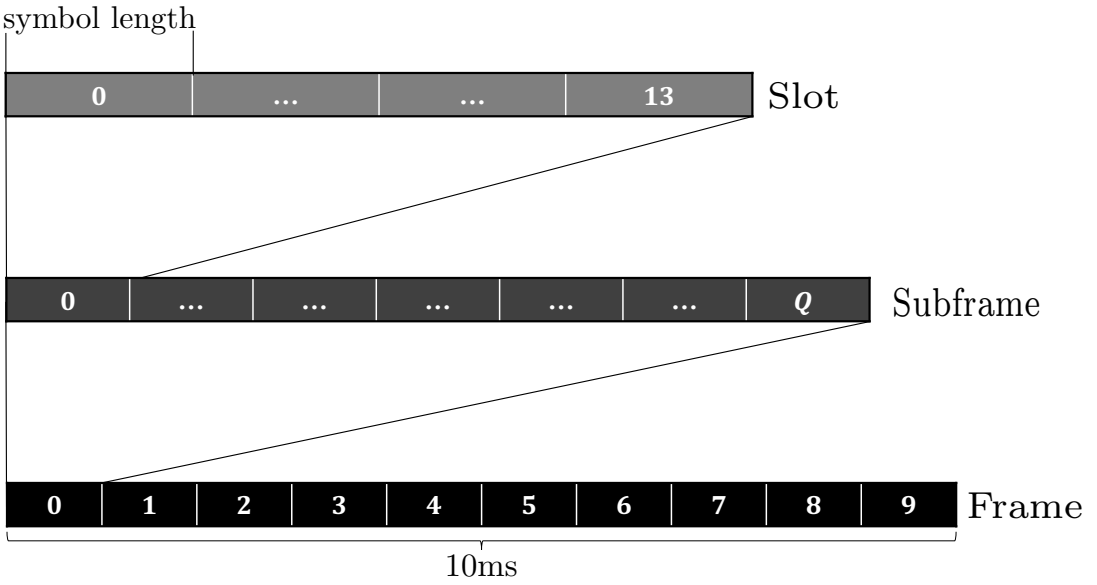
3.3 CF-MIMO

3.3.1 System Model

In this section, we consider a cell-free large MIMO system composed of N spatially distributed single-antenna APs connected by wired fronthaul links to a common high-performance CPU, serving M synchronized single-antenna users in a grant-free fashion, as depicted in Figure 3.2a. Due to the dynamic nature of grant-free systems, it is assumed that a fraction of the total M single-antenna users become active within a



(a) A model illustration of cell-free MIMO systems with distributed single-antenna APs serving uplink users which access the system in a grant-free basis.



(b) Radio frame structure in 5G NR [142].

Figure 3.2: System and Signal Model.

given 5G NR OFDM symbol frame [142], whereas the rest of uplink users remain silent during that period. As shown in Figure 3.2b, in the 5G NR signaling structure, each frame consists of 10 subframes and each subframe consists of Q slots. Furthermore, each slot is composed of 14 symbols. Taking advantage of this signaling design, one may utilize a part of the radio frame structure as reference signals and the rest as data streams. We also remark that the number of slots within a certain subframe (*i.e.*, Q) depends on the subcarrier spacing employed in a system.

Given the above, let K be the total number of discrete time indices within an OFDM frame and $\mathcal{C} \triangleq \{c_1, c_2, \ldots, c_{2^b}\}$ represent a given constellation of symbols with b denoting the number of bits per symbol. Introducing the user index set $\mathcal{M} \triangleq \{1, 2, \ldots, M\}$ and a set of active users $\mathcal{A}$, the received signal vector $\boldsymbol{y}_k \in \mathbb{C}^{N\times 1}$ at the k-th time index with $k \in \{1, 2, \ldots, K\}$ can be written as

$$\boldsymbol{y}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{w}_k, \tag{3.8}$$

where $\boldsymbol{x}_k \in \mathbb{C}^{M\times 1}$ denotes a possibly sparse transmitted signal vector, such that only its elements of indices $a \in \mathcal{A}$ are non-zero; $\boldsymbol{w}_k \in \mathbb{C}^{N\times 1}$ is a circularly symmetric zero-mean i.i.d. AWGN, *i.e.*, $\boldsymbol{w}_k \sim \mathcal{CN}_N(\boldsymbol{0}, N_0\mathbf{I}_N)$; and $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ denotes a flat block fading communication channel matrix assumed to be constant during K successive transmissions.

Assuming that the user activity remains consistent within an OFDM frame, we can concatenate the K consecutive symbol transmissions into the transmitted signal matrix $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K]$. In turn, the row sparse nature of $\boldsymbol{X}$ can be translated to a column-sparsity in the channel matrix $\boldsymbol{H}$, such that the m-th column $\boldsymbol{h}_m$ of the channel matrix can be modeled as a multivariate Bernoulli-Gaussian random variable, that is

$$\boldsymbol{h}_m \sim (1-\lambda)\delta(\boldsymbol{h}_m) + \lambda\mathcal{CN}_N(\boldsymbol{0}, \boldsymbol{\Gamma}_m), \tag{3.9}$$

where λ is the activity factor, $\delta(\boldsymbol{h}_m)$ denotes the Dirac delta function that takes the value 0 everywhere except at $\boldsymbol{h}_m = \boldsymbol{0}$ where it takes the value 1, and $\boldsymbol{\Gamma}_m$ is the covariance matrix of the channel.

In light of the above, the received signal matrix $\boldsymbol{Y} \in \mathbb{C}^{N\times K}$ concatenating the received signal vectors given in equation (3.8) over K successive time indices, can be readily expressed as

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W}, \tag{3.10}$$

where $\boldsymbol{Y} \triangleq [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K]$ and $\boldsymbol{W} \triangleq [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]$.

Let us employ the subscripts $\cdot_p$ and $\cdot_d$ to indicate pilot and data signals, respectively, and in order to explicitly express the pilot and data sequences within the transmitted and received signal, let us define, without loss of generality,

$$\boldsymbol{Y} \triangleq [\boldsymbol{Y}_p, \boldsymbol{Y}_d] \quad \text{and} \quad \boldsymbol{X} \triangleq [\boldsymbol{X}_p, \boldsymbol{X}_d], \tag{3.11}$$

where $\boldsymbol{Y}_p \in \mathbb{C}^{N\times K_p}$, $\boldsymbol{Y}_d \in \mathbb{C}^{N\times K_d}$, $\boldsymbol{X}_p \in \mathbb{C}^{M\times K_p}$ and $\boldsymbol{X}_d \in \mathbb{C}^{M\times K_d}$, with $K_p + K_d = K$ and $K_p \ll K_d < K$; and where each element of $\boldsymbol{X}_d$ is assumed to be drawn from the constellation set $\mathcal{C}$ in a similar way to conventional coherent MIMO systems.

At this point, we stress that although cell-free systems are our primary interest in this section, the signal model given in equation (3.10), and consequently the

proposed JACDE method presented later, are not limited to cell-free architectures. For instance, a conventional MIMO system model follows directly by setting $\mathbf{\Gamma}_m \triangleq \text{diag}([\gamma_{1m}, \ldots, \gamma_{Nm}])$, with $\gamma_m = \gamma_{1m} = \ldots = \gamma_{Nm}$, where γ_m denotes the power of the channel from the m-th user. In contrast, in order to model a cell-free MIMO system the receive channel powers γ_{nm} are simply allowed to be different for each link between the n-th AP and m-th user. Having said that, in centralized MIMO systems the spatial correlation needs to be involved in the channel model, for which the algorithm may need some technical adjustments to properly handle such correlation, although most of the message passing rules remain as it is.

We remark that unlike most grant-free systems found in literature, *e.g.* [107–109, 143–145], where data symbols are transmitted after spreading by pilot sequences, our approach can be seen, in light of equation (3.10) as a cell-free-implementable activity-aware variant of conventional MIMO systems, having the potential to enable both grant-free random access and pilot length reduction by jointly performing user activity, channel state, and data detection. These features make our approach better suited to meet the demand of higher spectrum efficiency per user than grant-free methods previously proposed, including those aforementioned.

3.3.2 Proposed Joint Estimation Method

In this section, we describe a joint activity, channel, and data estimation mechanism via bilinear Gaussian belief propagation (GaBP) for large cell-free MIMO architectures, whose belief propagation can be modeled as the graph schematized in Figure 3.3.

In order to further clarify the process of the proposed method, a work flow chart of the proposed detection mechanism is also illustrated in Figure 3.4, where the detection procedure is split into two algorithms, the belief consensus and hard decision blocks. As shown in the figure, the work flow starts with an initialization where a first guess of the channel estimate is obtained using only the pilot sequences, the result of which is however limited in accuracy due to the severely non-orthogonal structure of the pilot matrix as $K_p \ll M$. Aiming at improving the channel estimation accuracy by jointly detecting the data as well as the activity, the initial channel guess obtained by the initialization process is fed to the belief consensus block, which as illustrated in Figure 3.3, is composed of two different stages; 1) soft interference cancellation (SIC) and beliefs generation based on tentative soft estimates and 2) combining beliefs and soft estimates generation, described in the next two subsections, respectively.

Followed by the belief consensus block, we proceed with the subsequent hard decision block , where the final hard decision of the data and activity is carried out, which will be described later.

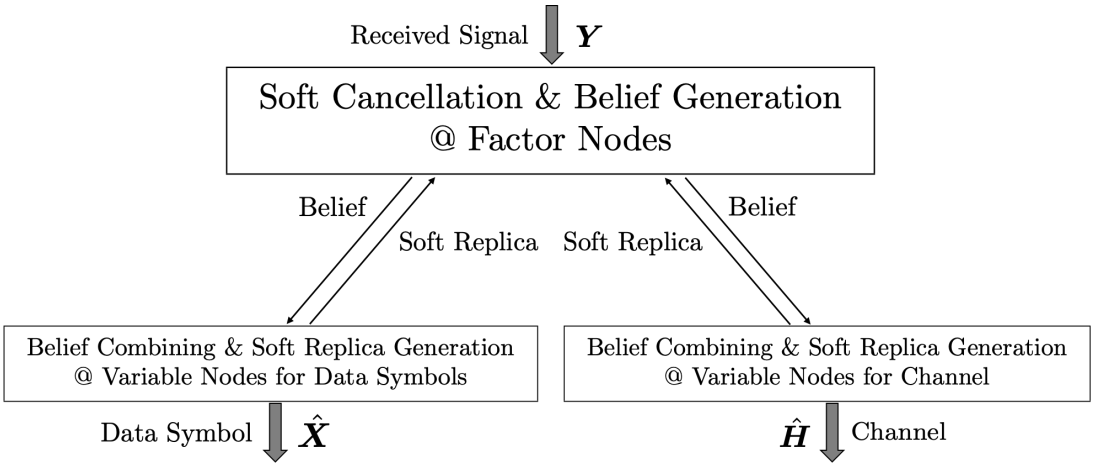


Figure 3.3: Belief generation and combining model.

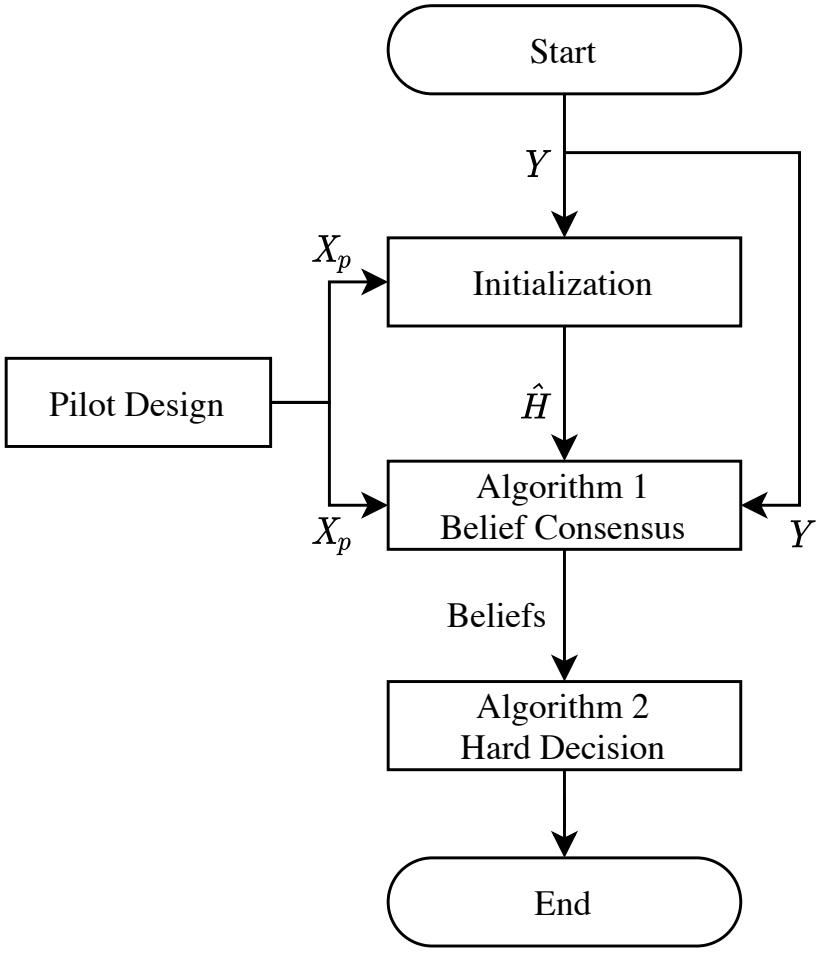


Figure 3.4: Work flow of the proposed detection process.

Factor nodes

Focusing on the received signal element y_{nk} at the n-th row and k-th column of $\boldsymbol{Y}$ with the aim of detecting x_{mk} at the m-th row and k-th column of $\boldsymbol{X}$, the received signal after SIC using soft estimates is given by

$$\tilde{y}_{m,nk} \triangleq y_{nk} - \overbrace{\sum_{i\neq m}^{M} \hat{h}_{k,ni}\hat{x}_{n,ik}}^{\text{Inter-user interference cancellation with soft-replicas}} \tag{3.12a}$$

$$= h_{nm}x_{mk} + \underbrace{\sum_{i\neq m}^{M}(h_{ni}x_{ik} - \hat{h}_{k,ni}\hat{x}_{n,ik})}_{\text{Residual interference}} + w_{nk}, \tag{3.12b}$$

where $\hat{x}_{n,ik}$ and $\hat{h}_{k,ni}$ with $i \in \{1, 2, \ldots, M\}$ are tentative estimates of x_{ik} and h_{ni}, respectively, generated at variable nodes in the previous iteration, and w_{nk} is the noise element at the n-th row and k-th column of $\boldsymbol{W}$.

Owing to the central limit theorem, the interference-plus-noise component can be approximated as a complex Gaussian random variable under large-system conditions, resulting in the fact that the conditional probability density function (PDF) of equation (3.12) for given x_{mk} and h_{nm} can be respectively expressed as

$$p_{\tilde{y}_{m,nk}|x_{mk}}(\tilde{y}_{m,nk}|x_{mk}) \propto e^{-\frac{|\tilde{y}_{m,nk} - \hat{h}_{k,nm}x_{mk}|^2}{v^x_{m,nk}}}, \tag{3.13a}$$

$$p_{\tilde{y}_{m,nk}|h_{nm}}(\tilde{y}_{m,nk}|h_{nm}) \propto e^{-\frac{|\tilde{y}_{m,nk} - h_{nm}\hat{x}_{n,mk}|^2}{v^h_{m,nk}}}, \tag{3.13b}$$

with

$$v^x_{m,nk} \triangleq \sum_{i\neq m}^{M} \left\{ |\hat{h}_{k,ni}|^2\psi^x_{n,ik} + (|\hat{x}_{n,ik}|^2 + \psi^x_{n,ik})\psi^h_{k,ni} \right\} + \psi^h_{k,nm} + N_0 \tag{3.14a}$$

$$v^h_{m,nk} \triangleq \sum_{i\neq m}^{M} \left\{ |\hat{h}_{k,ni}|^2\psi^x_{n,ik} + (|\hat{x}_{n,ik}|^2 + \psi^x_{n,ik})\psi^h_{k,ni} \right\} + \gamma_{nm}\psi^x_{n,mk} + N_0, \tag{3.14b}$$

where $\psi^x_{n,ik}$ and $\psi^h_{k,ni}$ denote expected error variances corresponding to $\hat{x}_{n,ik}$ and $\hat{h}_{k,ni}$, respectively, and we remark that $\mathbb{E}\big[|x_{mk}|^2\big] = 1$.

Variable nodes

Given the SIC and belief generation above, one can combine the Gaussian beliefs given in equation (3.13a), yielding the PDF of an extrinsic belief $l^x_{n,mk}$ for x_{mk}

$$p_{l^x_{n,mk}|x_{mk}}(l^x_{n,mk}|x_{mk}) = \prod_{i\neq n}^{N} p_{\tilde{y}_{m,ik}|x_{mk}}(\tilde{y}_{m,ik}|x_{mk}) \propto e^{-\frac{|x_{mk} - \hat{r}_{n,mk}|^2}{\psi^r_{n,mk}}}, \tag{3.15}$$

with

$$\psi^r_{n,mk} \triangleq \left(\sum_{i \neq n}^{N} \frac{|\hat{h}_{k,im}|^2}{v^x_{m,ik}} \right)^{-1} \tag{3.16a}$$

$$\hat{r}_{n,mk} \triangleq \psi^r_{n,mk} \sum_{i \neq n}^{N} \frac{\hat{h}^*_{k,im} \tilde{y}_{m,ik}}{v^x_{m,ik}}. \tag{3.16b}$$

In turn, since the activity can be expressed as column-sparsity in the channel matrix $\boldsymbol{H}$, combining beliefs of the m-th column of the channel matrix (*i.e.*, $\boldsymbol{h}_m$) needs to be jointly performed over $n \in \{1, 2, \ldots, N\}$. Thus, the PDF of the extrinsic belief $\boldsymbol{l}^h_{k,m}$ for given $\boldsymbol{h}_m$ is given by

$$p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m) = \prod_{n=1}^{N} \prod_{i \neq n}^{K} p_{\tilde{y}_{m,ni}|h_{nm}}(\tilde{y}_{m,ni}|h_{nm}) \tag{3.17a}$$

$$\propto e^{-\left(\boldsymbol{h}_m - \boldsymbol{\mu}^h_{k,m}\right)^{\mathrm{H}} \boldsymbol{\Sigma}^{h^{-1}}_{k,m} \left(\boldsymbol{h}_m - \boldsymbol{\mu}^h_{k,m}\right)},$$

$$\propto \frac{e^{-\left(\boldsymbol{h}_m - \boldsymbol{\mu}^h_{k,m}\right)^{\mathrm{H}} \boldsymbol{\Sigma}^{h^{-1}}_{k,m} \left(\boldsymbol{h}_m - \boldsymbol{\mu}^h_{k,m}\right)}}{\pi^N |\boldsymbol{\Sigma}^h_{k,m}|}. \tag{3.17b}$$

Notice that the expression in equation (3.17b) is a complex multi-variate Gaussian PDF, that is

$$p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m) \propto \underbrace{\mathcal{CN}_N\left(\boldsymbol{\mu}^h_{k,m}, \boldsymbol{\Sigma}^h_{k,m}\right)}_{N\text{-multivariate complex Gaussian distribution}}, \tag{3.18}$$

where

$$\boldsymbol{\mu}^h_{k,m} \triangleq [\hat{q}_{k,1m}, \ldots, \hat{q}_{k,Nm}]^{\mathrm{T}} \tag{3.19a}$$

$$\boldsymbol{\Sigma}^h_{k,m} \triangleq \mathrm{diag}\big(\psi^q_{k,1m}, \ldots, \psi^q_{k,Nm}\big), \tag{3.19b}$$

with

$$\psi^q_{k,nm} \triangleq \left(\sum_{i \neq k}^{K} \frac{|\hat{x}_{n,mi}|^2}{v^h_{m,ni}} \right)^{-1} \tag{3.20a}$$

$$\hat{q}_{k,nm} \triangleq \psi^q_{k,nm} \sum_{i \neq k}^{K} \frac{\hat{x}^*_{n,mi} \tilde{y}_{m,ni}}{v^h_{m,ni}}. \tag{3.20b}$$

Following the instruction given in Section 1, soft estimates of x_{mk} and $\boldsymbol{h}_m$ can be obtained by taking the expectation over the PDFs of the extrinsic beliefs given in

equations (3.15) and (3.18), respectively, as

$$\hat{x}_{n,mk} = \sum_{x_q \in \mathcal{C}} \frac{x_q \cdot p_{l^x_{n,mk}|x_{mk}}(l^x_{n,mk}|x_q) p_{x_{mk}}(x_q)}{\sum_{x'_q \in \mathcal{C}} p_{l^x_{n,mk}|x_{mk}}(l^x_{n,mk}|x'_q) p_{x_{mk}}(x'_q)}, \tag{3.21a}$$

$$\hat{\boldsymbol{h}}_{k,m} = \int_{\boldsymbol{h}_m} \boldsymbol{h}_m \frac{p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m) p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)}{\int_{\boldsymbol{h}'_m} p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m) p_{\boldsymbol{h}_m}(\boldsymbol{h}'_m)}, \tag{3.21b}$$

where the denominators are introduced to normalize the integral of the posterior PDFs to 1.

Although a closed-form expression of equation (3.21a) is not known for arbitrary discrete constellations, for Gray-coded QPSK[3] it can be written as [147]

$$\hat{x}_{n,mk} = \frac{1}{\sqrt{2}} \Big(\tanh\Big(\frac{\sqrt{2}\Re(\hat{r}_{n,mk})}{\psi^r_{n,mk}} \Big) + j \tanh\Big(\frac{\sqrt{2}\Im(\hat{r}_{n,mk})}{\psi^r_{n,mk}} \Big) \Big), \tag{3.22}$$

with the corresponding error variance estimate given by

$$\psi^x_{n,mk} = 1 - |\hat{x}_{n,mk}|^2. \tag{3.23}$$

A closed-form of equation (3.21b) can be obtained as follows. First, define the effective PDF

$$\begin{aligned} P^h_{k,m}(\boldsymbol{h}_m) &\triangleq p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m) p_{\boldsymbol{h}_m}(\boldsymbol{h}_m) \\ &= \frac{1}{\pi^N} \times \Bigg[\frac{\lambda e^{-{\boldsymbol{\mu}^h_{k,m}}^{\mathrm{H}} \left(\boldsymbol{\Sigma}^h_{k,m} + \boldsymbol{\Gamma}_m\right)^{-1} \boldsymbol{\mu}^h_{k,m}} \mathcal{CN}_N\left(\boldsymbol{\Gamma}_m \left(\boldsymbol{\Sigma}^h_{k,m} + \boldsymbol{\Gamma}_m\right)^{-1}, \boldsymbol{\Gamma}_m \left(\boldsymbol{\Sigma}^h_{k,m} + \boldsymbol{\Gamma}_m\right)^{-1} \boldsymbol{\Sigma}^h_{k,m}\right)}{|\boldsymbol{\Gamma}_m + \boldsymbol{\Sigma}^h_{k,m}|} \\ &\qquad + \frac{(1-\lambda)\delta(\boldsymbol{h}_m) e^{-{\boldsymbol{\mu}^h_{k,m}}^{\mathrm{H}} {\boldsymbol{\Sigma}^h_{k,m}}^{-1} \boldsymbol{\mu}^h_{k,m}}}{|\boldsymbol{\Sigma}^h_{k,m}|} \Bigg]. \end{aligned} \tag{3.24}$$

Then, the normalizing factor in the denominator of equation (3.21b) can be calculated by integrating the latter over the entire N-dimensional complex field, which yields

$$C^h_{k,m} \triangleq \int_{\boldsymbol{h}'_m} P^h_{k,m}(\boldsymbol{h}'_m) = \frac{\lambda \exp\Big(-{\boldsymbol{\mu}^h_{k,m}}^{\mathrm{H}} \Big(\boldsymbol{\Sigma}^h_{k,m} + \boldsymbol{\Gamma}_m \Big)^{-1} \boldsymbol{\mu}^h_{k,m} \Big)}{\pi^N |\boldsymbol{\Gamma}_m + \boldsymbol{\Sigma}^h_{k,m}|} \tau_{k,m}, \tag{3.25}$$

with the activity detection factor

$$\tau_{k,m} \triangleq 1 + \frac{1-\lambda}{\lambda} \exp\Big(-\big(\pi^h_{k,m} - \psi^h_{k,m} \big) \Big), \tag{3.26a}$$

[3]For higher-order modulations, by running the algorithm on the equivalent PAM real-valued model [146], the computational cost required to evaluate equations (3.21a) and its MSE grows with $\mathcal{O}\left(\sqrt{Q}NMK\right)$, where Q denotes the modulation order.

where

$$\pi_{k,m}^{h} \triangleq \boldsymbol{\mu}_{k,m}^{h\,\mathrm{H}}(\boldsymbol{\Sigma}_{k,m}^{h\,-1} - \left(\boldsymbol{\Sigma}_{k,m}^{h} + \boldsymbol{\Gamma}_{m}\right)^{-1})\boldsymbol{\mu}_{k,m}^{h} \tag{3.26b}$$

$$\psi_{k,m}^{h} \triangleq \log\left(\left|\boldsymbol{\Sigma}_{k,m}^{h\,-1}\boldsymbol{\Gamma}_{m} + \mathbf{I}_{N}\right|\right). \tag{3.26c}$$

With possession of equations (3.24) and (3.25), whose detailed derivations are given in Appendix C, the soft replica of $\boldsymbol{h}_m$ for a given effective distribution $P_{k,m}^{h}(\boldsymbol{h}_m)$ can be obtained as

$$\hat{\boldsymbol{h}}_{k,m} = \frac{\boldsymbol{\Gamma}_{m}\left(\boldsymbol{\Sigma}_{k,m}^{h} + \boldsymbol{\Gamma}_{m}\right)^{-1}}{\tau_{k,m}}\boldsymbol{\mu}_{k,m}^{h}, \tag{3.27}$$

and for the corresponding error variance, introducing $\boldsymbol{\Psi}_{k,m}^{h} \triangleq \mathrm{diag}\left(\psi_{k,1m}^{h}, \ldots, \psi_{k,Nm}^{h}\right)$ yields

$$\begin{aligned}\boldsymbol{\Psi}_{k,m}^{h} &= \mathrm{diag}\left(\int_{\boldsymbol{h}_m} \boldsymbol{h}_m\boldsymbol{h}_m^{\mathrm{H}}\frac{P_{k,m}^{h}(\boldsymbol{h}_m)}{C_{k,m}} - \hat{\boldsymbol{h}}_{k,m}\hat{\boldsymbol{h}}_{k,m}^{\mathrm{H}}\right) \\ &= (\tau_{k,m}-1)\mathrm{diag}(\hat{\boldsymbol{h}}_{k,m}\hat{\boldsymbol{h}}_{k,m}^{\mathrm{H}}) + \frac{\boldsymbol{\Sigma}_{k,m}^{h}\boldsymbol{\Gamma}_{m}(\boldsymbol{\Sigma}_{k,m}^{h} + \boldsymbol{\Gamma}_{m})^{-1}}{\tau_{k,m}}.\end{aligned} \tag{3.28}$$

3.3.3 Algorithm Description

In this subsection we summarize the belief propagation and consensus mechanisms described above and schematized in Figure 3.4, offering also detailed discussion on the algorithmic flow. For starters, the schemes are concisely described in the form of pseudo-codes in Algorithm 3 and Algorithm 4, respectively. For the sake of brevity, we define two sets of integers (*i.e.*, $\mathcal{K}_p$ and $\mathcal{K}_d$), which are respectively given by $\mathcal{K}_p \triangleq \{1, 2\ldots, K_p\}$ and $\mathcal{K}_d \triangleq \{K_p+1, K_p+2\ldots, K\}$.

As shown in the pseudo-codes, Algorithm 3 requires five different inputs: the received signal matrix $\boldsymbol{Y}$, the pilot sequence $\boldsymbol{X}_p$, an initial guess of the channel matrix $\hat{\boldsymbol{H}}$, an initial guess of the estimation error variance $\hat{\boldsymbol{\Psi}}^{h}$ corresponding to $\hat{\boldsymbol{H}}$, and the maximum number of iterations $t_{\max}$; while Algorithm 4 is fed with the beliefs obtained from Algorithm 3. We point out that rough estimates $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}^{h}$ can be obtained via state-of-the-art algorithms proposed for grant-free systems [47, 114, 115], although such mechanisms suffer from estimation inaccuracy in case of severely non-orthogonal pilot sequence ($K_p \ll M$). We emphasize again that reducing the overhead is however desired from a system-level perspective in terms of time resource efficiency. The initialization process adopted in this section will be discussed in detail later.

In turn, the outputs of Algorithm 4 are the following three quantities: an estimated symbol matrix $\hat{\boldsymbol{X}}$, an estimated channel matrix $\hat{\boldsymbol{H}}$, and estimates of active-user indexes $\hat{\mathcal{A}}$. As for $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{H}}$, they are obtained by combining all the beliefs (*i.e.*, consensus), while $\hat{\mathcal{A}}$ is determined by following a certain activity detection policy based on the

Algorithm 3 Bilinear GaBP (Part 1: Belief Consensus)

Inputs: $\boldsymbol{Y}$, $\boldsymbol{X}_p$, $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{\Psi}}^h$, λ, $t_{\max}$
Outputs: $\forall k \in \mathcal{K}_d$, $\forall m$, $\forall n$: $\hat{x}_{n,mk}$, $\psi^x_{n,mk}$, $\forall k$, $\forall m$: $\hat{\boldsymbol{h}}_{k,m}$, $\boldsymbol{\Psi}^h_{k,m}$, $\forall m$: $\boldsymbol{\Gamma}_m$

$\forall k \in \mathcal{K}_p$, $\forall m$, $\forall n$: $\hat{x}_{n,mk}(1) = \big[\boldsymbol{X}_p\big]_{mk}$, $\psi^x_{n,mk}(1) = 0$
$\forall k \in \mathcal{K}_d$, $\forall m$, $\forall n$: $\hat{x}_{n,mk}(1) = 0$, $\psi^x_{n,mk}(1) = 1$
$\forall k$, $\forall m$, $\forall n$: $\hat{h}_{k,nm}(1) = \big[\hat{\boldsymbol{H}}\big]_{nm}$, $\psi^h_{k,nm}(1) = \big[\hat{\boldsymbol{\Psi}}^h\big]_{nm}$
repeat
  $\forall k, \forall m, \forall n$: $\tilde{y}_{m,nk}(t) = y_{nk} - \sum_{i\neq m}^{M} \hat{h}_{k,ni}(t)\hat{x}_{n,ik}(t)$
  $\forall k, \forall m, \forall n$: $v^y_{m,nk}(t) = \sum_{i\neq m}^{M} |\hat{h}_{k,ni}(t)|^2 \psi^x_{n,ik}(t) + (|\hat{x}_{n,ik}(t)|^2 + \psi^x_{n,ik}(t))\psi^h_{k,ni}(t) + N_0$
  $\forall k \in \mathcal{K}_d, \forall m, \forall n$: $v^x_{m,nk}(t) = v^y_{m,nk}(t) + \psi^h_{k,nm}(t)$
  $\forall k, \forall m, \forall n$: $v^h_{m,nk}(t) = v^y_{m,nk}(t) + \gamma_{nm}\psi^x_{n,mk}(t)$
  $\forall k \in \mathcal{K}_d, \forall m, \forall n$: $\psi^r_{n,mk}(t) = \big(\sum_{i\neq n}^{N} \frac{|\hat{h}_{k,im}(t)|^2}{v^x_{m,ik}(t)}\big)^{-1}$
  $\forall k \in \mathcal{K}_d, \forall m, \forall n$: $\hat{r}_{n,mk}(t) = \psi^r_{n,mk}(t) \sum_{i\neq n}^{N} \frac{\hat{h}^*_{k,im}(t)\tilde{y}_{m,ik}(t)}{v^x_{m,ik}(t)}$
  $\forall k, \forall m, \forall n$: $\psi^q_{k,nm}(t) = \big(\sum_{i\neq k}^{K} \frac{|\hat{x}_{n,mi}(t)|^2}{v^h_{m,ni}(t)}\big)^{-1}$
  $\forall k, \forall m, \forall n$: $\hat{q}_{k,nm}(t) = \psi^q_{k,nm}(t) \sum_{i\neq k}^{K} \frac{\hat{x}^*_{n,mi}(t)\tilde{y}_{m,ni}(t)}{v^h_{m,ni}(t)}$
  $\forall k, \forall m$: $\boldsymbol{\mu}^h_{k,m}(t) = [\hat{q}_{k,1m}(t), \ldots, \hat{q}_{k,Nm}(t)]^{\mathrm{T}}$
  $\forall k, \forall m$: $\boldsymbol{\Sigma}^h_{k,m}(t) = \mathrm{diag}(\psi^q_{k,1m}(t), \ldots, \psi^q_{k,Nm}(t))$
  $\forall k, \forall m$: $\pi^h_{k,m}(t) = \boldsymbol{\mu}^{h\ \mathrm{H}}_{k,m}(t) \cdot (\boldsymbol{\Sigma}^{h\,-1}_{k,m}(t) - (\boldsymbol{\Sigma}^h_{k,m}(t) + \boldsymbol{\Gamma}_m)^{-1})\boldsymbol{\mu}^h_{k,m}(t)$
  $\forall k, \forall m$: $\psi^h_{k,m}(t) = \log(|\boldsymbol{\Sigma}^{h\,-1}_{k,m}(t)\boldsymbol{\Gamma}_m + \mathbf{I}_N|)$
  $\forall k, \forall m$: $\tau_{k,m}(t) = 1 + \frac{1-\lambda}{\lambda}\exp(-(\pi^h_{k,m}(t) - \psi^h_{k,m}(t)))$
  $\forall k, \forall m$: $\bar{\boldsymbol{h}}_{k,m}(t+1) = \frac{\boldsymbol{\Gamma}_m\left(\boldsymbol{\Sigma}^h_{k,m}(t) + \boldsymbol{\Gamma}_m\right)^{-1}}{\tau_{k,m}(t)} \boldsymbol{\mu}^h_{k,m}(t)$
  $\forall k, \forall m$: $\hat{\boldsymbol{h}}_{k,m}(t+1) = \eta\bar{\boldsymbol{h}}_{k,m}(t+1) + (1-\eta)\hat{\boldsymbol{h}}_{k,m}(t)$
  $\forall k, \forall m$: $\bar{\boldsymbol{\Psi}}^h_{k,m}(t+1) = (\tau_{k,m}(t) - 1)\mathrm{diag}(\bar{\boldsymbol{h}}_{k,m}(t)\bar{\boldsymbol{h}}^{\mathrm{H}}_{k,m}(t)) + \frac{\boldsymbol{\Sigma}^h_{k,m}(t)\boldsymbol{\Gamma}_m(\boldsymbol{\Sigma}^h_{k,m}(t) + \boldsymbol{\Gamma}_m)^{-1}}{\tau_{k,m}(t)}$
  $\forall k, \forall m$: $\boldsymbol{\Psi}^h_{k,m}(t+1) = \eta\bar{\boldsymbol{\Psi}}^h_{k,m}(t+1) + (1-\eta)\boldsymbol{\Psi}^h_{k,m}(t)$
  $\forall k \in \mathcal{K}_d, \forall m, \forall n$:
  $\bar{x}_{n,mk}(t+1) = \frac{\tau^{-1}_{k,m}(t)}{\sqrt{2}} \cdot \big(\tanh\big(\frac{\sqrt{2}\gamma(t)\Re(\hat{r}_{n,mk}(t))}{\psi^r_{n,mk}(t)}\big) + j \cdot \tanh\big(\frac{\sqrt{2}\gamma(t)\Im(\hat{r}_{n,mk}(t))}{\psi^r_{n,mk}(t)}\big)\big)$
  $\forall k \in \mathcal{K}_d, \forall m, \forall n$: $\hat{x}_{n,mk}(t+1) = \eta\bar{x}_{n,mk}(t+1) + (1-\eta)\hat{x}_{n,mk}(t)$
  $\forall k \in \mathcal{K}_d, \forall m, \forall n$: $\psi^x_{n,mk}(t+1) = \eta \cdot \tau^{-1}_{k,m}(1 - |\bar{x}_{n,mk}(t)|^2) + (1-\eta)\psi^x_{n,mk}(t)$
until $t = t_{max}$

Algorithm 4 Bilinear GaBP (Part 2: Hard Decision)

Inputs: $\boldsymbol{Y}$, $\forall k, \forall m, \forall n$: $\hat{x}_{n,mk}$, $\psi^{x}_{n,mk}$, $\forall k, \forall m$: $\hat{\boldsymbol{h}}_{k,m}$, $\boldsymbol{\Psi}^{h}_{k,m}$, $\forall m$: $\boldsymbol{\Gamma}_m$

Outputs: $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{H}}$, $\hat{\mathcal{A}}$

$\forall k, \forall m, \forall n$: $\tilde{y}_{m,nk} = y_{nk} - \sum_{i \neq m}^{M} \hat{h}_{k,ni} \hat{x}_{n,ik}$

$\forall k, \forall m, \forall n$: $v^{y}_{m,nk} = \sum_{i \neq m} |\hat{h}_{k,ni}|^2 \psi^{x}_{n,ik} + (|\hat{x}_{n,ik}|^2 + \psi^{x}_{n,ik}) \psi^{h}_{k,ni} + N_0$

$\forall k \in \mathcal{K}_d, \forall m, \forall n$: $v^{x}_{m,nk} = v^{y}_{m,nk} + \psi^{h}_{k,nm}$

$\forall k, \forall m, \forall n$: $v^{h}_{m,nk} = v^{y}_{m,nk} + \gamma_{nm} \psi^{x}_{n,mk}$

$\forall k \in \mathcal{K}_d, \forall m$: $\psi^{r}_{mk} = \left(\sum_{i=1}^{N} \frac{|\hat{h}_{k,im}|^2}{v^{x}_{m,ik}} \right)^{-1}$

$\forall k \in \mathcal{K}_d, \forall m$: $\hat{r}_{mk} = \psi^{r}_{mk} \sum_{i=1}^{N} \frac{\hat{h}^{*}_{k,im} \tilde{y}_{m,ik}}{v^{x}_{m,ik}}$

$\forall m, \forall n$: $\psi^{q}_{nm} = \left(\sum_{i=1}^{K} \frac{|\hat{x}_{n,mi}|^2}{v^{h}_{m,ni}} \right)^{-1}$

$\forall m, \forall n$: $\hat{q}_{nm} = \psi^{q}_{nm} \cdot \sum_{i=1}^{K} \frac{\hat{x}^{*}_{n,mi} \tilde{y}_{m,ni}}{v^{h}_{m,ni}}$

$\forall k \in \mathcal{K}_d, \forall m$: $\bar{x}_{mk} = \frac{1}{\sqrt{2}} \left(\tanh\left(\frac{\sqrt{2} \Re(\hat{r}_{mk})}{\psi^{r}_{mk}} \right) + j \cdot \tanh\left(\frac{\sqrt{2} \Im(\hat{r}_{mk})}{\psi^{r}_{mk}} \right) \right)$

$\forall k \in \mathcal{K}_d, \forall m$: $\hat{x}_{mk} = \underset{x_q \in \mathcal{C}}{\operatorname{argmin}} \ |x_q - \bar{x}_{mk}|$

$\forall m$: $\boldsymbol{\mu}^{h}_{m} = [\hat{q}_{1m}, \ldots, \hat{q}_{Nm}]^{\mathrm{T}}$

$\forall m$: $\boldsymbol{\Sigma}^{h}_{m} = \operatorname{diag}(\psi^{q}_{1m}, \ldots, \psi^{q}_{Nm})$

$\forall m$: $\pi^{h}_{m} = \boldsymbol{\mu}^{h\,\mathrm{H}}_{m} ({\boldsymbol{\Sigma}^{h}_{m}}^{-1} - (\boldsymbol{\Sigma}^{h}_{m} + \boldsymbol{\Gamma}_m)^{-1}) \boldsymbol{\mu}^{h}_{m}$

$\forall m$: $\psi^{h}_{m} = \log(|{\boldsymbol{\Sigma}^{h}_{m}}^{-1} \boldsymbol{\Gamma}_m + \mathbf{I}_N|)$

$\forall m$: $\tau_m = 1 + \frac{1-\lambda}{\lambda} \exp(-(\pi^{h}_{m} - \psi^{h}_{m}))$

$\forall m$: $\hat{\boldsymbol{h}}_m = \frac{\boldsymbol{\Gamma}_m \left(\boldsymbol{\Sigma}^{h}_{m} + \boldsymbol{\Gamma}_m \right)^{-1}}{\tau_m} \boldsymbol{\mu}^{h}_{m}$

$\hat{\mathcal{A}} = \text{ActivityDetectionPolicy}(\hat{h}_{nm})$

estimated channel $\hat{\boldsymbol{H}}$. The activity detection scheme considered in this section will be described in detail later.

The flow of Algorithms 3 and 4 generally follows the belief exchange steps described in Section 3.3.2, but in Algorithm 3, two belief manipulation techniques, namely, damping [146] and scaling [147] were introduced, with the objective of further improving the estimation accuracy and escaping from local minima. This is due to the fact that when the Gaussian approximation assumed in equation (3.12) does not sufficiently describe the actual stochastic behavior of the effective noise, the accuracy of soft-replicas is degraded by resultant belief outliers caused by the aforementioned approximation gap, leading to non-negligible estimation performance deterioration. Such approximation gap results from the increase in uncertainty that follows especially when the length of

pilot sequences decreases, which is the very aim of this section.

Following [99,103], we have applied damping to line 19, 21, 23 and 24 of Algorithm 3 with the damping factor $\eta \in [0, 1]$, which tends to prevent the algorithm from converging to local minima by forcing a slow update of soft-replicas, whereas belief scaling is adopted in line 22 of Algorithm 3 with parameter $\gamma(t)$, which in turn adjust the reliability of beliefs (*i.e.*, harnessing harmful outliers). The dynamics is designed to be a linear function of the number of iterations, that is,

$$\gamma(t) = \frac{t}{t_{\max}}. \tag{3.29}$$

Besides the above, as shown above, soft estimates of x_{mk} are obtained without considering the user activity (*i.e.*, row-sparsity) by imposing user activity detection upon the channel estimation process as described in equation (3.21b), indicating that such row-sparsity of $\boldsymbol{X}$ needs to be incorporated so as to avoid inconsistency with column-sparsity of $\boldsymbol{H}$.

Ironing out this issue, we leverage the sparsity factor $\tau_{k,m}$ given in equation (3.26) in line 22 and 24 of Algorithm 3, which tends to be 1 when active and to be ∞ otherwise, such that rows of $\boldsymbol{X}$ corresponding to non-active columns of $\boldsymbol{H}$ are suppressed, maintaining the consistency.

Initialization

Due to the fact that bilinear inference problems are strongly affected by the initial values of the solution variables, a reasonable initialization method is required so that the algorithm accurately estimates the channel, the informative data and the activity pattern simultaneously. However, one may also notice that due to the severe non-orthogonality of the pilot sequence (*i.e.*, $K_p \ll |\mathcal{A}| \ll M$) for overhead reduction, the accuracy of such an initial guess is not reliable enough. In light of the above, although several approaches developed for grant-free access such as covariance-based methods [47, 148] can be considered to produce initial $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}^h$, we have leveraged MMV-AMP [115] from a computational complexity perspective[4], which can be simply applied to equation (3.10) by regarding the first K_p columns of $\boldsymbol{Y}$ and the pilot matrix as the effective received signal matrix and its measurement matrix, respectively.

Activity Detection Policy

Below we describe how to identify active users based on the belief consensus performed in Algorithm 4. Due to the fact that miss-detections (MDs) and false alarms (FAs) are in a trade-off relationship as shown in the grant-free literature, such an activity

[4]Please refer to [112] for complexity analyses between existing grant-free schemes for further details.

detection policy is affected by system and user requirements, indicating that one needs to adopt a suitable criterion depending on the situation in practical implementations.

With that in mind, we consider the log likelihood ratio method based on estimated channel quantities, which thanks to uncorrelated Gaussianity of the channel and residual estimation error, can be written as

$$\mathrm{LLR}_m \triangleq \ln \frac{\prod_{n=1}^{N} \mathcal{CN}\left(0, \gamma_{nm} + \psi_{nm}^{h} | \hat{h}_{nm}\right)}{\prod_{n=1}^{N} \mathcal{CN}\left(0, \psi_{nm}^{h} | \hat{h}_{nm}\right)}, \tag{3.30}$$

where LLR_m is the log likelihood ratio corresponding to the m-th column and $\prod_{n=1}^{N}$ is introduced to perform consensus over the receive antenna dimension.

After basic manipulations, the log likelihood criterion can then be simplified to

$$\mathrm{LLR}_m = p_{\mathrm{active}}(m) - p_{\mathrm{inactive}}(m), \tag{3.31}$$

where

$$p_{\mathrm{active}}(m) \triangleq \sum_{n=1}^{N} \frac{-|\hat{h}_{nm}|^2}{\gamma_{nm} + \psi_{nm}^{h}} + \ln\big(\frac{1}{\pi(\gamma_{nm} + \psi_{nm}^{h})}\big) \tag{3.32a}$$

$$p_{\mathrm{inactive}}(m) \triangleq \sum_{n=1}^{N} \frac{-|\hat{h}_{nm}|^2}{\psi_{nm}^{h}} + \ln\big(\frac{1}{\pi\psi_{nm}^{h}}\big). \tag{3.32b}$$

With basis on the above, the set of active users is then determined as follows

$$\mathcal{A} = \big\{ m \,\big|\, p_{\mathrm{active}}(m) > p_{\mathrm{inactive}}(m) \big\}. \tag{3.33}$$

3.3.4 Performance Assessment

Below we evaluate via software simulation the proposed method in terms of BER, effective throughput, NMSE, and AUD performance.

Simulation Setup

Throughout this performance assessment section, we consider the following simulation setup unless otherwise specified. The number of receive antennas is set to $N = 100$, which are distributed over a square of side of 1000 [m] in a square mesh fashion, where $M = 100$ potential users are accommodated in each subcarrier. It is assumed that 50% of the M users become active during each OFDM frame, *i.e.*, $|\mathcal{A}| = M/2 = 50$, while K and K_p are considered to be $K \in \{140, 280\}$ and $K_p = 14$ depending on the employed subcarrier spacing[5]. It is further worth-noting that since we accommodate $M = 100$

[5]A scenario with $K = 140$ and $K = 280$ corresponds to the subcarrier spacing of 15 [kHz] and 30 [kHz], respectively. As shown in Figure 3.2, $K_p = 14$ indicates that only one OFDM slot is utilized as

users with overhead length $K_p = 14$, a significant amount of overhead reduction can be achieved. Although one might concern about performance degradation due to the resultant severe non-orthogonality, we dispel such concerns throughout this section by demonstrating that the bilinear inference method employed here is able to handle the non-orthogonality.

The covariance matrix employed in the multivariate Bernoulli-Gaussian model of equation (3.9) is constructed following the 3GPP urban microcell model [69], *i.e.*, $\mathbf{\Gamma}_m = \mathrm{diag}\left([\gamma_{1m}, \ldots, \gamma_{Nm}]\right)$, with each diagonal element obeying the relation $\gamma_{nm} \triangleq 10^{-\frac{\beta_{nm}}{10}}$ and β_{nm} [dB] $= 30.5 + 36.7 \log_{10}(d_{nm}) + \mathcal{N}\left(0, 4^2\right)$ where d_{nm} denotes the distance between the n-th AP and m-th user, given by $d_{nm} = \sqrt{(\rho_{AP} - \rho_U)^2 + \rho_R^2}$, with $\rho_{AP} = 10$ [m] and $\rho_U = 1.65$ [m] corresponding to the heights of the APs and users, respectively, while ρ_R is a random quantity.

The transmit power range at each uplink user is determined based on the experimental study presented in [149], where the transmit power of each uplink user is limited by 16 [dBm]. Furthermore, Gray-coded QPSK modulation is assumed to be employed at each user, whereas the noise floor N_0 at each AP is assumed to be modeled as

$$\sigma_u^2 = 10 \log_{10}(1000 \kappa T) + \mathrm{NF} + 10 \log_{10}(W) \,[\mathrm{dBm}], \tag{3.34}$$

where κ is the Boltzmann's constant, $T = 293.15$ denotes the physical temperature at each AP in kelvins, the noise figure NF is assumed to be 5 [dB] and W expresses the subcarrier bandwidth.

The pilot structure is designed via quadratic quadratic CSIDCO (QCSIDCO) in order to mitigate pilot contamination effects as much as possible even in case of severely non-orthogonal scenarios such as one considered in this section. Regarding the effective throughput performance, we adopt the definition proposed in [150, Def. 1], which is given by

$$\mathrm{R}_{\mathrm{eff}} \triangleq (1 - P_e) \cdot K_d \cdot b, \tag{3.35}$$

where P_e denotes the block (packet) error rate and b is the number of bits per symbol.

Finally, the maximum number of iterations in Algorithm 3 is set to $t_{\max} = 32$ and the damping factor η is 0.5, while the belief scaling factor $\gamma(t)$ is defined in equation (3.29).

Computational Complexity

Before proceeding to the performance evaluation via software simulations, we describe the computational complexity per iteration of the proposed JACDE algorithm, com-

pilot and the rest as data transmission, indicating that this situation imposes the most severe scenario as the pilot length is minimum.

paring it against those of different reference methods considered.

As can be seen from Algorithms 3 and 4, all the calculations required by the proposed method can be carried out in an element-wise manner. Taking into account the fact that the matrices $\mathbf{\Sigma}_{k,m}^{h}$ and $\mathbf{\Gamma}_{m}$ are both diagonal, it follows that Algorithm 3 has complexity of order $\mathcal{O}(NMK)$ both for multiplication/division and for addition/subtraction operators, resulting also in the complexity order per iteration of $\mathcal{O}(NMK)$ in total, including all steps to jointly perform CE, AUD and MUD.

For the sake of comparison, we consider the MMV-AMP algorithm [107–110] as state-of-the-art for JACE, and either the GaBP algorithm or the conventional ZF method as state-of-the-art for MUD. The complexity order per iteration for JACE via MMV-AMP is of $\mathcal{O}(NMK_p)$, whereas the complexity of $\mathcal{O}(NK_d|\hat{\mathcal{A}}_{\text{MMV}}|)$ and $\mathcal{O}(|\hat{\mathcal{A}}_{\text{MMV}}|^3 + |\hat{\mathcal{A}}_{\text{MMV}}|^2 K_d)$ is imposed for MUD by the GaBP algorithm and the ZF method [147], respectively, where $|\hat{\mathcal{A}}_{\text{MMV}}| \leq M$ denotes the number of active users estimated by MMV-AMP.

Assuming that MMV-AMP is capable of estimating active patterns with reasonable accuracy in the considered setup (*i.e.*, $|\hat{\mathcal{A}}_{\text{MMV}}| \approx M/2$), the complexity of the MMV-AMP algorithm followed by the GaBP method can be approximated written as $\mathcal{O}(NM(K_p+K_d/2))$. Therefore, it can be concluded that the proposed JACDE method is approximately in the same order of complexity of state-of-the-art alternatives.

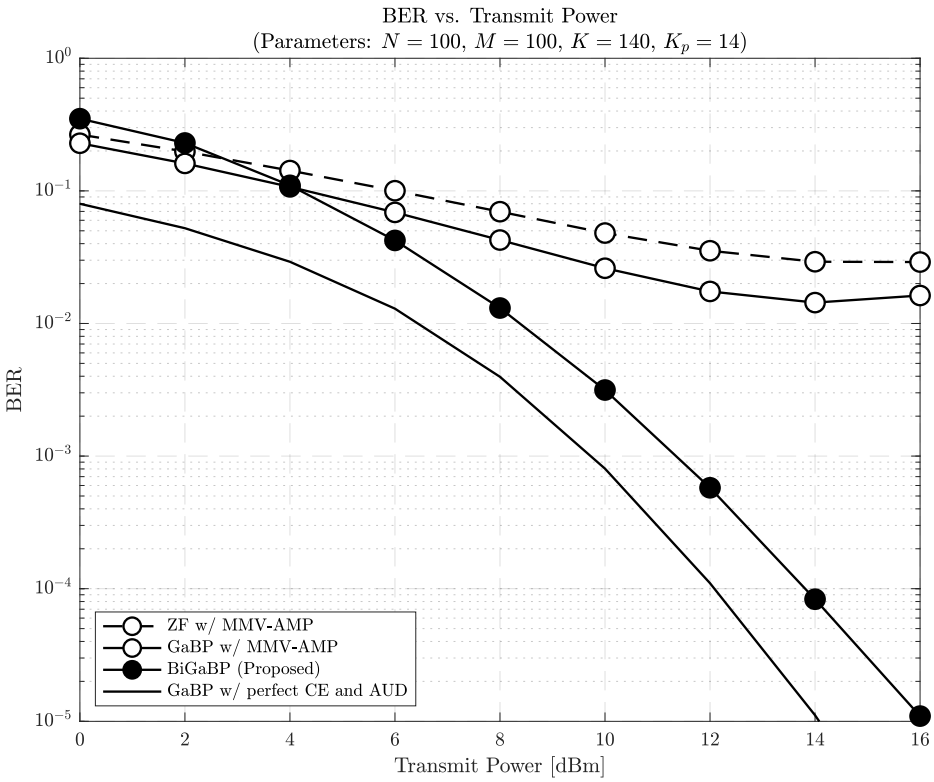
Multi-user Detection

The MUD performance of the proposed algorithm is studied in terms of uncoded BER as a function of transmit power. In order to take into account MD effects on data detection, we count not only bits received in error but also the number of lost bits due to missing user activity, *i.e.*,

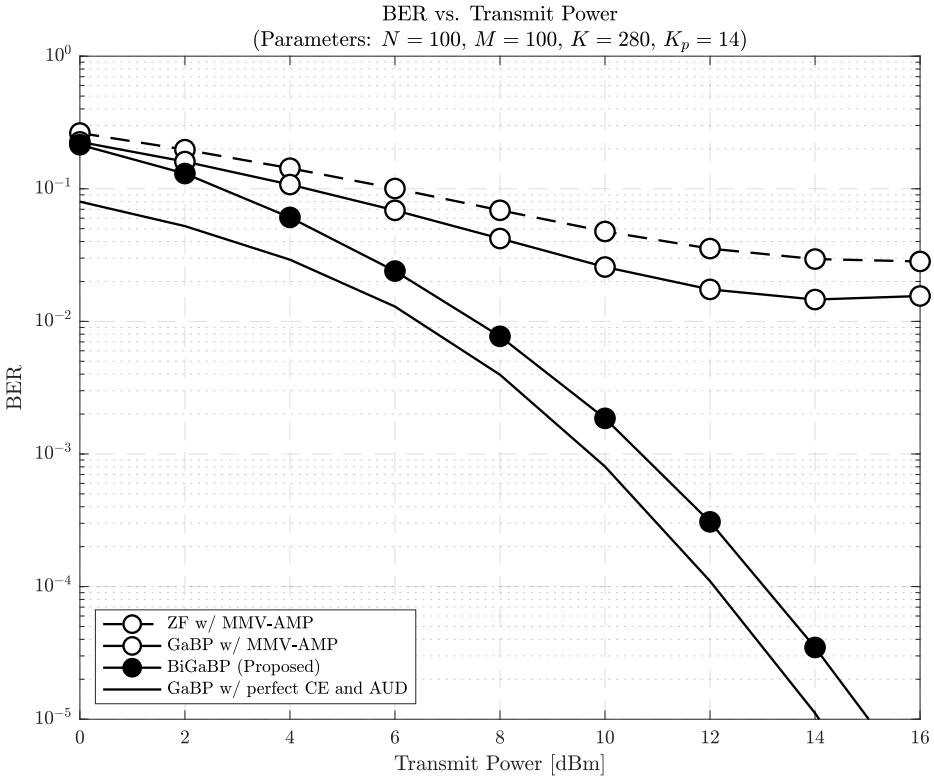
$$\text{BER} = \frac{P_e^1 + P_e^2}{\text{Total number of bits}}, \tag{3.36}$$

where P_e^1 denotes the number of errors due to failure of symbol detection and P_e^2 is the number of bits that have been lost due to failure of user detection.

Since there are no existing cell-free scheme with grant-free access which do not rely on spreading data sequences as described in equation (3.10), we consider the MMV-AMP algorithm as a state-of-the-art method to carry out JACE with basis of non-orthogonal pilot sequences $\boldsymbol{X}_p$, remarking that this receiver is widely employed in related literature [107–110]. For the same reason (of lack of a direct equivalent competitor), we also compare the performance of our method against an idealized system in which perfect CE and AUD is assumed, with signal detection performed by the GaBP algorithm.



(a) 15 kHz subcarrier spacing.



(b) 30 kHz subcarrier spacing.

Figure 3.5: BER comparisons as a function of transmit power.

Our assessment starts with Figure 3.5, where the BER performance of the proposed method is compared not only to the state-of-the-art but also to the ideal performance, for different subcarrier spacing scenarios (*i.e.*, $K \in \{140, 280\}$). The state-of-the-art methods compared are the linear ZF MIMO detector and the GaBP message passing MIMO detector, followed by MMV-AMP-based CE with the aid of the non-orthogonal pilot sequence. Results obtained from the GaBP MIMO detector with perfect CE and AUD provide lower bounds on the evaluated methods.

As can be seen from both figures, state-of-the-art methods suffer from high error floors, which stem from the poor CE and AUD performances by MMV-AMP, caused by the severely overloaded condition. In fact the aspect ratio (number of users over number of pilot symbols) of the pilot matrix is $\frac{M}{K_p} = \frac{100}{14} \approx 7.1428$, indicating a highly non-orthogonal condition. Although only $m = 50$ users out of $M = 100$ are assumed to be active in each coherent frame, the overloading ratio is still sufficiently high to hinder CE and AUD.

In contrast, the proposed method enjoys a water-falling curve in terms of BER for the both situations, which can be achieved by taking advantage of the pseudo-orthogonality of the data structure. This advantage can be confirmed from the fact that increasing the total symbol length from $K = 140$ to $K = 280$ while fixing the pilot length to be $K_p = 14$ can indeed enhance the detection performance as shown in Figure 3.5a and 3.5b. One may readily notice from the above that the corresponding CE performance can be also improved due to the same logic, which is offered below.

Effective Throughput

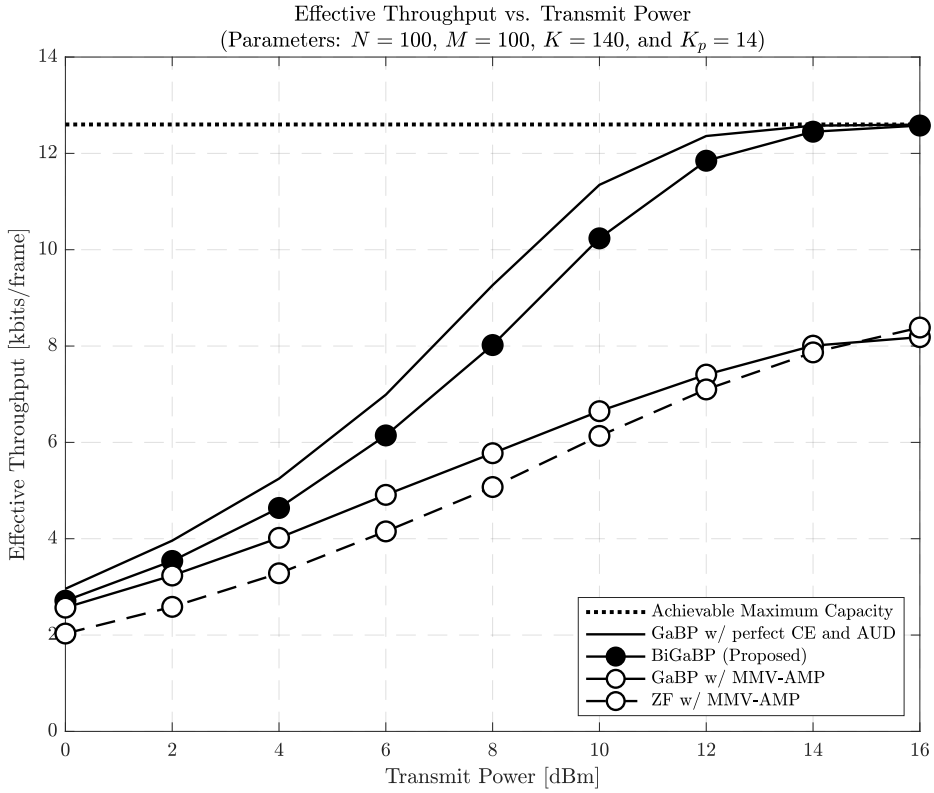
In light of the definition given in equation (3.35), we next investigate the effective throughput performance per each OFDM frame of the proposed method.

Simulation results showing the effective throughput achieved with the proposed scheme and compared alternatives are offered in Figure 3.6 for different subcarrier spacing setups. Note that the unit of the vertical axis is set to *kilobits* per frame for the sake of readability. Furthermore, we also implicitly measure the packet (block) error performance of the methods as shown in equation (3.35), which is often used for practical performance assessment. Finally, in addition to the three counterparts considered in the previous section, we also offer in both figures the system-level achievable maximum data rate as reference, which is determined by

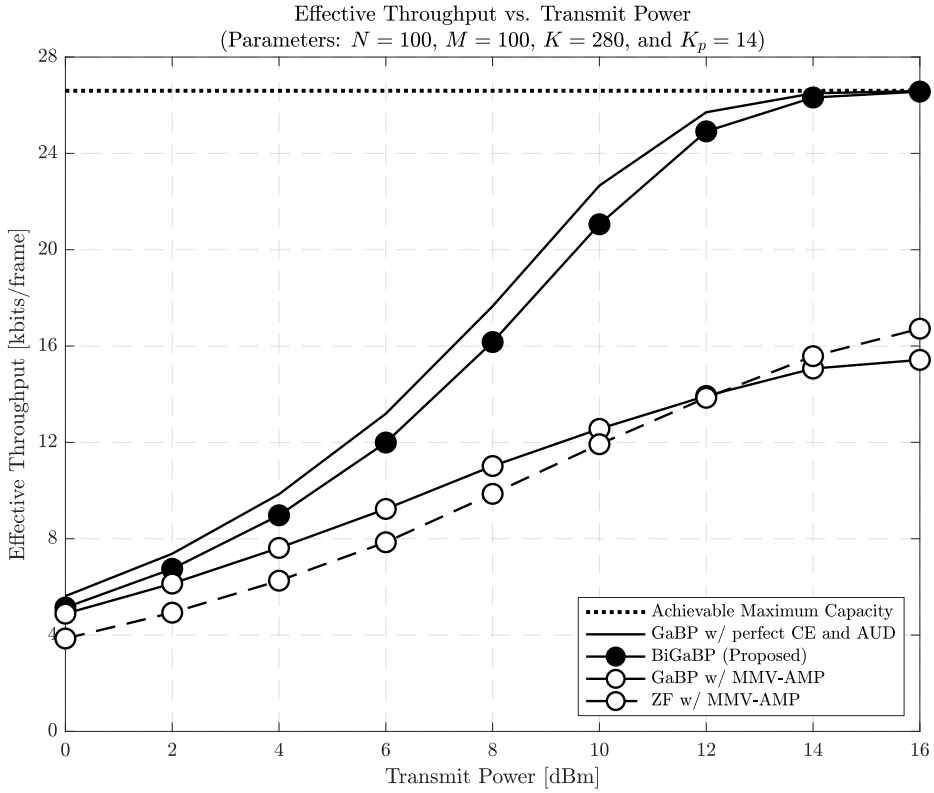
$$\text{Maximum Capacity} \triangleq K_d \cdot |\mathcal{A}| \cdot b \quad \text{[bits/frame]}, \tag{3.37}$$

where b denotes the number of bits per symbol.

As expected from the discussion of the previous section, it is found that the two state-of-the-art alternatives are incapable of successfully delivering bits transmitted by



(a) 15 kHz subcarrier spacing.



(b) 30 kHz subcarrier spacing.

Figure 3.6: Effective throughput comparisons as a function of transmit power.

the active users even with sufficiently high transmit power.

In contrast, the proposed method dynamically follows the same improvement in performance with transmit power as the idealized receiver, approaching the achievable capacity as the transmit power increases. It is furthermore seen that the throughput gap from the idealized scheme narrows as the subcarrier spacing increases, which is again due to the pseudo-orthogonality of the data sequence.

Channel Estimation

In addition to the above, the CE performance of the proposed method is assessed in this section as a function of transmit power for different data lengths, so that one may observe that the performance improvement described above is, at least in part, induced by the resultant CE. To this end, the NMSE performance of the proposed method for $K = 140$ and $K = 280$ is offered in Figure 3.7a and 3.7b, respectively, where the NMSE is defined as

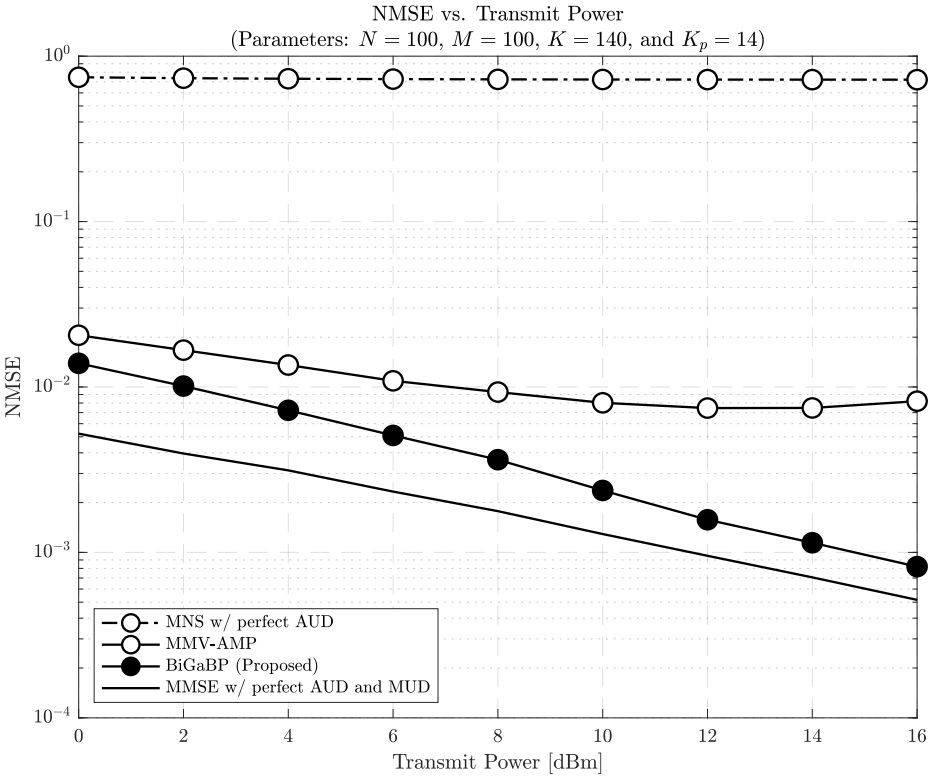
$$\text{NMSE} \triangleq \frac{\|\boldsymbol{H} - \hat{\boldsymbol{H}}\|_{\text{F}}^2}{\|\boldsymbol{H}\|_{\text{F}}^2}, \tag{3.38}$$

assuming a fixed pilot length $K_p = 14$.

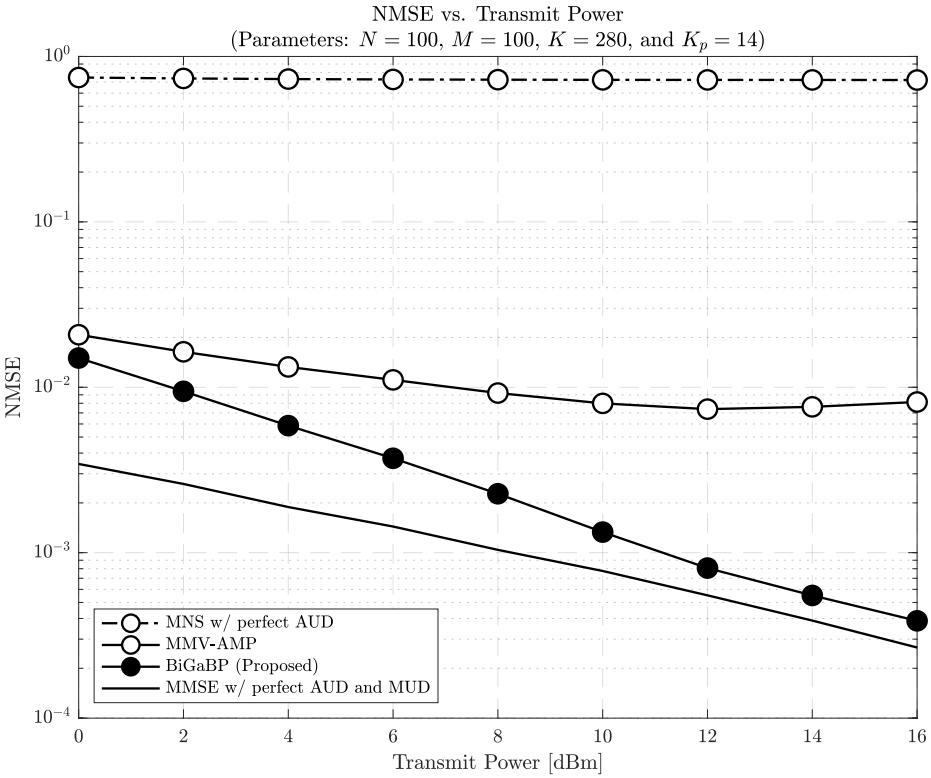
As for methods to compare, we have adopted not only MMV-AMP but also minimum norm solution (MNS) that is known to be a method to seek a closed-form unique CE solution in case of a non-orthogonal pilot sequence [103], while employing the MMSE performance with perfect knowledge of AUD and MUD at the receiver as reference. Please note that since the non-Bayesian approach, which takes advantage of the sample covariance of the received signals in order to detect user activity patterns, aims at only AUD, the resultant performance in terms of CE can be lower-bounded by MNS with perfect AUD.

With that in mind, it can be observed from Figure 3.7a and 3.7b that the proposed method can indeed improve the CE performance and approach the unachievable MMSE performance with perfect AUD and MUD, maintaining a similar gradient with that of the MMSE, whereas MMV-AMP and MNS suffer from a relatively high error floor due to the non-orthogonality of the pilot, although MMV-AMP appears to offer moderate performance in comparison with MNS.

Thanks to the pseudo-orthogonality of the data structure, the proposed method with 30 kHz subcarrier spacing again outperforms its own NMSE with 15 kHz subcarrier spacing. Furthermore, it can be mentioned that due to the sufficiently high CE accuracy of the proposed method (*i.e.*, NMSE $\in [10^{-3}, 10^{-4}]$), the considered non-coherent transmission architecture is comparable to the CE performance of the conventional grant-based MIMO systems (please refer to, for instance, [151]) to verify this claim.

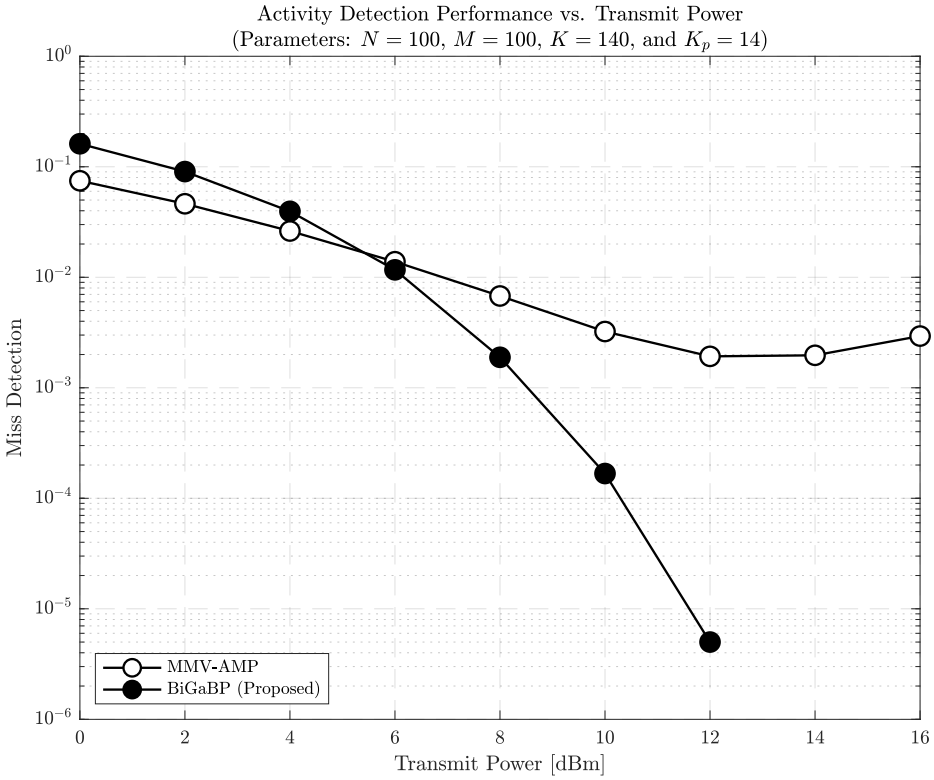


(a) 15 kHz subcarrier spacing.

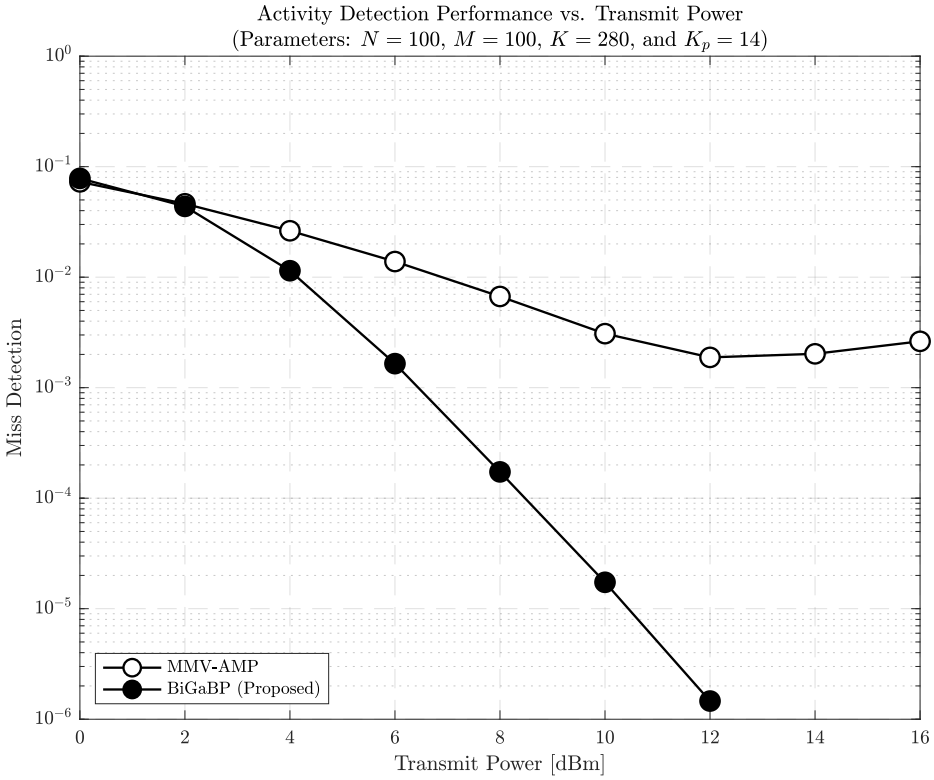


(b) 30 kHz subcarrier spacing.

Figure 3.7: NMSE comparisons as a function of transmit power.

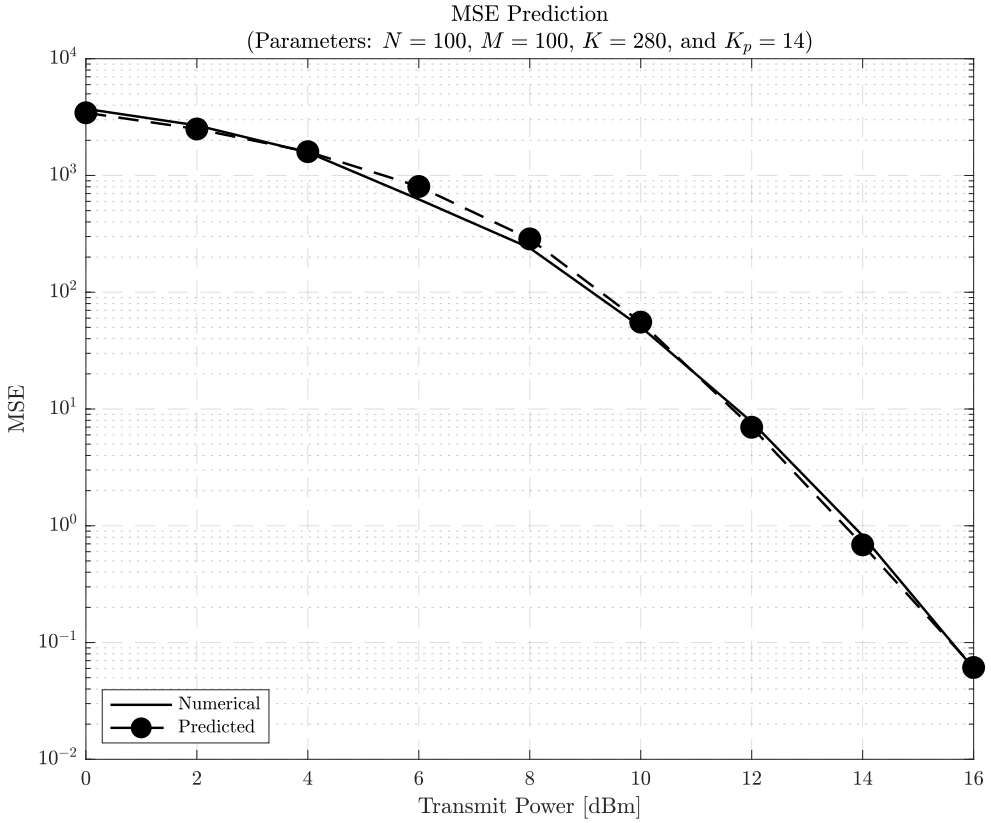


(a) 15 kHz subcarrier spacing.

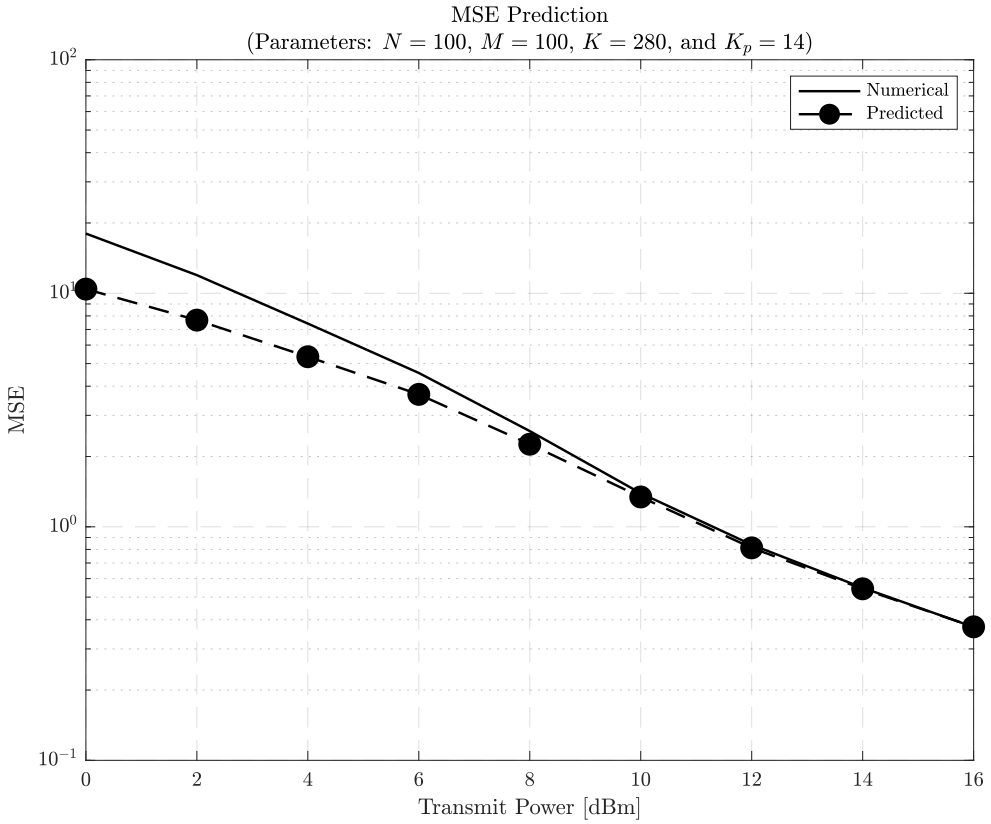


(b) 30 kHz subcarrier spacing.

Figure 3.8: MD comparisons as a function of transmit power.



(a) MSE error of $\hat{\boldsymbol{X}}$.



(b) MSE error of $\hat{\boldsymbol{H}}$.

Figure 3.9: MSE performance prediction via the state evolution of Algorithm 3 in comparison with the actual numerical evaluation as a function of transmit power.

Active User Detection

In this section, we evaluate the AUD performance of the proposed BiGaBP method. Although the AUD performance may be examined in terms of either FA, MD, or both, FA can be removed at higher layers by leveraging cyclic redundancy check codes [108], which are widely employed in practice. In light of the above, we adopt the occurrence of MDs as an AUD performance index.

In Figure 3.8, the MD probabilities of the proposed BiGaBP and MMV-AMP algorithms are illustrated for different symbol lengths K as a function of transmit power at each uplink user, while assuming $K_p = 14$ for both scenarios. It is perceived from Figure 3.8 that the proposed method can exponentially reduce the occurrence of MDs as transmit power increases, the reason of which can be explained from the discussions given in the preceding sections as follows.

As observed in Figures 3.5-3.7, the proposed BiGaBP algorithm starts to gradually recover the data and the channel from the observations $\boldsymbol{Y}$ as transmit power increases, which stems from the fact that the residual noise variances given in equation (3.23) and (3.28) are also accordingly reduced. Consequently, the resultant LLR given in (3.30) intends to be positive when $|\hat{h}_{nm}|$ is not sufficiently close to 0 and negative when $\prod_{n=1}^{N} \mathcal{CN}\left(0, \gamma_{nm} + \psi_{nm}^{h} | \hat{h}_{nm}\right) \approx 0$ in comparison with $\prod_{n=1}^{N} \mathcal{CN}\left(0, \psi_{nm}^{h} | \hat{h}_{nm}\right)$ for a small ψ_{nm}^{h}. Furthermore, the reason why the MD performance of MMV-AMP deteriorates in high transmit power regions can be explained as follows.

Besides the insufficient observations due to a non-orthogonal pilot structure, MMV-AMP suffers from the fact that in such a high SNR region, its estimation error noise variance becomes indistinguishable from the AWGN noise level at the receiver, leading to a tendency to regard inactive users as active and vice versa. In contrast, the proposed method mitigates this bottleneck by taking advantage of DoFs in the time domain.

MSE Performance Prediction and Its Accuracy

Finally, we evaluate the accuracy of MSE tracking via the state evolution of the proposed BiGaBP algorithm[6], where the predicted MSE performances of the data and channel are obtained by equation (3.23) and (3.28), respectively. In particular, the predicted MSE for $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{H}}$ is compared with the corresponding simulated counterpart in Figures 3.9a and 3.9b, respectively, with $K = 280$ and $K_p = 14$ and where the solid line and the dashed line with markers correspond to the simulated and predicted performance, respectively.

It can be observed from the figures that the state evolution can track the error

[6]Note that since the GaBP algorithm is a generalization of AMP, the resultant error level can be predicted in a similar fashion to the state evolution in AMP. For the sake of consistency, we call the corresponding error predicting quantities as BiGaBP's state evolution for the MSE performances.

performance of the proposed BiGaBP for both data and channel estimates, since the predicted MSEs follow approximately the same trajectory of its simulated counterpart.

As a final remark, it has been observed throughout the experiments that the proposed algorithm shows its operating-stability; thus, no unstable numerical calculation such as negative variance calculations and inversion of a singular matrix, which is often the case with algorithms based on expectation propagation, is needed.

3.4 XL-MIMO

3.4.1 System Model

In this section, we in turn consider another type of recently-emerging distributed MIMO architectures (*i.e.*, XL-MIMO) with the aim of reducing the overhead while addressing spatial non-stationarity, which consists of S sub-arrays, each equipped with N_s antenna elements, such that the total number of antenna array elements is given by $N = \sum_{s=1}^{S} N_s$, and let $\boldsymbol{G} \in \mathbb{C}^{N \times M}$ be the effective channel matrix between the XL-MIMO array and M single-antenna users, which jointly depicts user activities, sub-array VRs, and the fading gains. Then, the corresponding system model as shown in Figure 3.10 is given by

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{X} + \boldsymbol{W} \in \mathbb{C}^{N \times L}, \tag{3.39}$$

where $\boldsymbol{X} \in \mathbb{C}^{M \times L}$ is a pilot matrix collecting the L signals transmitted by each user, while $\boldsymbol{W} \in \mathbb{C}^{N \times L}$ denotes zero-mean unit-variance i.i.d. AWGN such that $\mathrm{vec}(\boldsymbol{W}) \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\mathbf{I})$.

In equation (3.39), it is assumed that only a small fraction of the M users is active, while the rest remains silent during the time interval of L transmissions. Letting K be a random variable that denotes the number of active users at a given time interval, the average user-activity probability can be expressed as $\lambda \triangleq \mathbb{E}[K/M]$. Furthermore, owing to *non-stationarities* observed in the XL-MIMO setting [118], the channel matrix $\boldsymbol{G}$ possesses block-sparsity that captures both user activity and the sub-arrays in their VRs (*i.e.* active sub-arrays), such that the m-th column of $\boldsymbol{G}$, relative to the m-th user, can be modeled as

$$\boldsymbol{g}_m = a_m \cdot \tilde{\boldsymbol{g}}_m \odot \boldsymbol{p}_m, \tag{3.40}$$

where $\odot$ denotes the Hadamard (element-wise) product, $a_m \in \{0, 1\}$ is the user activity indicator, $\tilde{\boldsymbol{g}}_m$ is the channel response vector, and $\boldsymbol{p}_m \triangleq [\boldsymbol{p}_{1m}^{\mathrm{T}}, \ldots, \boldsymbol{p}_{Sm}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{N \times 1}$ denotes a sub-array activity indicator defined by

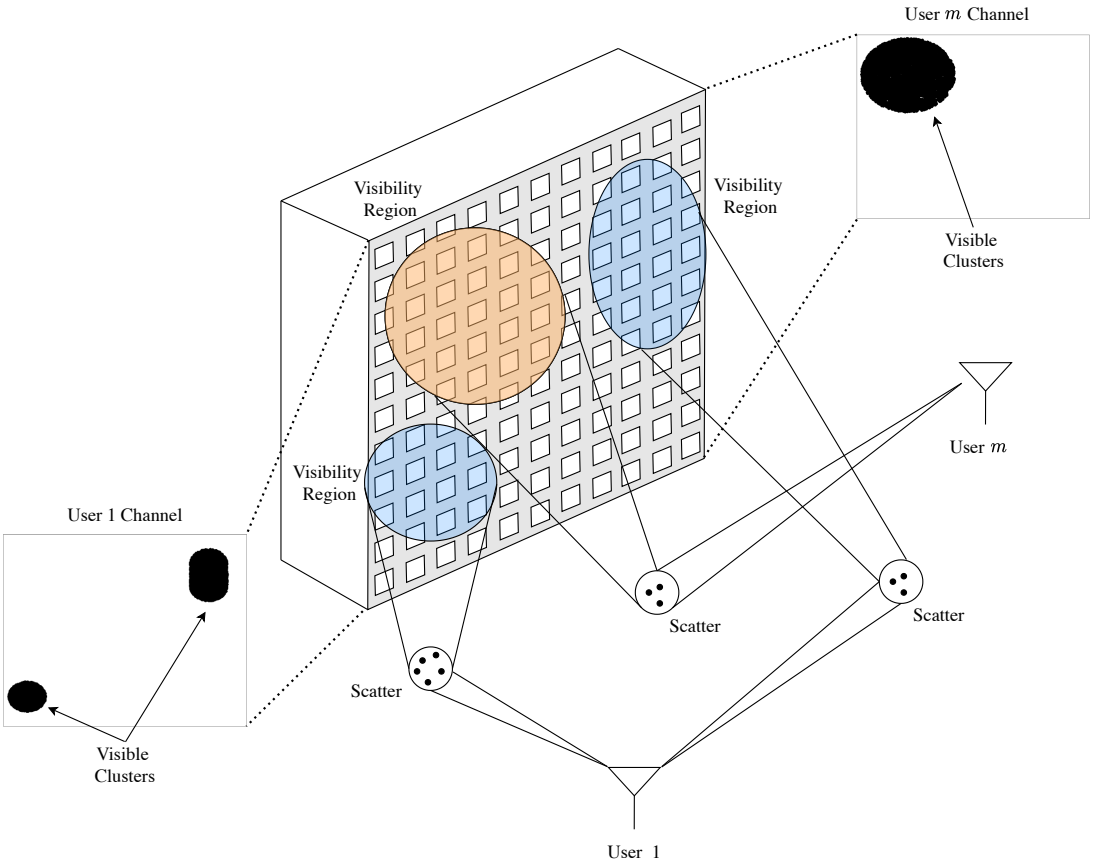


Figure 3.10: Illustration of the uplink of a multiuser XL-MIMO system with spatial non-stationarity, whereby each user independently activates different subarrays of the XL-MIMO array depending propagation conditions. ©2022 IEEE

$$\boldsymbol{p}_{sm} \triangleq \begin{cases} \mathbf{1}_{N_s \times 1} & \text{if the } s\text{-th sub-array is in the VR of the } m\text{-th user,} \\ \mathbf{0}_{N_s \times 1} & \text{if the } s\text{-th sub-array is outside the VR of the } m\text{-th user.} \end{cases} \tag{3.41}$$

Assuming that $\tilde{\boldsymbol{g}}_m$ is Gaussian, the distribution of $\boldsymbol{g}_m$ can be written as

$$\boldsymbol{g}_m \sim p_{\boldsymbol{g}_m}(\boldsymbol{g}_m) \triangleq \overbrace{(1-\lambda)\delta(\boldsymbol{g}_m)}^{\text{captures user activity}} + \lambda \overbrace{\prod_{s=1}^{S} \underbrace{f(\boldsymbol{g}_{\Phi(s)m}|\phi_{sm}, \mathbf{0}, \boldsymbol{\Gamma}_{sm})}_{\text{captures fading and sub-array activity}}}^{\text{jointing all sub-arrays}}, \tag{3.42}$$

where $\delta(\cdot)$ denotes the Dirac delta function, $\boldsymbol{\Gamma}_{sm}$ is the covariance matrix of the m-th user's channel to the s-th sub-array, ϕ_{sm} depicts the mean activity of the s-th sub-array, with respect to the m-th user and

$$f(\boldsymbol{z}|\phi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq (1-\phi)\delta(\boldsymbol{z}) + \phi \cdot \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.43}$$

where ϕ is an active probability, $\boldsymbol{\mu}$ denotes a certain mean, and $\boldsymbol{\Sigma}$ is a given covariance matrix.

3.4.2 Proposed Non-Stationarity Aware Detection Method

In this subsection, we propose a novel bilinear message passing algorithm for joint activity and channel estimation in XL-MIMO systems subjected to spatial non-stationarity. To this end, a decomposition of the system model given in equation (3.39) is carried out, followed by detailed derivations of message passing rules.

We remark that for the sake of generality, throughout the section it is assumed that the elements of the channel vectors $\tilde{\boldsymbol{g}}_m$ are independently but *not* identically distributed, which is equivalent to saying that the covariance matrices $\boldsymbol{\Gamma}_{sm}$ are all diagonal, but have different norms. This is motivated by the fact that the VRs of each user at the XL-MIMO array in general result from the impinging of signals from different propagation paths [126], as illustrated in Figure 3.10.

Reformulation

For the sake of future convenience, let us first reformulate equation (3.42) as

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{W} \in \mathbb{C}^{N\times L}, \tag{3.44}$$

with $\boldsymbol{G} \triangleq \boldsymbol{H}\boldsymbol{A} \in \mathbb{C}^{N\times M}$, where $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ is the block fading channel matrix and the $M \times M$ diagonal matrix $\boldsymbol{A}$, with $\text{diag}\,(\boldsymbol{A}) = [a_1, a_2, \ldots, a_M] \in \{0,1\}^M$, captures user activity indicators.

For notation simplicity, we hereafter introduce the quantity $\nu(s) \triangleq \sum_{i=1}^{s} N_i$, with $\nu(0) \triangleq 0$, to denote the cumulative collection of sub-array antenna indices at the s-th sub-array. In order to gain a closer insight into the effect of non-stationarity onto the column vectors of the channel matrix $\boldsymbol{H}$, consider the anatomized m-th column of $\boldsymbol{H}$, which is given by

$$\boldsymbol{h}_m = [\overbrace{p_{1m}\bar{\boldsymbol{h}}^{\mathrm{T}}_{\Phi(1)m}}^{\text{1st sub-array}},\ \overbrace{p_{2m}\bar{\boldsymbol{h}}^{\mathrm{T}}_{\Phi(2)m}}^{\text{2nd sub-array}},\ \ldots,\ \overbrace{p_{Sm}\bar{\boldsymbol{h}}^{\mathrm{T}}_{\Phi(S)m}}^{S\text{-th sub-array}}]^{\mathrm{T}}, \tag{3.45}$$

where $p_{ms} \in \{0,1\}$, with $s \in \{1,2,\ldots,S\}$, denotes the sub-array activity indicator, the vectors $\bar{\boldsymbol{h}}_{\Phi(s)m} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Gamma}_{sm})$, and $\Phi(s) \triangleq \{\nu(s-1)+1, \nu(s-1)+2, \ldots, \nu(s)\}$ is a set of antenna indices corresponding to the s-th sub-array.

More conveniently, the m-th column and s-th sub-array of the channel matrix can be modeled as a Bernoulli-Gaussian random variable, that is,

$$\boldsymbol{h}_{\Phi(s)m} \sim p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m}) \triangleq (1-\phi_{sm})\delta(\boldsymbol{h}_{\Phi(s)m}) + \phi_{sm}\cdot\mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Gamma}_{sm}), \tag{3.46}$$

where ϕ_{sm} denotes the mean of p_{ms}.

From the above, one can readily notice that the problem of jointly estimating users and sub-arrays activity indicators, as well as channel coefficients, belongs to the class of

bilinear inference problems. More precisely, given the received signal matrix $\boldsymbol{Y}$ and the predetermined reference signal matrix $\boldsymbol{X}$, our goal is to jointly estimate a_m, p_{ms} and $\boldsymbol{h}_{\Phi(s)m}$ for all $m \in \{1, 2, \ldots, M\}$ and $s \in \{1, 2, \ldots, S\}$, which are linearly multiplied by one another. In the next subsections we proceed to derive message passing rules designed to tackle this challenging problem, proposing a new joint activity and channel estimation for XL-MIMO subject to non-stationarity.

Factor Node

Focusing on the received signal element $y_{n\ell}$ at the n-th row and ℓ-th column of $\boldsymbol{Y}$, the received signal after SIC using tentative estimates can be written as

$$\tilde{y}_{m,n\ell} = y_{n\ell} - \overbrace{\sum_{i\neq m}^{M} \hat{h}_{\ell,ni}\hat{a}_{n\ell,i}x_{i\ell}}^{\text{SIC}} = h_{nm}a_m x_{m\ell} + \overbrace{\sum_{i\neq m}^{M}(h_{ni}a_i - \hat{h}_{\ell,ni}\hat{a}_{n\ell,i})x_{i\ell} + w_{n\ell}}^{\text{Residual interference plus noise}}, \quad (3.47)$$

where the soft estimates $\hat{h}_{\ell,nm}$ and $\hat{a}_{n\ell,i}$ are generated in variable nodes at the previous iteration, while $w_{n\ell}$ denotes the noise element at the n-th row and ℓ-th column of the AWGN matrix $\boldsymbol{W}$.

Assuming that the residual interference plus noise component of equation (3.47) can be approximated as a complex Gaussian random variable in conformity to the central limit theorem, the conditional PDF of equation (3.47) for given h_{nm} can be written as

$$p_{\tilde{y}_{m,n\ell}|h_{nm}}(\tilde{y}_{m,n\ell}|h_{nm}) \propto \exp\left(-\frac{|\tilde{y}_{m,n\ell} - h_{nm}\hat{a}_{n\ell,m}x_{m\ell}|^2}{v^h_{m,n\ell}}\right), \quad (3.48)$$

where the error variance is given by

$$\begin{aligned} v^h_{m,n\ell} &= \mathbb{E}\left[\left|(a_m - \hat{a}_{n\ell,m})h_{nm}x_{m\ell} + \sum_{i\neq m}^{M}(h_{ni}a_i - \hat{h}_{\ell,ni}\hat{a}_{n\ell,i})x_{i\ell} + w_{n\ell}\right|^2\right] \quad (3.49)\\ &= \psi^a_{n\ell,m}\gamma_{nm}|x_{m\ell}|^2 + \underbrace{\sum_{i\neq m}^{M}\left(|\hat{h}_{\ell,ni}|^2\psi^a_{n\ell,i} + (|\hat{a}_{n\ell,i}|^2 + \psi^a_{n\ell,i})\psi^h_{\ell,ni}\right)|x_{i\ell}|^2 + \sigma^2}_{\triangleq v^y_{m,n\ell}}, \end{aligned}$$

with γ_{nm} denoting the variance of the n-th row and m-th column of $\boldsymbol{H}$, and where we implicitly defined the residual error variance $v^y_{m,n\ell}$ for future convenience.

Similarly, the conditional PDF of $\tilde{y}_{m,n\ell}$ given a_m can be approximated as

$$p_{\tilde{y}_{m,n\ell}|a_m}(\tilde{y}_{m,n\ell}|a_m) \propto \exp\Big(-\frac{|\tilde{y}_{m,n\ell} - \hat{h}_{\ell,nm}a_m x_{m\ell}|^2}{v^a_{m,n\ell}}\Big), \quad (3.50)$$

with variance given by

$$\begin{aligned} v^a_{m,n\ell} &= \mathbb{E}\left[\left|(h_{nm}-\hat{h}_{\ell,nm})a_m x_{m\ell} + \sum_{i\neq m}^{M}(h_{ni}a_i - \hat{h}_{\ell,ni}\hat{a}_{n\ell,i})x_{i\ell} + w_{n\ell}\right|^2\right] \\ &= \psi^h_{\ell,nm}\lambda|x_{m\ell}|^2 + v^y_{m,n\ell}, \end{aligned} \tag{3.51}$$

where we utilized the fact that $\mathbb{E}\big[a_m^2\big] = \mathbb{E}[a_m] = \lambda$, since $a_m \in \{0,1\}$.

Variable Node

Taking advantage of the SIC mechanism and its resultant statistics shown above, the beliefs corresponding to the s-th sub-array combined over all available time resources except the ℓ-th time index, yields the PDF of the extrinsic belief $\boldsymbol{\xi}^h_{\ell,\Phi(s)m}$ given $\boldsymbol{h}_{\Phi(s)m}$, which is given by

$$\begin{aligned} &p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}) \\ &= \prod_{n=\nu(s-1)+1}^{\nu(s)}\prod_{i\neq\ell}^{L} p_{\tilde{y}_{m,ni}|h_{nm}}(\tilde{y}_{m,ni}|h_{nm}) \\ &\propto \prod_{n=\nu(s-1)+1}^{\nu(s)} \exp\left(-\frac{|h_{nm}-\mu^h_{\ell,nm}|^2}{\theta^h_{\ell,nm}}\right) \propto \mathcal{CN}(\boldsymbol{\mu}^h_{\ell,\Phi(s)m}, \boldsymbol{\Theta}^h_{\ell,\Phi(s)m}), \end{aligned} \tag{3.52}$$

with

$$\boldsymbol{\mu}^h_{\ell,\Phi(s)m} \triangleq [\mu^h_{\ell,(\nu(s-1)+1)m}, \ldots, \mu^h_{\ell,\nu(s)m}]^{\mathrm{T}} \in \mathbb{C}^{N_s\times 1}, \tag{3.53a}$$

$$\boldsymbol{\Theta}^h_{\ell,\Phi(s)m} \triangleq \mathrm{diag}\left([\theta^h_{\ell,(\nu(s-1)+1)m}, \ldots, \theta^h_{\ell,\nu(s)m}]\right) \in \mathbb{R}^{N_s\times N_s}, \tag{3.53b}$$

where

$$\theta^h_{\ell,nm} \triangleq \left(\sum_{i\neq\ell}^{L}\frac{|\hat{a}_{ni,m}x_{mi}|^2}{v^h_{m,ni}}\right)^{-1}, \tag{3.54a}$$

$$\mu^h_{\ell,nm} \triangleq \theta^h_{\ell,nm}\sum_{i\neq\ell}^{L}\frac{\tilde{y}_{m,ni}\hat{a}^*_{ni,m}x^*_{mi}}{v^h_{m,ni}}. \tag{3.54b}$$

In turn, the PDF of the extrinsic belief $\xi^a_{n\ell,m}$ given a_m can be similarly obtained as

$$\begin{aligned} p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|a_m) &= \prod_{j\neq n}^{N}\prod_{i\neq\ell}^{L} p_{\tilde{y}_{m,ji}|a_m}(\tilde{y}_{m,ji}|a_m) \\ &\propto \exp\left(-\frac{|a_m-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right) \propto \mathcal{CN}(\mu^a_{n\ell,m}, \theta^a_{n\ell,m}), \end{aligned} \tag{3.55}$$

where

$$\theta^a_{n\ell,m} \triangleq \left(\sum_{j\neq n}^{N}\sum_{i\neq \ell}^{L}\frac{|\hat{h}_{i,jm}x_{mi}|^2}{v^a_{m,ji}}\right)^{-1}, \tag{3.56a}$$

$$\mu^a_{n\ell,m} \triangleq \theta^a_{n\ell,m}\sum_{j\neq n}^{N}\sum_{i\neq \ell}^{L}\frac{\tilde{y}_{m,ji}\hat{h}^*_{i,jm}x^*_{mi}}{v^a_{m,ji}}. \tag{3.56b}$$

Combining the PDF in equation (3.52) with the prior channel PDF in equation (3.46) yields the posterior distribution of the channel. Therefore, taking the expectation of $\boldsymbol{h}_{\Phi(s)m}$ over the latter yields the corresponding soft estimate $\hat{\boldsymbol{h}}_{\ell,\Phi(s)m}$ at the ℓ-th variable node, which is given by

$$\hat{\boldsymbol{h}}_{\ell,\Phi(s)m} = \int_{\boldsymbol{h}_{\Phi(s)m}} \boldsymbol{h}_{\Phi(s)m}\frac{p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})}{\int_{\boldsymbol{h}'_{\Phi(s)m}} p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m})p_{\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{h}'_{\Phi(s)m})}, \tag{3.57}$$

where the denominator in the integrand is introduced for normalization purposes.

The error covariance associated with $\hat{\boldsymbol{h}}_{\ell,\Phi(s)m}$ is given by

$$\begin{aligned}\boldsymbol{\Psi}^h_{\ell,\Phi(s)m} \triangleq \operatorname{diag}\Bigg(\int_{\boldsymbol{h}_{\Phi(s)m}} \boldsymbol{h}_{\Phi(s)m}\boldsymbol{h}^{\mathrm{H}}_{\Phi(s)m}\frac{p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})}{\int_{\boldsymbol{h}'_{\Phi(s)m}} p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m})p_{\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{h}'_{\Phi(s)m})}\Bigg)\\ - \operatorname{diag}\left(\hat{\boldsymbol{h}}_{\ell,\Phi(s)m}\hat{\boldsymbol{h}}^{\mathrm{H}}_{\ell,\Phi(s)m}\right),\end{aligned} \tag{3.58}$$

such that $\boldsymbol{\Psi}^h_{\ell,\Phi(s)m} = \operatorname{diag}(\psi^h_{\ell,(\nu(s-1)+1)m}, \psi^h_{\ell,(\nu(s-1)+2)m}, \ldots, \psi^h_{\ell,\nu(s)m})$.

In turn, the soft replica $\hat{a}_{n\ell,m}$ of the user activity indicator can be similarly obtained as

$$\hat{a}_{n\ell,m} = \sum_{\alpha\in\{0,1\}} \alpha \cdot \frac{p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha)}{\sum_{\alpha'\in\{0,1\}} p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha')p_{a_m}(\alpha')}, \tag{3.59}$$

with $p_{a_m}(\cdot)$ denoting the Bernoulli probability mass function (PMF) with intensity λ, which is accompanied by its MSE given by

$$\psi^a_{n\ell,m} = \sum_{\alpha\in\{0,1\}} \alpha^2 \cdot \frac{p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha)}{\sum_{\alpha'\in\{0,1\}} p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha')p_{a_m}(\alpha')} - \hat{a}^2_{n\ell,m}. \tag{3.60}$$

In order to compute equations (3.57) - (3.60) above, it is essential to analyze the effective distributions, namely,

$$p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m}) \quad \text{and} \quad p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha) \tag{3.61}$$

and such that the resultant calculations become tractable. To that end, plugging equations (3.46) and (3.52) into the effective distribution of $\boldsymbol{h}_{\Phi(s)m}$ yields

$$
\begin{aligned}
&p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})\\
&=\big((1-\phi_{sm})\delta(\boldsymbol{h}_{\Phi(s)m})+\phi_{sm}\cdot\mathcal{CN}(\mathbf{0},\boldsymbol{\Gamma}_{sm})\big)\times\mathcal{CN}(\boldsymbol{\mu}^h_{\ell,\Phi(s)m},\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})\\
&=(1-\phi_{sm})\delta(\boldsymbol{h}_{\Phi(s)m})\frac{\exp\left(-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}\boldsymbol{\Theta}^{h-1}_{\ell,\Phi(s)m}\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\right)}{\pi^{N_s}|\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}|}\\
&\quad+\phi_{sm}\frac{\exp\big(-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\big)}{\pi^{N_s}|\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}|}\\
&\quad\times\underbrace{\mathcal{CN}\Big(\boldsymbol{\Gamma}_{sm}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}\boldsymbol{\mu}^h_{\ell,\Phi(s)m},\boldsymbol{\Gamma}_{sm}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}\Big)}_{\triangleq f(\boldsymbol{h}_{\Phi(s)m}|\boldsymbol{\mu}^h_{\ell,\Phi(s)m},\boldsymbol{\Theta}^h_{\ell,\Phi(s)m},\boldsymbol{\Gamma}_{sm})}.
\end{aligned}
\tag{3.62}
$$

Thus, the corresponding normalization factor can be written as

$$
\begin{aligned}
&\int_{\boldsymbol{h}_{\Phi(s)m}}p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})\\
&=\int_{\boldsymbol{h}_{\Phi(s)m}}(1-\phi_{sm})\delta(\boldsymbol{h}_{\Phi(s)m})\frac{\exp\left(-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}\boldsymbol{\Theta}^{h-1}_{\ell,\Phi(s)m}\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\right)}{\pi^{N_s}|\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}|}\\
&+\phi_{sm}\frac{\exp\left(-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\right)}{\pi^{N_s}|\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}|}\overbrace{\int_{\boldsymbol{h}_{\Phi(s)m}}f(\boldsymbol{h}_{\Phi(s)m}|\boldsymbol{\mu}^h_{\ell,\Phi(s)m},\boldsymbol{\Theta}^h_{\ell,\Phi(s)m},\boldsymbol{\Gamma}_{sm})}^{=1}\\
&=\phi_{sm}\frac{\exp\big(-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\big)}{\pi^{N_s}|\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}|}\tau_{\ell,ms},
\end{aligned}
\tag{3.63}
$$

where the activity detection factor $\tau_{\ell,ms}$ is given by

$$
\begin{aligned}
\tau_{\ell,ms}\triangleq 1+\frac{1-\phi_{sm}}{\phi_{sm}}\exp\Big(&-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}\Big(\boldsymbol{\Theta}^{h-1}_{\ell,\Phi(s)m}-(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}\Big)\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\\
&+\log\left(|\boldsymbol{\Theta}^{h-1}_{\ell,\Phi(s)m}\boldsymbol{\Gamma}_{sm}+\mathbf{I}_{N_s}|\right)\Big).
\end{aligned}
\tag{3.64}
$$

Taking advantage of equations (3.62) and (3.63), the soft replica of $\boldsymbol{h}_{\Phi(s)m}$ at the ℓ-th node can be re-written as

$$
\bar{\boldsymbol{h}}_{\ell,\Phi(s)m}=\frac{\boldsymbol{\Gamma}_{sm}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}}{\tau_{\ell,ms}}\boldsymbol{\mu}^h_{\ell,\Phi(s)m},
\tag{3.65}
$$

while its MSE $\bar{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m}\triangleq\mathrm{diag}\left(\psi^h_{\ell,(\nu(s-1)+1)m},\ldots,\psi^h_{\ell,\nu(s)m}\right)$ can be expressed as

$$
\bar{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m}=(\tau_{\ell,ms}-1)\mathrm{diag}\left(\hat{\boldsymbol{h}}_{\ell,\Phi(s)m}\hat{\boldsymbol{h}}^{\mathrm{H}}_{\ell,\Phi(s)m}\right)+\frac{\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}\boldsymbol{\Gamma}_{sm}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}}{\tau_{\ell,ms}}.
\tag{3.66}
$$

In turn, the effective distribution $p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha)$ can be simplified to

$$p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha) = \exp\left(-\frac{|\alpha-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right) \times \overbrace{(\lambda\alpha+(1-\lambda)(1-\alpha))}^{\text{Bernoulli PDF with intensity } \lambda}, \quad (3.67)$$

where $\alpha \in \{0,1\}$ and the associated normalization factor can then be written as

$$\sum_{\alpha\in\{0,1\}} p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha) = \lambda\exp\left(-\frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)+(1-\lambda)\exp\left(-\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right). \quad (3.68)$$

Thus, the soft replica of α_m can be obtained as

$$\begin{aligned}\hat{a}_{n\ell,m} &= \sum_{\alpha\in\{0,1\}} \alpha\cdot\frac{p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha)}{\sum_{\alpha'\in\{0,1\}} p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha')p_{a_m}(\alpha')} \\ &= \underbrace{\frac{1}{1+\frac{1-\lambda}{\lambda}\exp\left(-\left(\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}-\frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)\right)}}_{\text{Weighted Sigmoid Function}},\end{aligned} \quad (3.69)$$

and its error variance can be written as

$$\begin{aligned}\psi^a_{n\ell,m} &= \sum_{\alpha\in\{0,1\}} \alpha^2\cdot\frac{p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha)p_{a_m}(\alpha)}{\sum_{\alpha'\in\{0,1\}} p_{\xi_{n\ell,m}|a_m}(\xi_{n\ell,m}|\alpha')p_{a_m}(\alpha')} - \hat{a}^2_{n\ell,m} \\ &= \frac{\frac{1-\lambda}{\lambda}\exp\left(-\left(\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}-\frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)\right)}{\left(1+\frac{1-\lambda}{\lambda}\exp\left(-\left(\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}-\frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)\right)\right)^2},\end{aligned} \quad (3.70)$$

from which it is readily found[7] that $0 \le \psi^a_{n\ell,m} \le 1/4$.

Learning Array Activity via Expectation Maximization

In this subsection, we consider an auto-parameterization approach aided by expectation-maximization to learn sub-array activity indicators. Sub-array activity indicators ϕ_{sm} are dependent upon instantaneous propagation environments, which may be difficult to obtain in practice and may not be compensated by the long-term statistical knowledge.

[7]The error variance can be upper-bounded by 1/4 under the Gaussian approximation.

We therefore propose to find an estimate of ϕ_{sm} via

$$\hat{\phi}_{sm} = \underset{\phi_{sm}}{\operatorname{argmax}} \sum_{\ell=1}^{L} \int_{\boldsymbol{h}_{\Phi(s)m}} \frac{\log(p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m};\phi_{sm}))p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})}{\int_{\boldsymbol{h}'_{\Phi(s)m}} p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m})p_{\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{h}'_{\Phi(s)m})}, \tag{3.71}$$

with the maximization constrained to satisfying the first-order necessary condition

$$\sum_{\ell=1}^{L} \int_{\boldsymbol{h}_{\Phi(s)m}} \frac{d\log(p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m};\phi_{sm}))}{d\phi_{sm}} \frac{p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})}{\int_{\boldsymbol{h}'_{\Phi(s)m}} p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m})p_{\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{h}'_{\Phi(s)m})} = 0, \tag{3.72}$$

where the derivative can be written as

$$\frac{d\log(p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m};\phi_{sm}))}{d\phi_{sm}} = \begin{cases} \frac{1}{\phi_{sm}} & \text{if } \boldsymbol{h}_{\Phi(s)m} \neq \mathbf{0}, \\ \frac{-1}{1-\phi_{sm}} & \text{if } \boldsymbol{h}_{\Phi(s)m} = \mathbf{0}. \end{cases} \tag{3.73}$$

Treating the neighborhood around $\boldsymbol{h}_{\Phi(s)m} = \mathbf{0}$ and the rest separately, equation (3.72) boils down to the following equality condition

$$\sum_{\ell=1}^{L} \frac{1}{\phi_{sm}\tau_{\ell,ms}} = \sum_{\ell=1}^{L} \frac{1}{1-\phi_{sm}} \frac{\tau_{\ell,ms}-1}{\tau_{\ell,ms}}, \tag{3.74}$$

which readily yields

$$\hat{\phi}_{ms} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{1}{\tau_{\ell,ms}}. \tag{3.75}$$

Assuming uniformity among sub-array activity indicators (*i.e.*, $\phi = \phi_{sm}$ for all m and s), the above derivation can be further generalized to combine all available information over the spatial and user domains, namely,

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}} \sum_{\ell=1}^{L} \sum_{m=1}^{M} \sum_{s=1}^{S} \int_{\boldsymbol{h}_{\Phi(s)m}} \frac{\log(p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m};\phi))p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}_{\Phi(s)m})p_{\boldsymbol{h}_{\Phi(s)m}}(\boldsymbol{h}_{\Phi(s)m})}{\int_{\boldsymbol{h}'_{\Phi(s)m}} p_{\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{\xi}_{\ell,\Phi(s)m}|\boldsymbol{h}'_{\Phi(s)m})p_{\boldsymbol{h}'_{\Phi(s)m}}(\boldsymbol{h}'_{\Phi(s)m})}, \tag{3.76}$$

which finally yields

$$\hat{\phi} = \frac{1}{MSL} \sum_{\ell=1}^{L} \sum_{m=1}^{M} \sum_{s=1}^{S} \frac{1}{\tau_{\ell,ms}}. \tag{3.77}$$

Activity Detection Policy

Finally, in this subsection we discuss the last step, corresponding to the refinement of the user activity estimates according to a pre-defined user activity policy. Different approaches to perform this step can be considered, such as those proposed in [47, 108, 114]. Here, we adopt the new log-likelihood ratio (LLR)-based approach taking both $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{A}}$ into consideration.

To that end, first recognize that the user activity pattern captured by the binary quantities on the diagonal elements of $\hat{\boldsymbol{A}}$ can equivalently be expressed as column-sparsity of $\hat{\boldsymbol{H}}$, which implies that an activity detection policy must jointly consider $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{A}}$ for AUD. Denoting the estimated effective channel matrix $\hat{\boldsymbol{G}} \triangleq \hat{\boldsymbol{H}}\hat{\boldsymbol{A}}$, the element-wise LLR can be written as

$$\Lambda_{nm} = \log \frac{\mathcal{CN}(0, \gamma_{nm} + \psi^h_{nm} | \hat{g}_{nm})}{\mathcal{CN}(0, \psi^h_{nm} | \hat{g}_{nm})}. \tag{3.78}$$

From the above, the user activity can be detected by combining the LLRs Λ_{nm} for all the receive antenna dimensions N. However, such a detection policy ignores the presence of the block-wise sparsity due to spatial non-stationarity, leading to detection performance degradation. To address this issue, we consider the following sub-array activity aware AUD policy.

$$\hat{a}_m = \begin{cases} 1 & \text{if } \max\limits_{s}(\sum_{n\in\Phi(s)} \Lambda_{nm}) \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.79}$$

Algorithm Description and Discussion

In this subsection we offer several remarks on the message passing and consensus mechanisms for JACE proposed above, which for convenience is concisely summarized in Algorithm 5. Referring to Algorithm 5, first notice that the procedure requires two initialization quantities, namely, initial values of the channel matrix $\hat{\boldsymbol{H}}$ and error covariance matrix $\hat{\boldsymbol{\Psi}}^h$, which can be obtained via a number of state-of-the art methods, such as the AUD-aware approximate BP algorithm proposed in [115], adopted here due to its complexity-performance tradeoff advantages. Besides that, the proposed JACE algorithm takes as inputs the received signal matrix $\boldsymbol{Y}$ and the pilot matrix $\boldsymbol{X}$; to which it outputs estimates of the channel matrix $\hat{\boldsymbol{H}}$ and of the user activity matrix $\hat{\boldsymbol{A}}$.

The algorithm has two essential stages, the iterative stage described by lines 3 to 19 within which the beliefs are propagated and exchanged between factor and variable nodes, and the consensus stage in line 20 where the output quantities are finally determined based on the obtained beliefs. Notice that lines 17 and 18 correspond to a well-known damping procedure (see *e.g.* [147]), which aims to avoid estimates being trapped at a local optimum, especially at the early stage of the iterations by allowing a slow update of the quantities $\hat{\boldsymbol{h}}_{\ell,\Phi(s)m}$, $\hat{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m}$, $\hat{a}_{n\ell,m}$, and $\psi^a_{n\ell,m}$. This is due to the fact that at the early stage of the iterative process, the Gaussian approximation assumed in equation (3.47) may not capture the actual statistics of the effective noise, which might lead to convergence to a local optimum point.

Algorithm 5 Proposed JACE in XL-MIMO with Non-Stationarity

Inputs: $\boldsymbol{Y}$ and $\boldsymbol{X}$, and initializers $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}^h$
Outputs: $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{A}}$

For all (n, m, ℓ)

$\hat{a}_{n\ell,m}(1) \leftarrow 0$, $\psi^a_{n\ell,m}(1) \leftarrow 0$

$\hat{h}_{\ell,nm}(1) \leftarrow \left[\hat{\boldsymbol{H}}\right]_{nm}$, $\psi^h_{\ell,nm}(1) \leftarrow \left[\hat{\boldsymbol{\Psi}}^h\right]_{nm}$

repeat

For all (n, m, ℓ)

Eq. (3.47): $\tilde{y}_{m,n\ell} \leftarrow y_{n\ell} - \sum_{i\neq m}^{M} \hat{h}_{\ell,ni}\hat{a}_{n\ell,i}x_{i\ell}$

Eq. (**??**): $v^y_{m,n\ell} \leftarrow \sum_{i\neq m}^{M} \left(|\hat{h}_{\ell,ni}|^2\psi^a_{n\ell,i} + (|\hat{a}_{n\ell,i}|^2 + \psi^a_{n\ell,i})\psi^h_{\ell,ni}\right)|x_{i\ell}|^2 + \sigma^2$

$v^h_{m,n\ell} \leftarrow \psi^a_{n\ell,m}\gamma_{nm}|x_{m\ell}|^2 + v^y_{m,n\ell}$

Eq. (3.51): $v^a_{m,n\ell} \leftarrow \psi^h_{\ell,nm}\lambda|x_{m\ell}|^2 + v^y_{m,n\ell}$

Eq. (3.54): $\theta^h_{\ell,nm} \leftarrow \left(\sum_{i\neq\ell}^{L} \frac{|\hat{a}_{ni,m}x_{mi}|^2}{v^h_{m,ni}}\right)^{-1}$ and $\mu^h_{\ell,nm} \leftarrow \theta^h_{\ell,nm}\sum_{i\neq\ell}^{L} \frac{\tilde{y}_{m,ni}\hat{a}^*_{ni,m}x^*_{mi}}{v^h_{m,ni}}$

Eq. (3.56): $\theta^a_{n\ell,m} \leftarrow \left(\sum_{j\neq n}^{N}\sum_{i\neq\ell}^{L} \frac{|\hat{h}_{i,jm}x_{mi}|^2}{v^a_{m,ji}}\right)^{-1}$

$\mu^a_{n\ell,m} \leftarrow \theta^a_{n\ell,m}\sum_{j\neq n}^{N}\sum_{i\neq\ell}^{L} \frac{\tilde{y}_{m,ji}\hat{h}^*_{i,jm}x^*_{mi}}{v^a_{m,ji}}$

Eq. (3.59): $\hat{a}_{n\ell,m} \leftarrow \left(1 + \frac{1-\lambda}{\lambda}\exp\left(-\left(\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}} - \frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)\right)\right)$

Eq. (3.60): $\psi^a_{n\ell,m} \leftarrow \frac{\frac{1-\lambda}{\lambda}\exp\left(-\left(\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}} - \frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)\right)}{\left(1+\frac{1-\lambda}{\lambda}\exp\left(-\left(\frac{|\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}} - \frac{|1-\mu^a_{n\ell,m}|^2}{\theta^a_{n\ell,m}}\right)\right)\right)^2}$

For all (m, s, ℓ)

Eq. (3.53): $\boldsymbol{\mu}^h_{\ell,\Phi(s)m} \leftarrow [\mu^h_{\ell,(\nu(s-1)+1)m}, \ldots, \mu^h_{\ell,\nu(s)m}]^{\mathrm{T}}$

$\boldsymbol{\Theta}^h_{\ell,\Phi(s)m} \leftarrow \mathrm{diag}\left([\theta^h_{\ell,(\nu(s-1)+1)m}, \ldots, \theta^h_{\ell,\nu(s)m}]\right)$

Eq. (3.64): $\tau_{\ell,ms} \leftarrow 1 + \frac{(1-\hat{\phi}_{sm})|\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h{}_{\ell,\Phi(s)m}|}{\hat{\phi}_{sm}|\boldsymbol{\Theta}^h{}_{\ell,\Phi(s)m}|}e^{\left(-\boldsymbol{\mu}^{h\mathrm{H}}_{\ell,\Phi(s)m}(\boldsymbol{\Theta}^{h\,-1}_{\ell,\Phi(s)m} - (\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1})\boldsymbol{\mu}^h_{\ell,\Phi(s)m}\right)}$

Eq. (3.71): $\hat{\phi}_{ms} = \hat{\phi} \leftarrow \frac{1}{MSL}\sum_{\ell=1}^{L}\sum_{m=1}^{M}\sum_{s=1}^{S}\frac{1}{\tau_{\ell,ms}}$

Eq. (3.65): $\bar{\boldsymbol{h}}_{\ell,\Phi(s)m} \leftarrow \frac{\boldsymbol{\Gamma}_{sm}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}}{\tau_{\ell,ms}}\boldsymbol{\mu}^h_{\ell,\Phi(s)m}$

Eq. (3.66): $\bar{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m} \leftarrow (\tau_{\ell,ms}-1)\mathrm{diag}\left(\bar{\boldsymbol{h}}_{\ell,\Phi(s)m}\bar{\boldsymbol{h}}^{\mathrm{H}}_{\ell,\Phi(s)m}\right) + \frac{\boldsymbol{\Theta}^h_{\ell,\Phi(s)m}\boldsymbol{\Gamma}_{sm}(\boldsymbol{\Gamma}_{sm}+\boldsymbol{\Theta}^h_{\ell,\Phi(s)m})^{-1}}{\tau_{\ell,ms}}$

Damping [147]: $\hat{\boldsymbol{h}}_{\ell,\Phi(s)m} \leftarrow \eta\bar{\boldsymbol{h}}_{\ell,\Phi(s)m} + (1-\eta)\hat{\boldsymbol{h}}_{\ell,\Phi(s)m}$

$\hat{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m} \leftarrow \eta\bar{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m} + (1-\eta)\hat{\boldsymbol{\Psi}}^h_{\ell,\Phi(s)m}$

until $t = t_{max}$

Make hard decision

It is also worth-noting that the number of iterations is fixed here to $t_{\max}$ only for the sake of the complexity analysis to be offered later. In practice, the process can be terminated at a fewer (also adaptively-determined) number of iterations, resulting in lower total complexity. The possibility of reducing the number of iterations is studied later via the convergence behavior of the algorithm, where it is shown that approximately 9 iterations are sufficient for convergence, regardless of SNR levels.

3.4.3 Performance Assessment

In this section, we assess the estimation performance of the proposed bilinear inference method under various system setups. In particular, we consider the NMSE and activity error rate (AER) as key performance metrics to measure, respectively, the estimation accuracy of channel coefficients, as well as user activity indicators.

The NMSE and AER are respectively defined as

$$\text{NMSE} \triangleq \frac{\|\boldsymbol{H}\boldsymbol{A} - \hat{\boldsymbol{H}}\hat{\boldsymbol{A}}\|_{\text{F}}^2}{\|\boldsymbol{H}\boldsymbol{A}\|_{\text{F}}^2}, \tag{3.80}$$

$$\text{AER} \triangleq \frac{|\mathcal{A} \setminus \hat{\mathcal{A}}|}{M} \leq 1, \tag{3.81}$$

where $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{A}}$ denoting estimated channel and user activity matrices, respectively, $\mathcal{A}$ denotes the true activity index set, $|\cdot|$ denotes the cardinality of a given set, and the operator $\setminus$ denotes the relative complement, such that $|\mathcal{A}| \leq M$, $|\hat{\mathcal{A}}| \leq M$, and $|\mathcal{A} \setminus \hat{\mathcal{A}}| \leq M$.

Throughout the section, the following parameters are utilized, unless specified otherwise. The number of total antenna elements and sub-arrays are respectively assumed to be $N = 400$ and $S = 100$, indicating that each sub-array possesses $N_s = 4$ antenna elements. This setup can be interpreted as an XL-MIMO system consisting of multiple sub-arrays with each being a 2×2 patch antenna array, for instance. The total number of time indices and potential users is set to $L = \{50, 70\}$ and $M = 200$, respectively. The user activity ratio is assumed to be $\lambda = 0.1$, while the number of active users at each channel realization is modeled as a binomial random variable with mean λM. The variance of channel coefficients is assumed to be identical and modeled as $\phi = \phi_{sm} = 1/M$ for all m and s, whereas different models for the non-stationarity phenomena are considered.

As for the algorithmic parameters, the maximum number of iterations is assumed to be $t_{max} = 32$, while the damping factor η is set to 0.5. The sub-array activity indicator ϕ is automatically learned over iterations via the EM framework presented above. It is assumed that initial estimates (*i.e.*, $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{\Psi}}^h$) are obtained via the low-complexity multiple measurement approximate belief propagation (MMVABP) algorithm.

Remarks on Complexity

Before presenting results on the actual performance of the proposed bilinear inference method for JACE, we analyze its computational complexity in terms of the number of flops required at each iteration of the algorithm. Since all the calculations in Algorithm 5 are scalar-by-scalar and the required inverse operation is performed with a diagonal matrix, the number of multiplication, division, subtraction, and addition operations is of order $\mathcal{O}(NML)$, which is linear with respect to each of available resource dimensions. In Algorithm 5, the number of iterations is set to t_{max}, implying that the total complexity is of order $\mathcal{O}(t_{max}NML)$.

Similarly, the computational complexity of existing linear inference algorithms such as those proposed in [114, 115, 152] is also of order $\mathcal{O}(t_{max}^{\text{SotA}}NML)$, where t_{max}^{SotA} is the total number of iterations of existing linear inference algorithms. And since both the proposed and existing JACE algorithms are based on the Bayesian message passing approach, the total number of iterations until convergence required by our method is comparable to those of existing alternatives ($i.e. t_{max} \approx t_{max}^{\text{SotA}}$), such that we can conclude that the proposed algorithm has the same order of complexity of existing JACE methods.

It will be shown in the sequel, however, that the proposed method outperform existing alternatives in terms of estimation accuracy, measured by NMSE and AER as defined in equations (3.80) and (3.81).

Uniformly Random Non-Stationarity

Aiming at evaluating the fundamental performance improvement attained by the proposed JACE algorithm, we first consider in this subsection an XL-MIMO system subjected to uniformly random sub-array activity pattern. In other words, the sub-array activity indicators p_{ms} for all m and s are independently generated as a Bernoulli random variable, with the corresponding mean ϕ_{sm} set to be $\phi_{sm} = 0.2$, such that the number of the total active sub-arrays at each channel realization follows the Binomial distribution, and 20% of the total subarrays are active at each channel realization in an average sense.

For the sake of comparison, we consider two state-of-the-art methods, namely the conventional linear MMSE estimator, and an MMVABP scheme, which is a generalization of the MMV-AMP algorithm of [115]. Comparing these three algorithms highlights performance gains due to awareness both to column-wise sparsity in the channel matrix resulting from grant free access, and to block-wise sparsity of active columns of the channel matrix, resulting from spatial non-stationarity.

For instance, consider the comparison of NMSE performances of the three distinct algorithms as a function of SNR in decibels as shown in Figure 3.11, for different

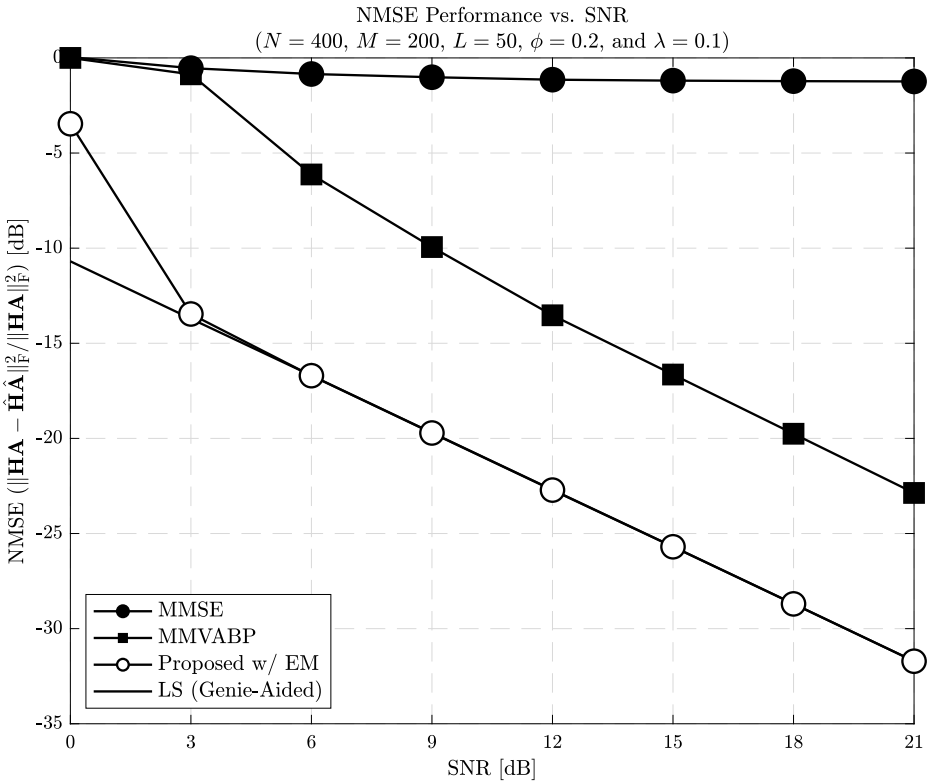
pilot lengths (*i.e.*, $L \in \{50, 70\}$), with pilot sequences designed via the QCSIDCO algorithm [131, 133, 153] in order to mitigate pilot contamination due to the non-orthogonal structure of $\boldsymbol{X} \in \mathbb{C}^{M \times L}$ with $M \gg L$. For the sake of a further reference, we also include curves (in solid line without markers) corresponding to lower-bounding NMSE performances obtained by the LS estimator aided by a genie, *i.e* with perfect knowledge of active user and sub-array activity indicators.

The figure clearly illustrates the impact of the two distinct factors which impose structured sparsity upon the channel matrix. In particular, it is found that regardless of the length of the pilot sequence, the MMSE estimator suffers from a high error floor in terms of its NMSE performance, while the MMVABP algorithm improves as the SNR increases. The gains of MMVABP over the MMSE method is due to the awareness to column-wise sparsity in the channel matrix – *i.e.*, awareness to user activity – which the MMVABP method incorporates, while the MMSE method does not. It is also found, however, that a large gap exists between the performance of MMVABP and the lower-bound. In comparison to the latter two methods, the proposed algorithm exhibits a substantial gain over the MMVABP approach, thanks to the fact that the proposed technique incorporates awareness not only to user activity, but also to the sub-array activity caused by spatial non-stationarity. As a result, the proposed method is found to actually reach the lower bound over a wide SNR range and starting from relatively low SNRs.

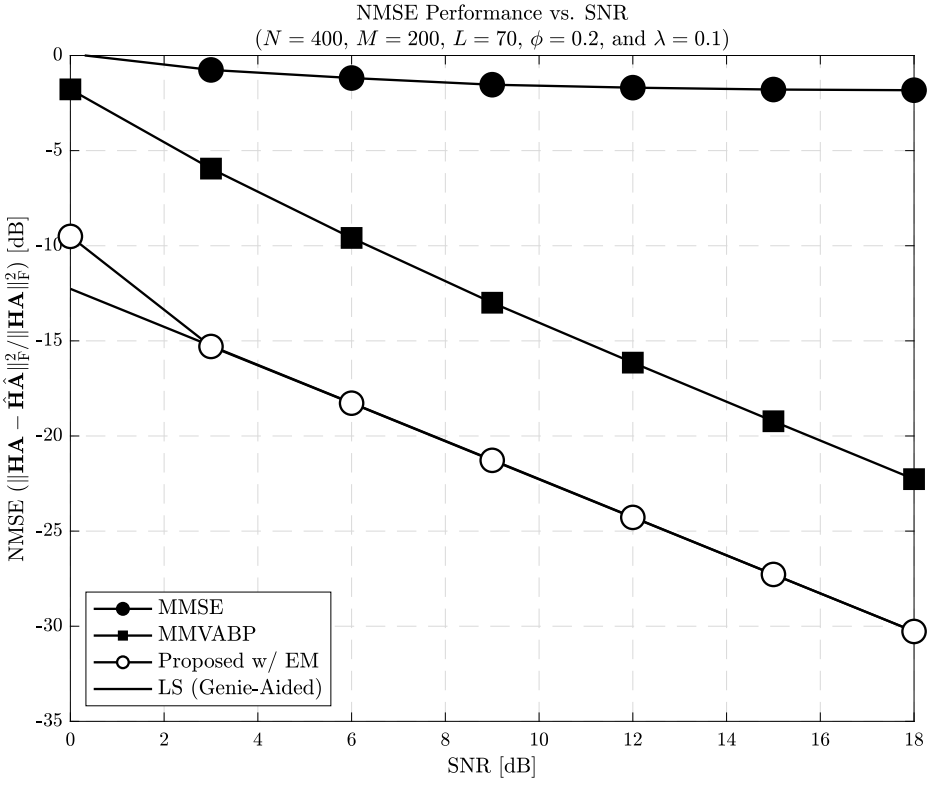
From these observations, one may conclude that the gain between the MMSE and the MMVABP methods results from awareness to user activity, while the gain between MMVABP and the proposed method is due to awareness to sub-array activity. It is also worth-mentioning that the sub-array activity indicators ϕ_{sm} are automatically learned for each channel realization via the EM framework presented above, such that estimating such parameters before transmission is not necessary, contributing to improving the efficiency of the XL-MIMO system.

Next, we consider the AER performances of the proposed and the best state-of-the-art methods (namely, the MMVABP), omitting results for the MMSE scheme as it was found to be inferior to the latter. In addition, notice that showing Genie-aided lower bounding results is not useful, since the the Genie-aided LS estimator has perfect knowledge of user and sub-array activities, such that AER performance is always 0.

The results are shown in Figure 3.12 for different pilot lengths as done in Figure 3.11. As expected, but interestingly, it is found that the proposed algorithm significantly outperforms MMVABP also in terms of the AER performance. For instance, it can be seen that the relative AER gain of the proposed method over MMVABP is approximately 4 [dB] in SNR at 10^{-5} of AER for both short and long pilot scenarios. In addition, it is interesting to see that the gradient of the AER curve of the proposed algorithm is steeper than that of the MMVABP approach.

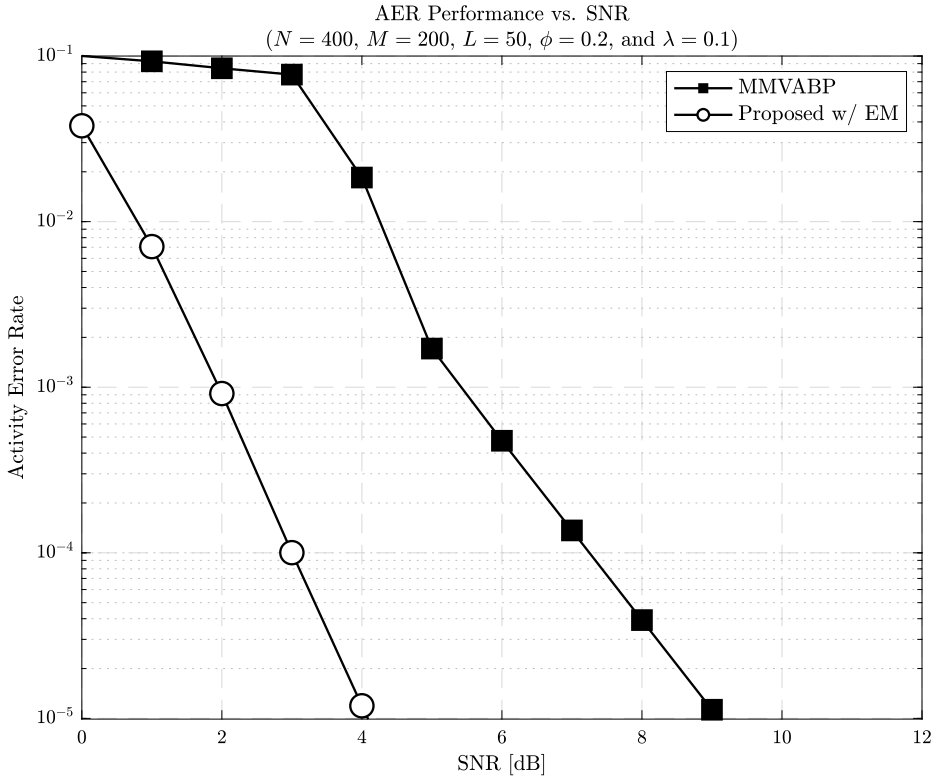


(a) $L = 50$.

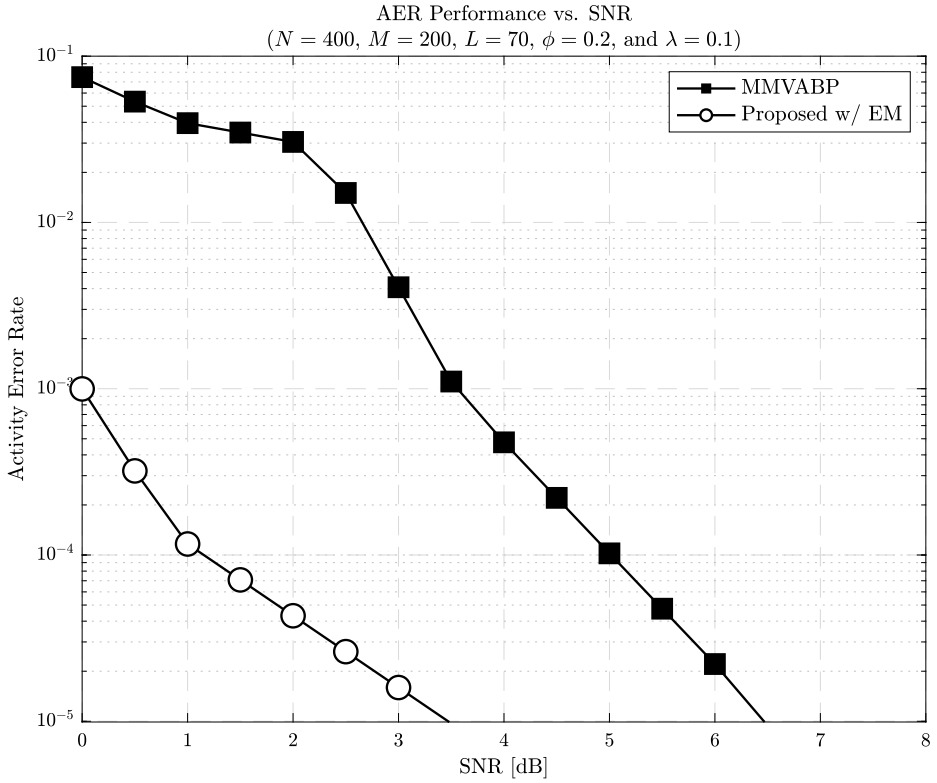


(b) $L = 70$

Figure 3.11: NMSE performance with respect to SNR with $N = 400$ and $M = 200$ for different pilot lengths. ©2022 IEEE

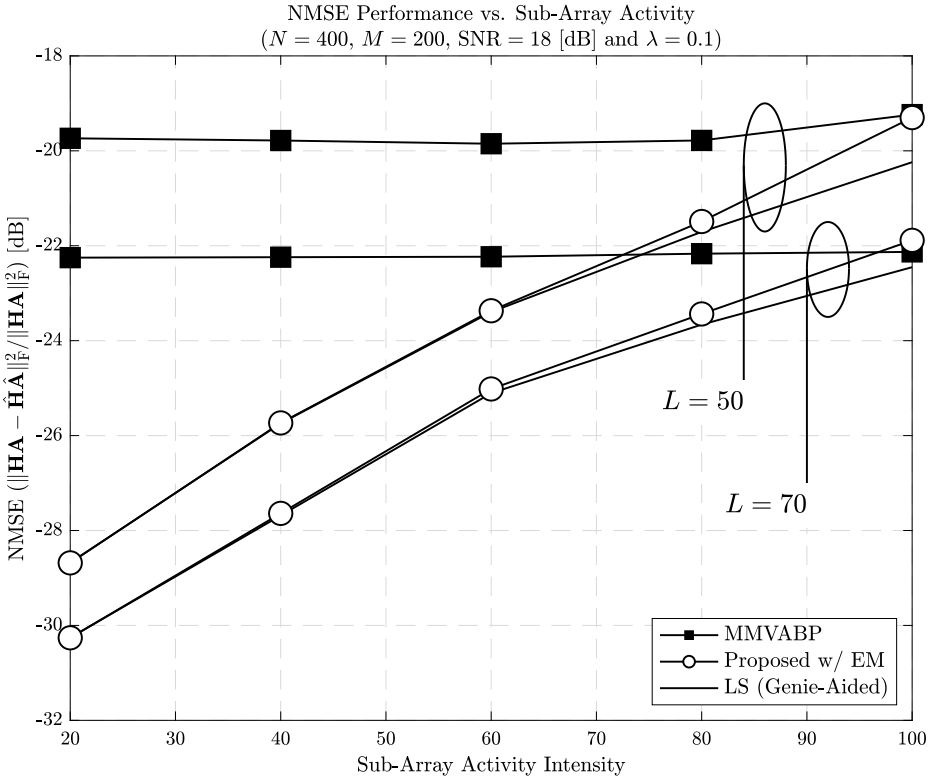


(a) $L = 50$.

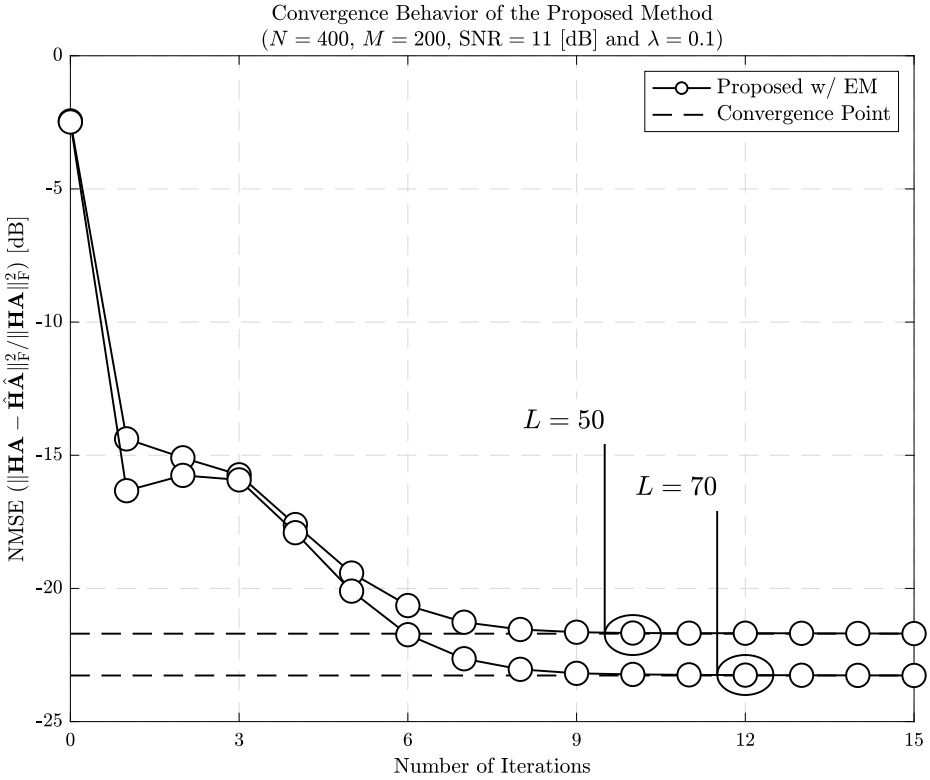


(b) $L = 70$.

Figure 3.12: AER performance with respect to SNR with $N = 400$ and $M = 200$ for different pilot lengths. ©2022 IEEE



(a) Resilience of the proposed algorithm against different sub-array activity indicators in percentage.



(b) Convergence behavior of the proposed algorithm with respect to the number of algorithmic iterations.

Figure 3.13: Convergence and scalability of the proposed method. ©2022 IEEE

Having clarified the NMSE and AER gains of the proposed algorithm, we turn our attention to resilience and convergence aspects of the proposed algorithm. To that end, we first compare in Figure 3.13a the NMSE performance of the MMVABP, the Genie-aided LS, and the proposed estimators as a function of the sub-array activity indicators. We remark that the channel is *stationary* at 100% sub-array intensity (*i.e.*, at the right edge of the figure), while non-stationarity effects becomes severer as the sub-array intensity decreases. The results suggest that the proposed algorithm is a generalization of the MMVABP method, capturing effects not only from the user activity but also from the sub-array activity. Thus, the performance of the proposed algorithm approaches that of MMVABP in case of a stationary channel (*i.e.*, $\phi_{sm} = 1$), while offering significant gains over the latter as the non-stationarity increases.

Finally, in Figure 3.13b, we show the convergence behavior of the proposed algorithm as a function of the number of the algorithmic iterations for different pilot lengths. It is shown in the figure that although we assume a relatively large value for the maximum number of iterations ($t_{max} = 32$), the algorithm converges within 10 iterations under both pilot lengths considered. This implies that the total complexity of the method can be further reduced (by more than a half) by setting a certain convergence criterion.

Matérn-Cluster Point Process Based Non-Stationarity

The comparison results shown in the previous subsections serve the purpose of quantifying the gains achievable by the proposed method over state-of-the-art alternatives, which stem in particular from the ability of the contributed scheme to detect both user and sub-array activity.

It can be argued, however, that the uniformly random sub-array activity pattern is somewhat artificial, since in realistic scenarios, such patterns are characterized by VRs. Indeed, in practice the likelihood of activation of a certain sub-array is highly correlated with that of neighboring sub-arrays, such that VRs tend to occur in clusters.

In this subsection, we therefore repeat the experiments reported above, utilizing this time a stochastic geometry approach to model the aforementioned geometrical correlation among the activity indicators of sub-arrays, so as to mimic the cluster-like nature of VRs, as illustrated in [118, Fig. 2]. To this end, we consider the Matérn-cluster point process (MCPP) – not to be confused with Matérn hardcore point process – to model the non-stationarity effects in XL-MIMO systems [154, 155].

In order to bring MCPP into the simulation setup, we consider a rectangular area in which sub-arrays are placed following an equispaced grid. Within this area, MCPP is leveraged to generate random clusters with a constant radius r and centers following a homogeneous Poisson point process (PPP) with an intensity μ. Each cluster generated

by MCPP is regarded as a VR, and therefore, sub-arrays located in the clusters are considered active, whereas sub-arrays located outside the clusters are assumed to be inactive.

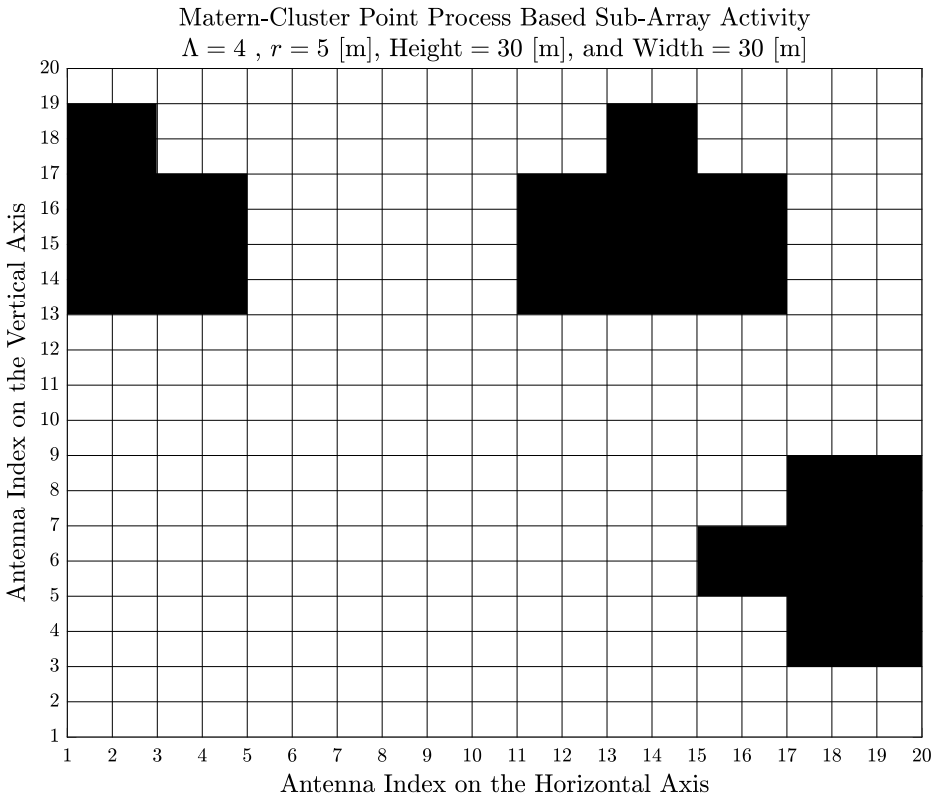
To visualize the difference between MCPP-based and uniformly random non stationarity models, we offer in Figure 3.14a and 3.14b a comparison of sub-array activity patterns for the two different models for a given realization, with the number of active antennas set to be identical in both cases. In the figures, white and black squares indicate inactive and active antennas, respectively, where we assumed a 2×2 square sub-array, Height = 30 [m], Width = 30 [m], $\mu = 4$ and $r = 5$. One can observe from the figures that the MCPP-based approach clearly illustrates clustered VRs, capturing more realistically the behavior of the non-stationarity, while the uniformly random counterpart shows a more scattered distribution of VRs.

With the stochastic-geometric VR generation model described, we proceed to the performance assessment of the JACE algorithms under this MCPP-based non-stationarity model. In this section, we evaluate the estimation performance of the proposed method in comparison with the two state-of-the-art estimators as well as the Genie-aided ideal performance for different cluster setups, by considering different cluster intensities μ and the radius r and studying the impact of both parameters on the detection performance.

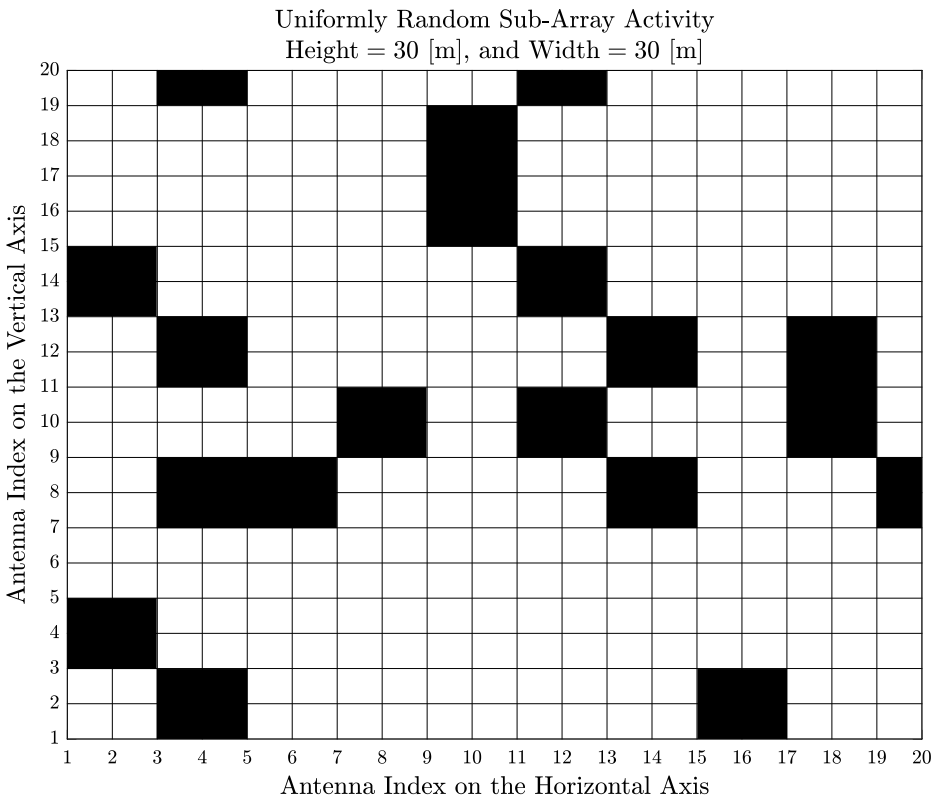
First, we compare in Figure 3.15 the NMSE performances of the three distinct detection algorithms for different cluster intensities μ and a fixed radius size $r = 5$ [m], with the Genie-aided ideal performance also included for reference. As expected based on the previous results of Figures 3.11 and 3.13a, the proposed method outperforms both the MMVABP and the conventional linear MMSE methods, although the performance gain diminishes slightly as the cluster intensity increases, which is expected since with for larger cluster intensities the number of active sub-arrays itself grows. It is also observed that once again the proposed algorithm reaches the Genie-aided ideal performance for a wide range of SNR regardless of the cluster intensity level.

Next, we compare in Figure 3.16 the AER performance of the proposed algorithm against that of the MMVABP scheme, again for different cluster intensity levels and as a function of SNR. The figure confirms that the proposed method is effective also in terms of the AER performance.

Having shown the effectiveness of the proposed algorithm even in case of clustered sub-array activity, we finally assess in Figures 3.17a and 3.17b the NMSE performance of the proposed algorithm as a function of cluster intensity μ and radius r, respectively, for two different SNR levels.

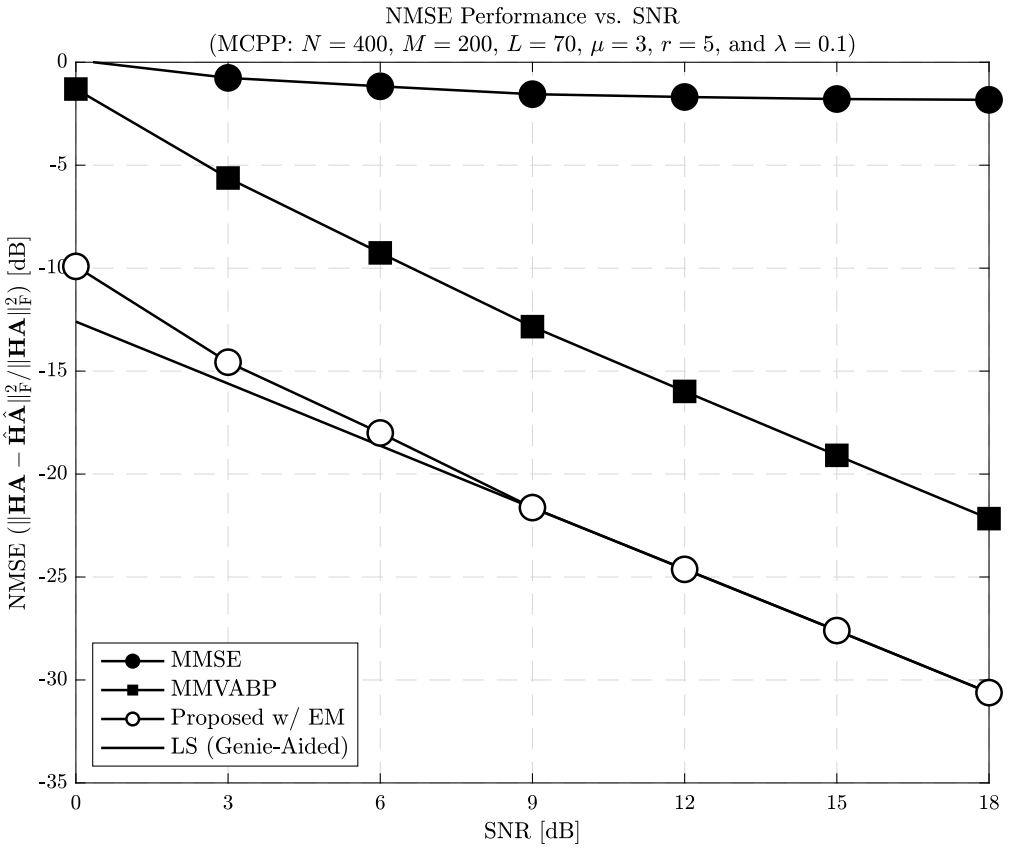


(a) MCPP-based subarray activity.

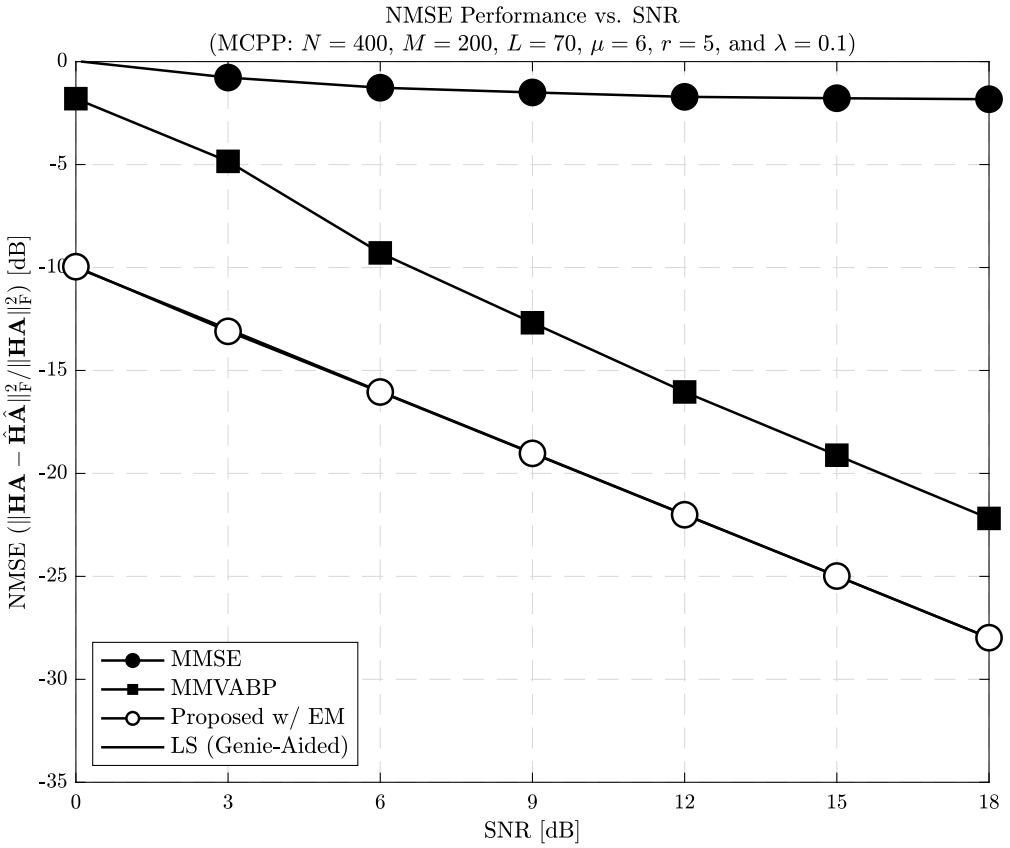


(b) Uniformly random subarray activity.

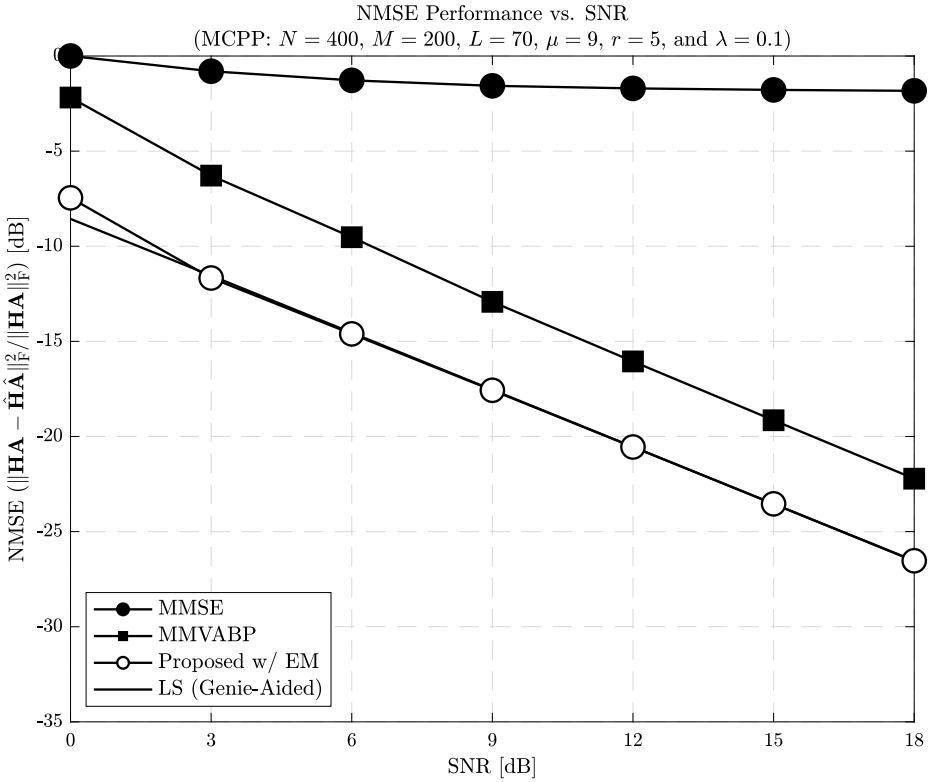
Figure 3.14: Comparison between MCPP and PPP based sub-array activity models. ©2022 IEEE


NMSE Performance vs. SNR
(MCPP: $N = 400$, $M = 200$, $L = 70$, $\mu = 3$, $r = 5$, and $\lambda = 0.1$)
NMSE ($\|\mathbf{HA} - \hat{\mathbf{H}}\hat{\mathbf{A}}\|_{\mathrm{F}}^2/\|\mathbf{HA}\|_{\mathrm{F}}^2$) [dB]
SNR [dB]
MMSE
MMVABP
Proposed w/ EM
LS (Genie-Aided)

(a) $\mu = 3$.

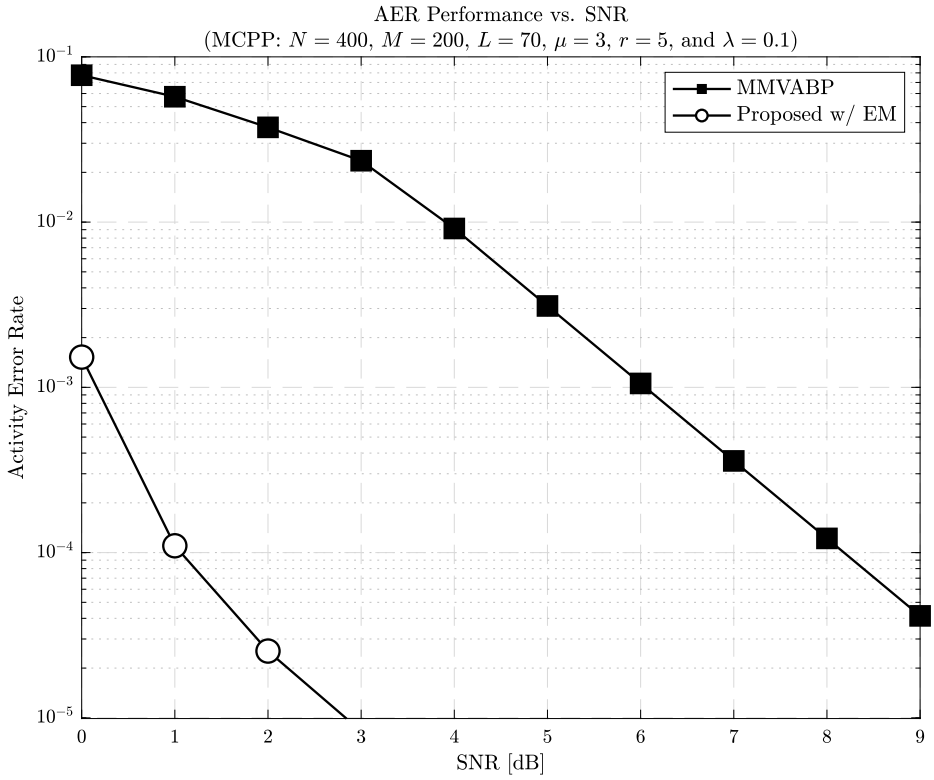

NMSE Performance vs. SNR
(MCPP: $N = 400$, $M = 200$, $L = 70$, $\mu = 6$, $r = 5$, and $\lambda = 0.1$)
NMSE ($\|\mathbf{HA} - \hat{\mathbf{H}}\hat{\mathbf{A}}\|_{\mathrm{F}}^2/\|\mathbf{HA}\|_{\mathrm{F}}^2$) [dB]
SNR [dB]
MMSE
MMVABP
Proposed w/ EM
LS (Genie-Aided)

(b) $\mu = 6$.

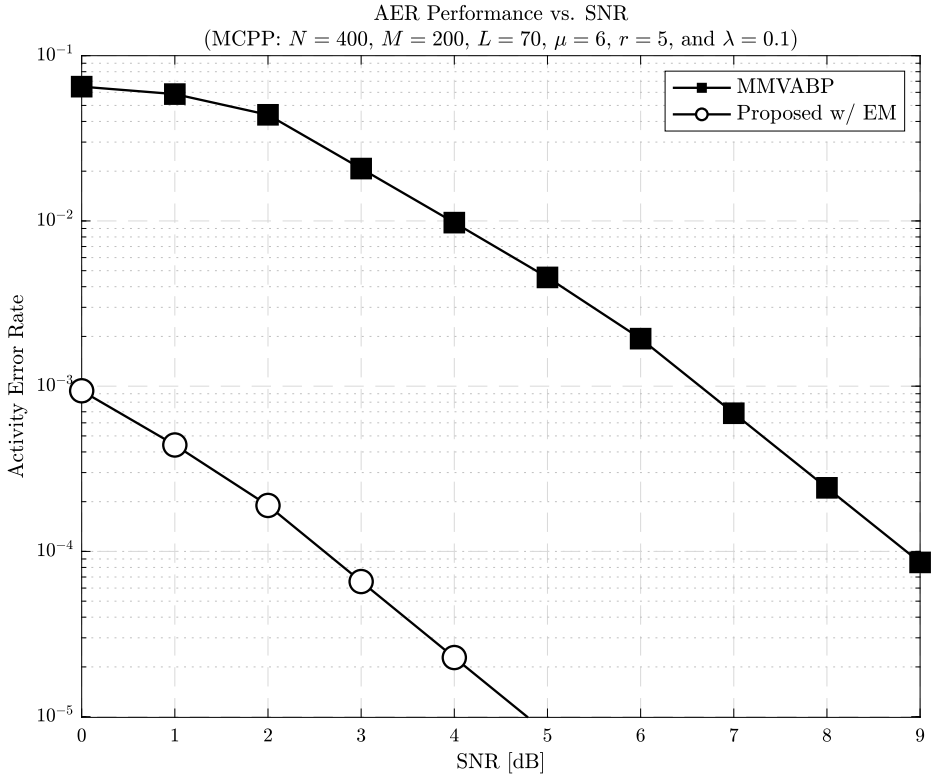


(c) $\mu = 9$.

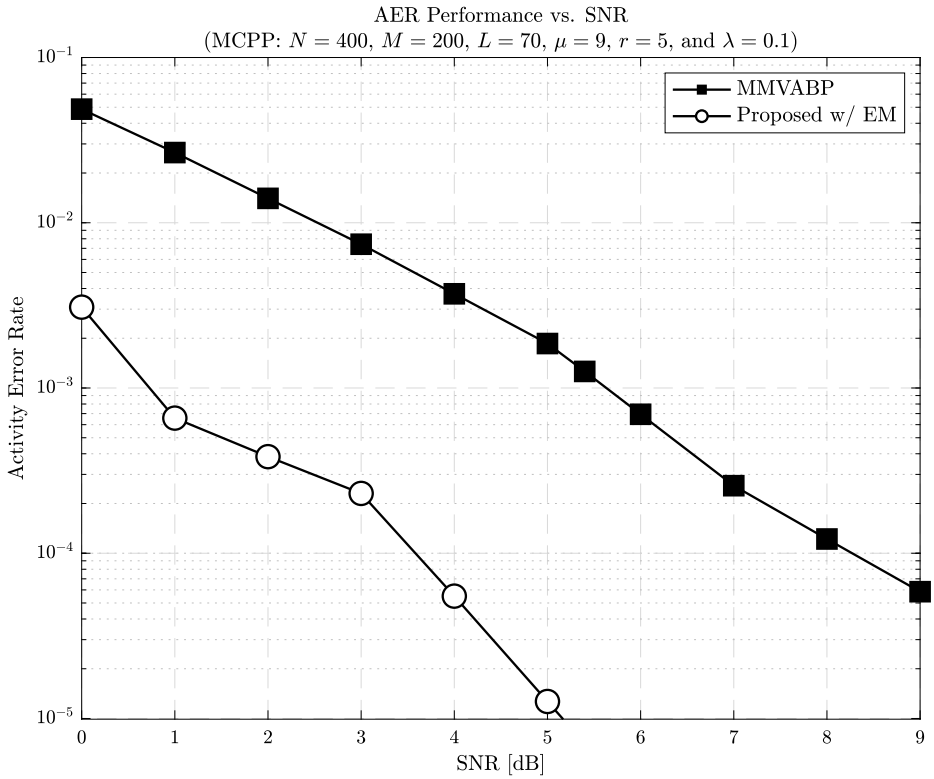
Figure 3.15: NMSE Performance with respect to SNR with $N = 400$, $M = 200$, and $L = 70$ with MCPP for different μ. ©2022 IEEE



(a) $\mu = 3$.

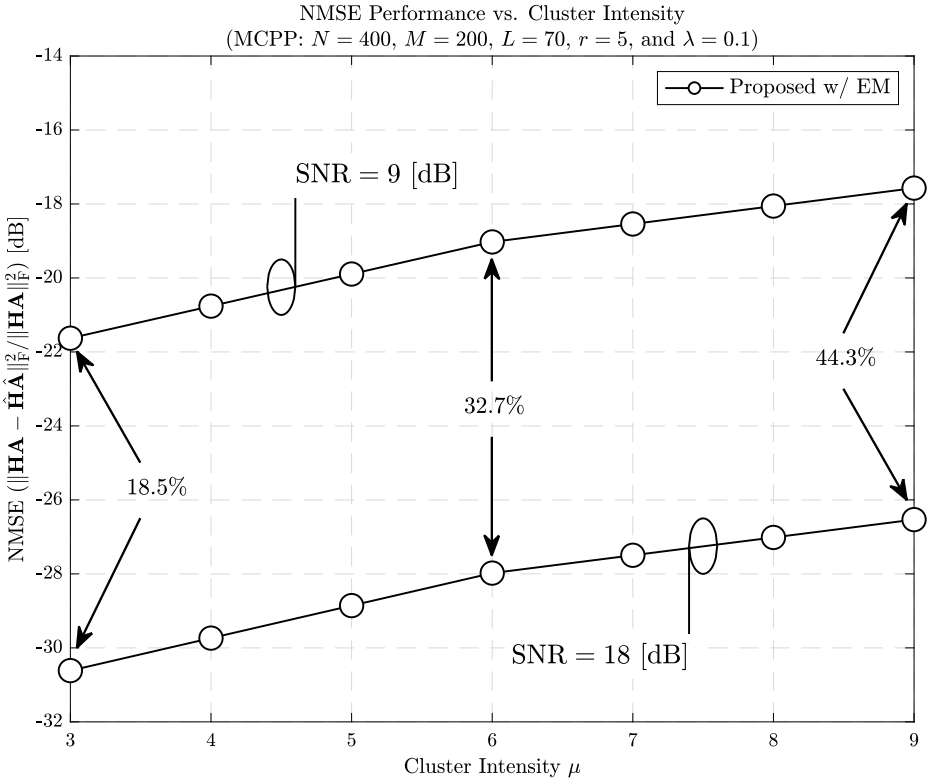


(b) $\mu = 6$.

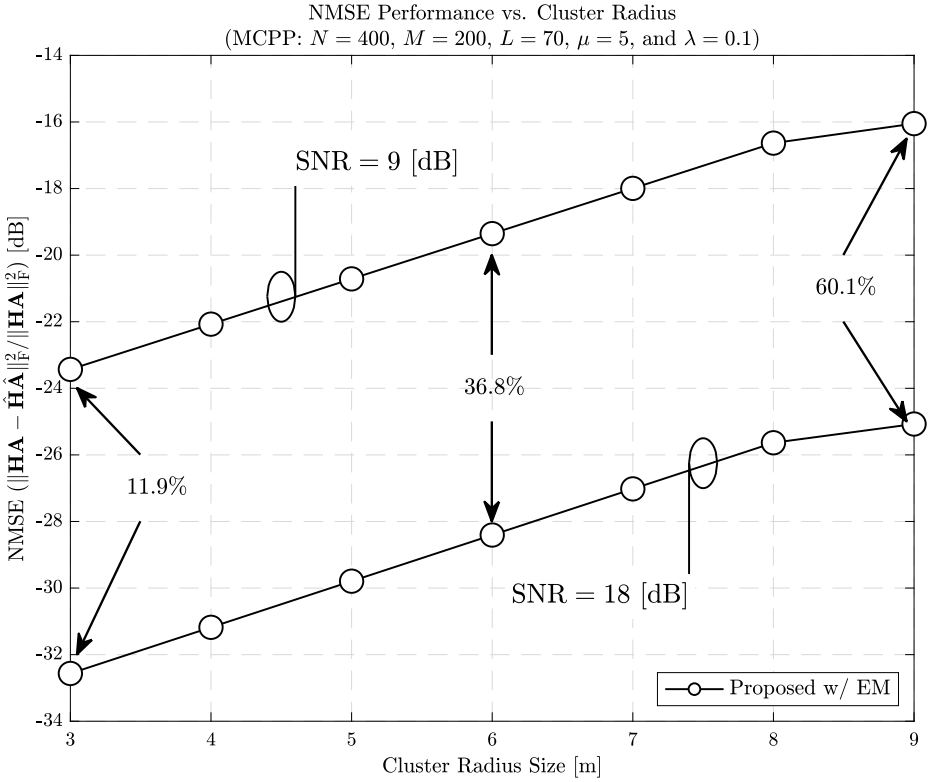


(c) $\mu = 9$.

Figure 3.16: AER Performance with respect to SNR with $N = 400$, $M = 200$, and $L = 70$ with MCPP for different μ. ©2022 IEEE



(a) NMSE performance of the proposed method for different cluster intensities. The actual array activity ratios for $\mu = 3, 6$ and 9 are annotated by allows.



(b) NMSE performance of the proposed method for different cluster radius sizes. The actual array activity ratios for $\mu = 3, 6$ and 9 are annotated by allows.

Figure 3.17: NMSE performance of the proposed algorithm for different cluster-related parameters. ©2022 IEEE

For ease of interpretation of the results, the actual sub-array activity ratios, defined as the ratio between the number of active sub-arrays and the total number of sub-arrays, are annotated by arrows for $\mu = 3, 6$ and 9. The figures illustrate the fact that the performance of the proposed JACE algorithm is mostly dependent on the actual activity indicator, rather than the shape of clusters, implying an inherent robustness against the nature of the spatial non-stationarity.

3.5 Conclusion

In this chapter, we proposed two joint estimation algorithms for the overhead reduction problem in distributed MIMO architectures. The first algorithm developed for CF-MIMO systems aims at the per-user throughput bottleneck issue raised in the related literature, while reducing the latency of massive uplink channels by jointly estimating AUD, CE, and MUD. In contrast, the second algorithm intends to tackle the spatial non-stationarity in XL-MIMO systems, while jointly addressing the same overhead reduction problem, leading to a joint estimation problem of AUD, CE, and active sub-arrays (*i.e.*, VRs). Two distinct bilinear Bayesian inference algorithms are proposed as the key enabler for each problem, which are designed by tailoring the BiGaBP framework. Also, new active user identification policies are proposed for each problem so as to capture their spacial features.

The performance of the proposed algorithms is evaluated and compared via quantitative computer simulations against state-of-the-art methods proposed in the grant-free literature. A set of the results shown in this chapter demonstrated the efficacy of the proposed methods in terms of its CE, AUD, and/or MUD detection performance.

As for possible future works, the proposed JACDE algorithm for CF-MIMO systems can be integrated with channel coding schemes. For such an extension, the soft outputs to the subsequent decoder can be directly computed based on the messgae passing mechanism proposed in this chapter. An open and debatable question is, however, whether active detection should be performed before or after the channel decoder. In addition, the design of the detector output with the active detection information is worth considering for coded grant-free systems.

Regarding that of the second algorithm, an interesting extension is to incorporate the data detection part, which is difficult to tackle with the present bilinear framework as such a problem results in a multilinear estimation problem. To address this difficulty, one can consider to extend the bilinear framework to a multilinear alternative by tailoring the message passing rules. Another alternative is to leverage a more flexible distribution instead of the Gaussian approximation such that the spatial non-stationarity and CSI acquisition are jointly treated as a single variable, which can be addressed by a bilinear inference algorithm.

Chapter 4

Blockage-Robust Beamforming for Reliable Millimeter Wave Communications

In this chapter, we turn our attention to the reliability issue in mmWave systems suffering from random path blockage due to the susceptibility of mmWave channels, proposing a new framework to minimize the total QoS violation among downlink users. To elaborate, we propose a stochastic optimization based approach to design a robust waveform to combat the susceptibility of mmWave channels to random path blockage. To begin with, we first formulate the outage minimization problem that can be seen as an empirical risk minimization (ERM) problem, which is then tackled via a stochastic gradient descent method so as to optimize the fully-digital transmit beamforming vectors. Then, we generalize the latter to a partially-connected hybrid beamforming architecture with the aim of reducing the implementation and hardware costs. The simulation results demonstrate that the proposed framework possesses robustness against such blockages for a wide range of blockage probabilities. Furthermore, the proposed hybrid mechanism is also shown to be close to its fully-digital counterpart in terms of the outage probability, while significantly reducing the number of RF chains.

Part of this chapter is reprinted and enhanced from the following publication:

- Hiroki Iimori, Giuseppe Thadeu Freitas de Abreu, Omid Taghizadeh, Razvan-Andrei Stoica, Takanori Hara, and Koji Ishibashi: "A Stochastic Gradient Descent Approach for Hybrid MmWave Beamforming with Blockage and CSI-Error Robustness," *IEEE Access*, vol. 9, pp. 74471–74487, May 2021.
- Hiroki Iimori, Giuseppe Thadeu Freitas de Abreu, Omid Taghizadeh, Razvan-Andrei Stoica, Takanori Hara and Koji Ishibashi,: "Stochastic Learning Robust Beamforming for Millimeter-Wave Systems with Path Blockage," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1557–1561, Sept. 2020.

4.1 Background and Contributions

The ever-growing demands for data-rate and massive wireless connectivity have driven the 5G standard to incorporate technologies that exploit the mmWave spectrum available in the 24–300 [GHz] bands [156–158]. This trend is expected to continue, with 5G New Radio (NR) aiming to support spectrum bands up to 71 [GHz] in Release 17. Higher frequencies in the sub-Terahertz (*i.e.* > 100 GHz) bands are planned to be added [159].

MmWave systems have much wider bandwidths than sub-6GHz, and they enable highly directional communications owing to the dense packing of antenna elements, which yields numerically large arrays of small physical dimensions, such that MIMO and beamforming techniques play key roles in mmWave technology [160]. Despite the advantages of higher achievable throughputs (data rates) and improved radio access (user capacity), mmWave systems suffer from communications-impairing effects such as increased path loss, higher atmospheric absorption, higher sensitivity to phase noise, and random path blockage [161–165]. These phenomena pose challenges to the practical implementation of mmWave, which over the years has attracted the attention of the research community, leading to several important advancements.

Early contributions concentrated on mitigating the path loss issue by exploiting the high directivity of mmWave antenna arrays [166, 167] and therefore aimed at taking advantage of the sparsity of the mmWave channel. To this end, novel sparse signal processing methods were developed, albeit under the assumption of perfect CSI and for a single-user single-carrier setup. These designs were then extended to multi-user and multi-carrier systems by incorporating fully-connected hybrid architectures, where the RF threads are connected to all the phase shifters equipped, *e.g.*, [168]. It was shown that beamforming can take advantage of the homoscedasticity and variance correlations of channel covariance across subcarriers.

Under the argument that accurate RF components at mmWave bands are expensive, cost reduction via the design of partially connected hybrid beamformers such as those in [169, 170] has been motivated. The same argument that lower-cost RF components lead to various imperfections, which in turn can be counterweighted by robust hybrid designs, motivated works such as [171] in which the authors proposed that the alternating maximization of a smooth SNR-driven optimization problem can be solved via orthogonal matching pursuit (OMP), for the realization of transceivers and amplify-and-forward (AF) relays with CSI uncertainty; in [75], a transmit beamforming scheme was designed via a fractional programming approach and was matched with a MMSE receive beamformer to mitigate hardware and CSI imperfection; and [172] aimed to maximize the sum-rate under imperfect CSI feedback, employing a robust convolutional neural-network approach.

Despite the contributions that helped solidify robust hybrid beamforming as the prevailing approach that ensures the feasibility of mmWave systems, the random path blockage problem inherent to the mmWave channel has not been addressed sufficiently compared to the other issues discussed previously. The fundamental challenge remaining is that mmWave signals are susceptible to random blockage of propagation paths with probabilities ranging between 20% and 60%, which if not counteracted can significantly detract from the high-throughputs that mmWave communication systems can potentially deliver [162, 163, 165].

An early proposal to combat such effects is the coordinated multipoint (CoMP) scheme described in [173], in which QoS was maintained despite path blockages by synchronizing transmissions from multiple BSs and APs. The CoMP approach is not only attractive in the context of mmWave systems owing to the reduced radio coverage, but it is also shown analytically to offer strong theoretical guarantees to achieve capacity in the presence of a path blockage [174], which, however, is unfortunately not accompanied by procedures for practical implementation.

This chapter intends to fill in the latter gap, which presents an extensive robust stochastic learning approach to minimize outage probability in mmWave CoMP systems subjected to random path blockage, thus effectively maintaining high-throughput service guarantees. To this end, it may be emphasized that the present work is compatible with some existing side-information-aided approaches such as the sub-6 [GHz] side-signaling assumed in [175, 176] or the visual (camera-based) information used in [177–179]. In addition, our contributions are also aligned with recent works on robust mmWave beamforming in which path blockage is assumed to be random and dealt-with dynamically, a few examples of which are discussed below.

In [180] a high-speed railway communications environment was considered, where it can be assumed that path blockages, detected during a first probing stage, remain constant during a subsequent transmission stage (*i.e.*, *coherent blockage assumption*). The article then proposed a corresponding two-stage (detect-and-avoid) mmWave hybrid beamforming mechanism based on greedy matching pursuit optimization, aimed at minimizing MMSE and maximizing sum-rate maximization (SRM) under the assumption of perfect CSI.

Besides the coherent blockage assumption, the latter contribution did not consider the benefit of CoMP transmission, which was previously shown in [173, 174] to be crucial for high-performing mmWave systems. In turn, [181] considered a robust CoMP setup under the assumption that blockage affects line-of-sight (LoS) paths in an incoherent manner. Subsequently, a greedy digital beamforming problem was formulated aimed at maximizing the system's minimum sum-rate, which was solved iteratively via convex approximations.

Contributions

In this chapter we contribute to this trending topic by proposing a stochastic framework for the design of outage-minimum robust mmWave CoMP systems, which has two main differences over the previous works. First, unlike the latter, which requires all conceivable QoS constraints to be considered thus leading to combinatorics[1], our proposed method is based on a stochastic optimization framework that enables us to avoid dealing with such combinatorial operations and yields decent solutions even in ill-conditioned scenarios. Secondly, unlike the sum-rate maximization problem considered in the previous works, the fundamental goal of robust beamforming design for mmWave systems subject to uncertain blockage is rather to minimize the required QoS violation (*i.e.*, outage probability).

In summary, the contributins offered in this chapter are as follows.

- We introduce a new stochastic-learning-based robust beamforming design framework for CoMP systems to combat unpredictable random blockages in mmWave channels. The complexity of our contributed methods does not scale with the number of active clusters, avoiding combinatorial computations in the optimization.

- Based on this framework, we first propose a novel full-digital beamforming design for CoMP mmWave systems subject to random blockages, showing its advantage over the conventinal SRM and maximum ratio transmission (MRT) methods.

- In realistic scenarios, the mmWave systems suffer not only from random path blockages but also CSI imperfection due to unavoidable channel estimation errors. In this context, we incorporate into a Bernoulli-Gaussian PDF the statistical features of both path blockages [161–163, 165, 166] and CSI errors [75, 171, 172], resulting in an integrated stochastic mmWave channel model that enables both challenges to be addressed simultaneously.

- Incorporating practical considerations into the preceding full-digital beamforming method, we then propose a robust virtually-configured partially-connected cooperative hybrid beamforming algorithm suitable to mitigate both path blockages and imperfect CSI in mmWave systems.

[1] In addition, the worst-case method requires that the target QoS be satisfied even in ill-conditioned scenarios, which may not admit a feasible solution in some realistic circumstances.

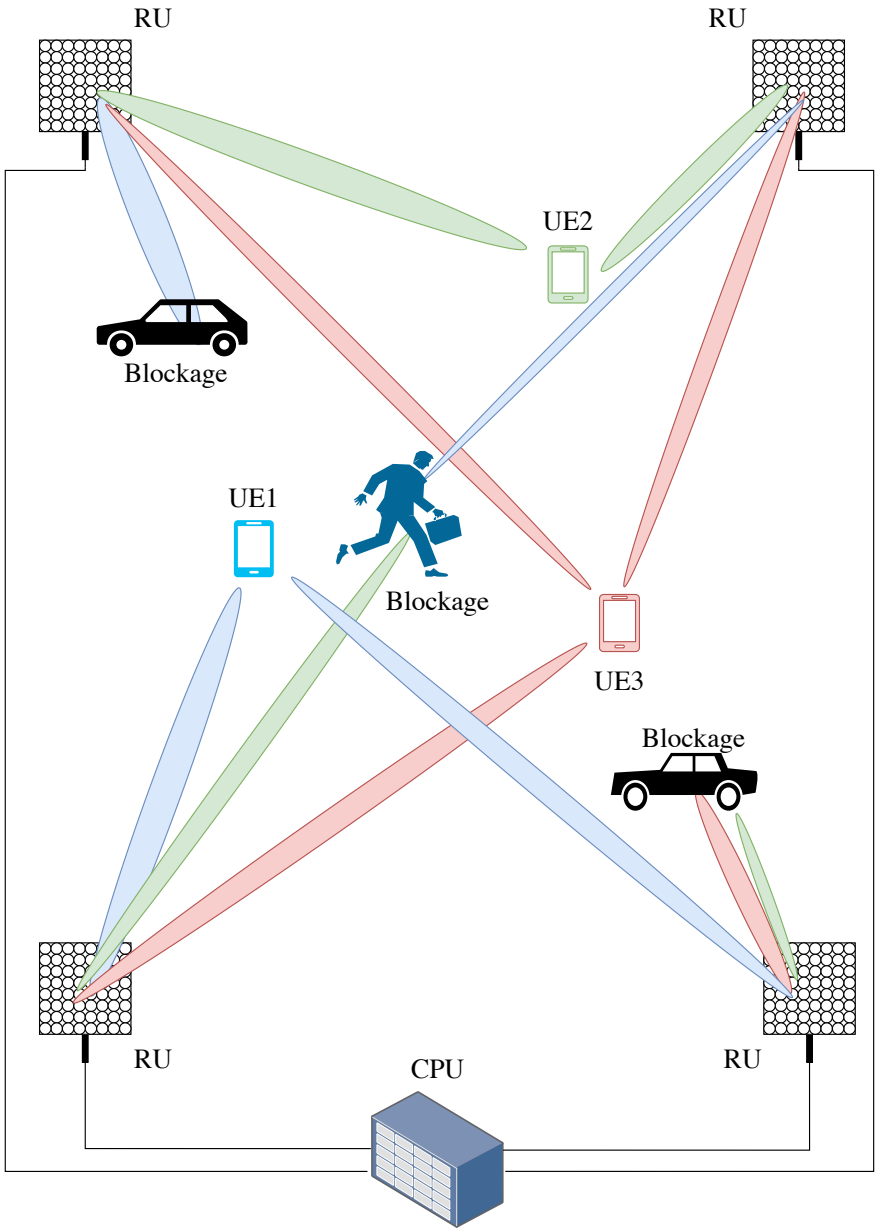


Figure 4.1: Illustration of the considered mmWave CoMP system subject to blockage.

4.2 Full Digital Beamforming

4.2.1 System Model

Consider the downlink of a single carrier CoMP narrowband mmWave system in which multiple synchronized BSs cooperatively serve single-antenna downlink users subject to unpredictable blockages, as shown in Figure 4.1. For such a scenario, let $b \in \mathcal{B} \triangleq \{1, 2, \ldots, B\}$ and $u \in \mathcal{U} \triangleq \{1, 2, \ldots, U\}$ denote the BS and downlink user indices, respectively, with B and U denoting the total number of BSs and downlink users.

It has been shown that unpredictable blockages occur with $20\% - 60\%$ probability, leading to significant loss in the achievable throughput. In order to abstract such non-ideal effects, we consider a probabilistic model, in which each path undergoes random and independent blockage, whose frequency of occurrence can however be obtained as side-information for the system optimization, as shown in [177]. In other words, the blockage probability of each component channel path can be assumed to be known at the BSs and employed in robust beamforming designs.

Following [160], it is assumed that the channel contains a random number $K_{b,u}$ of clusters, which is modeled as [161] $K_{b,u} \sim \max\left(1, \text{Possion}\left(\lambda\right)\right)$ with the intensity parameter λ, where one of the number of clusters $K_{b,u}$ corresponds to LoS and the rest of them are regarded as non-line-of-sight (NLoS). It is furthermore assumed that CSI is acquired and tracked continuously exploiting the reciprocity between uplink and downlink of standard time division duplex (TDD) systems, such that channel estimates can be modeled as[2]

$$\hat{\boldsymbol{h}}_{b,u} = \sqrt{\frac{1}{K_{b,u}}} \sum_{m=1}^{K_{b,u}} g_{b,u}^{m} \boldsymbol{a}_T(\phi_{b,u}^{m}), \tag{4.1}$$

where $\phi_{b,u}^{m}$ denotes the angles of departure (AoD) of the m-th cluster from the b-th BS towards the u-th downlink user and $\boldsymbol{a}_T(\phi_{b,u}^{m})$ is an array response vector at the transmitter, while $g_{b,u}^{m}$ is the associated channel gain modeled as $g_{b,u}^{m} \sim \mathcal{CN}\Big(0, 10^{\frac{-\text{PL}_{b,u}^{m}}{10}}\Big)$ with $\text{PL}_{b,u}^{m} = \alpha + 10\beta \log_{10}\left(d_{b,u}\right) + \xi$ [dB] in which $d_{b,u}$ is the distance (in meters) between the b-th BS and the u-th user and the parameters α, β and ξ are determined according to [161, Table I].

In spite of the knowledge of $\hat{\boldsymbol{h}}_{b,u}$, during actual downlink the system might be subjected to partial outage if and when any or some of the LoS and NLoS clusters become temporarily blocked, such that the actual channel between the b-th BS and the u-th user can then be modeled as

$$\boldsymbol{h}_{b,u} = \sqrt{\frac{1}{K_{b,u}}} \sum_{k=1}^{K_{b,u}} \omega_{b,u}^{k} g_{b,u}^{k} \boldsymbol{a}_T(\phi_{b,u}^{k}), \tag{4.2}$$

where $\omega_{b,u}^{k} \in \{0,1\}\ \forall k \in \{1,\ldots,K_{b,u}\}$ denotes a Bernoulli random variable with mean $p_{b,u}^{k}$ depicting the corresponding blockage probability[3].

Assuming that each BS is equipped with a uniform linear array (ULA) with N_t transmit antenna elements and half-wavelength spacing, let $\boldsymbol{f}_{b,u} \in \mathbb{C}^{N_t \times 1}$ denote the transmit beamforming vector from the b-th BS towards the u-th user subject to a

[2]For the sake of simplicity, in this section we first consider a perfect CSI scenario, which is later extended to an imperfect CSI counterpart.

[3]Although one may consider the case that some paths are blocked at the channel estimation and unblocked for the data transmission, we assumed that such scenarios are negligible as the channel would not simply be considered for data transmission due to the few number of mmWave paths [161].

maximum transmit power constraint $\sum_{u\in\mathcal{U}} \|\boldsymbol{f}_{b,u}\|_2^2 \leq P_{\max,b}$, such that the received signal y_u at the u-th downlink user can be written as

$$\begin{aligned} y_u &= \sum_{b\in\mathcal{B}} \boldsymbol{h}_{b,u}^{\mathrm{H}} \boldsymbol{f}_{b,u} x_u + \sum_{u'\in\mathcal{U}\setminus u} \sum_{b\in\mathcal{B}} \boldsymbol{h}_{b,u}^{\mathrm{H}} \boldsymbol{f}_{b,u'} x_{u'} + n_u \\ &= \boldsymbol{h}_u^{\mathrm{H}} \boldsymbol{f}_u x_u + \sum_{u'\in\mathcal{U}\setminus u} \boldsymbol{h}_u^{\mathrm{H}} \boldsymbol{f}_{u'} x_{u'} + n_u, \end{aligned} \tag{4.3}$$

where x_u is the unit-power transmit signal intended to the u-th user, n_u denotes i.i.d. circularly symmetric AWGN at the u-th user, *i.e.* $n_u \sim \mathcal{CN}(0, \sigma_u^2)$, and the vectors $\boldsymbol{h}_u$ and $\boldsymbol{f}_u$ are respectively defined for all $u \in \mathcal{U}$ as $\boldsymbol{h}_u \triangleq \left[\boldsymbol{h}_{1,u}^{\mathrm{T}}, \ldots, \boldsymbol{h}_{B,u}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\boldsymbol{f}_u \triangleq \left[\boldsymbol{f}_{1,u}^{\mathrm{T}}, \ldots, \boldsymbol{f}_{B,u}^{\mathrm{T}}\right]^{\mathrm{T}}$.

4.2.2 Problem Formulation

Given the system model described above, in this subsection we offer an optimization problem formulation corresponding to the mmWave robust CoMP downlink beamforming design subject to unpredictable path blockage, which does not rely on a standard deterministic robust optimization framework that imposes combinatorial operations.

In particular, our formulation consists of the following stochastic sum-outage-probability minimization problem:

$$\underset{\boldsymbol{f}}{\text{minimize}} \quad \sum_{u\in\mathcal{U}} \Pr\left\{\Gamma_u\left(\boldsymbol{h}_u, \boldsymbol{f}\right) < \gamma_u\right\} \tag{4.4a}$$

$$\text{subject to} \quad \sum_{u\in\mathcal{U}} \|\boldsymbol{f}_{b,u}\|_2^2 \leq P_{\max,b}\, \forall\, b, \tag{4.4b}$$

where γ_u denotes the target SINR for the u-th user and the corresponding effective SINR Γ_u can be written as

$$\Gamma_u\left(\boldsymbol{h}_u, \boldsymbol{f}\right) = \frac{|\boldsymbol{h}_u^{\mathrm{H}} \boldsymbol{f}_u|^2}{\sum_{u'\in\mathcal{U}\setminus u} |\boldsymbol{h}_u^{\mathrm{H}} \boldsymbol{f}_{u'}|^2 + \sigma_u^2}, \tag{4.5}$$

where $\boldsymbol{f} \triangleq \left[\boldsymbol{f}_1^{\mathrm{T}}, \ldots, \boldsymbol{f}_U^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{BUN_t \times 1}$.

Notice that equation (4.4) is stochastic and has only as many constraints as the number of BSs due to the power constraints. In addition, the objective of the mmWave optimization problem formulated in equation (4.4) is to preserve the system's continuity subject to QoS guarantees and under maximum transmit power constraints, which we argue is a more direct robust solution to the problem posed by random blockage.

4.2.3 Full Digital OutMin

In this section, we propose a novel robust beamforming algorithm based on stochastic learning, aiming at solving the intended nondeterministic optimization problem given in equation (4.4). To that end, we first introduce an indicator function $\mathbb{1}_{\Gamma_u(\boldsymbol{h}_u,\boldsymbol{f})<\gamma_u}$ defined as

$$\mathbb{1}_{\Gamma_u(\boldsymbol{h}_u,\boldsymbol{f})<\gamma_u} = \begin{cases} 1 & \text{if } \Gamma_u\left(\boldsymbol{h}_u, \boldsymbol{f}\right) < \gamma_u \\ 0 & \text{otherwise} \end{cases}, \tag{4.6}$$

which yields

$$\underset{\boldsymbol{f}}{\text{minimize}} \quad \sum_{u\in\mathcal{U}} \mathbb{E}_{\omega_{b,u}^m}\left[\mathbb{1}_{\Gamma_u(\boldsymbol{h}_u,\boldsymbol{f})<\gamma_u}\right] \tag{4.7a}$$

$$\text{subject to} \quad \sum_{u\in\mathcal{U}} \|\boldsymbol{f}_{b,u}\|_2^2 \leq P_{\max,b} \ \forall b. \tag{4.7b}$$

One may notice that equation (4.7) can be seen as an ERM problem [182], studied thoroughly in the machine learning and stochastic optimization literature and known to be solved efficiently via stochastic optimization approaches, which are widely adopted due to their complexity and memory requirements, as well as their easy-to-implement nature [183].

In light of the fact that the channel gains and corresponding AoDs are assumed to be known at BSs, as described in equations (4.1) and (4.2), we furthermore remark that possible combinations of blockage patterns due to $\omega_{b,u}^m$ can be randomly generated under the assumption that path blockage probabilities are somehow predictable [177] and can be utilized during stochastic optimization as part of the training set design.

Let $\tilde{\boldsymbol{h}}_u^m$ be the m-th data batch for $\boldsymbol{h}_u$ and $m \in \{1, 2, \ldots, M\}$ with M denoting the size of training data. Then, equation (4.7) can be further rewritten in an ERM fashion as

$$\underset{\boldsymbol{f}}{\text{minimize}} \quad \frac{1}{M}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}}\left[\mathbb{1}_{\Gamma_u(\tilde{\boldsymbol{h}}_u^m,\boldsymbol{f})<\gamma_u}\right] \tag{4.8a}$$

$$\text{subject to} \quad \sum_{u\in\mathcal{U}} \|\boldsymbol{f}_{b,u}\|_2^2 \leq P_{\max,b} \ \forall b. \tag{4.8b}$$

A universal technique to deal with variety of ERM problems such as that described by equation (4.8) is the gradient descent (GD) approach. This approach requires, however, the successive evaluation of sum-gradients which imposes high complexity especially in setups with large data sizes. In order to circumvent this issue, an efficient alternative to GD is stochastic gradient descent (SGD) [184], in which the solution at

each m-th iteration is updated based on the recursion

$$\boldsymbol{f}^{(m)} = \boldsymbol{f}^{(m-1)} - \alpha_m \sum_{u \in \mathcal{U}} \nabla \mathbb{1}_{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right) < \gamma_u}, \tag{4.9}$$

where α_m is a suitable stepsize at the m-th iteration.

Notice that it is difficult to directly compute the gradient direction in equation (4.9) due to the non-smoothness of the indicator function given by equation (4.6). Hence, we further introduce the generalized smooth hinge surrogate function [185], which is given by

$$\nu\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right) = \begin{cases} 0 & \text{if } 1 - \frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} < 0 \\ \frac{1}{2\varepsilon}\left(1 - \frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u}\right)^2 & \text{otherwise} \\ 1 - \frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} - \frac{\varepsilon}{2} & \text{if } 1 - \frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} > \varepsilon, \end{cases} \tag{4.10}$$

where $0 < \varepsilon \ll 1$ and its gradient with respect to $\boldsymbol{f}$ is

$$\nabla\nu\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right) = \begin{cases} 0 & \text{if } 1 - \frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} < 0 \\ \frac{\frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} - 1}{\varepsilon} \frac{\nabla\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} & \text{otherwise} \\ -\frac{\nabla\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} & \text{if } 1 - \frac{\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right)}{\gamma_u} > \varepsilon. \end{cases} \tag{4.11}$$

From the above, equation (4.8) can be rewritten as

$$\underset{\boldsymbol{f}}{\text{minimize}} \quad \frac{1}{M} \sum_{m=1}^{M} \sum_{u \in \mathcal{U}} \nu\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right) \tag{4.12a}$$

$$\text{subject to} \quad \sum_{u \in \mathcal{U}} \|\boldsymbol{f}_{b,u}\|_2^2 \leq P_{\max,b} \ \forall b, \tag{4.12b}$$

which can be solved by the projected SGD algorithm with the following update

$$\boldsymbol{f}^{(m)} = \boldsymbol{f}^{(m-1)} - \alpha_m \sum_{u \in \mathcal{U}} \nabla\nu\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}^{(m-1)}\right), \tag{4.13}$$

where $\boldsymbol{f}^{(m-1)}$ indicates the solution at $(m-1)$-th iteration.

Note that in equation (4.13), we omit the normalizing factor $1/M$ as it is independent from the optimal condition. In order to compute the gradient $\nabla\Gamma_u$ with respect to the stacked beamforming vector $\boldsymbol{f}$, it is necessary to equivalently reformulate the SINR formula given in equation (4.5) as

$$\Gamma_u\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\right) = \frac{\boldsymbol{f}^{\mathrm{H}} \tilde{\boldsymbol{H}}_u^m \boldsymbol{f}}{\boldsymbol{f}^{\mathrm{H}} \tilde{\boldsymbol{H}}_{\bar{u}}^m \boldsymbol{f} + \sigma_u^2}, \tag{4.14}$$

Algorithm 6 Full Digital OutMin

Inputs: Initial estimate: $\boldsymbol{f}^{(0)}$, Data set: $\mathcal{H}$
Outputs: Optimized beamforming vector: $\boldsymbol{f}$

1: Set $m = 1$.
2: **repeat**
3: Sample data batch $\tilde{\boldsymbol{h}}_u^m \ \forall u$ from $\mathcal{H}$.
4: Compute $\sum_{u\in\mathcal{U}} \nabla\nu\left(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}^{(m-1)}\right)$ from eq. (4.11).
5: Update $\boldsymbol{f}^{(m)}$ according to eq. (4.13).
6: Project $\boldsymbol{f}^{(m)}$ onto the feasible set (4.12b)
7: **until** $m = M$

with

$$\tilde{\boldsymbol{H}}_u^m = \mathrm{diag}\left(\boldsymbol{e}_u\right) \otimes \tilde{\boldsymbol{h}}_u^m \tilde{\boldsymbol{h}}_u^{m\mathrm{H}}, \tag{4.15a}$$

$$\tilde{\boldsymbol{H}}_{\bar{u}}^m = \mathrm{diag}\left(\bar{\boldsymbol{e}}_u\right) \otimes \tilde{\boldsymbol{h}}_u^m \tilde{\boldsymbol{h}}_u^{m\mathrm{H}}, \tag{4.15b}$$

where $\boldsymbol{e}_u \in \{0,1\}^{U\times 1}$ is the standard basis of length U in which only its u-th element is 1, $\bar{\boldsymbol{e}}_u$ denotes the ones' complement of $\boldsymbol{e}_u$ (*i.e.*, negation of $\boldsymbol{e}_u$) and $\otimes$ expresses the Kronecker product operator.

In light of all the above, the gradient $\nabla\Gamma_u\big(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\big)$ with respect to $\boldsymbol{f}$ can be finally written in closed-form as

$$\nabla\Gamma_u\big(\tilde{\boldsymbol{h}}_u^m, \boldsymbol{f}\big) = \frac{\tilde{\boldsymbol{H}}_u^m \boldsymbol{f}}{\boldsymbol{f}^{\mathrm{H}}\tilde{\boldsymbol{H}}_{\bar{u}}^m \boldsymbol{f} + \sigma_u^2} - \frac{\boldsymbol{f}^{\mathrm{H}}\tilde{\boldsymbol{H}}_u^m \boldsymbol{f}}{\left(\boldsymbol{f}^{\mathrm{H}}\tilde{\boldsymbol{H}}_{\bar{u}}^m \boldsymbol{f} + \sigma_u^2\right)^2}\tilde{\boldsymbol{H}}_{\bar{u}}^m \boldsymbol{f}. \tag{4.16}$$

A summary of the here-proposed SGD-based outage-minimum robust beamforming design for mmWave CoMP systems with unpredictable blockages, dubbed the SGD-OutMin for short, is offered as a pseudo code in Algorithm 6.

4.2.4 Performance Assessment

In this section we evaluate the proposed method via software simulations both in terms of achieved outage probability and of the corresponding effective throughput defined as $R_{\mathrm{eff},u} \triangleq W\mathbb{E}[\log_2(1+x)]$, where $x = \Gamma_u(\boldsymbol{h}_u, \boldsymbol{f})$ if $\Gamma_u(\boldsymbol{h}_u, \boldsymbol{f}) \geq \gamma_u$, $x = 0$ otherwise, and W is the sub-carrier bandwidth. Also we define $P_{\mathrm{out},u}$ being the outage probability of the u-th user.

As already remarked earlier, state-of-the-art SRM methods rely on a worst case optimization framework where all conceivable blockage patterns are taken into account as constraints so that the minimum throughput among such combinations can be

maximized. As a consequence, probabilistic blockage was considered thereby only at LoS for a fixed number of clusters, because the complexity of the approach makes it prohibitive to apply to a more general and stochastic model such as the one considered here.

In light of the above, we compare the performance of our proposed beamforming design with the well-known MRT beamforming (also known as conjugate beamforming), SRM beamforming without CoMP, and SRM beamforming with CoMP as a special case of the state-of-the-art. In order to maintain comparisons in line with the latter references, it is assumed that $B = 4$ BSs are respectively placed at each corner of a square cell with inter-cell spacing 100 [m], with each BS subject to a transmit power cap of $P_{\max,b} \leq 30$ [dbm], where the blockage probability can be accurately predictable.

Downlink users are assumed to be randomly placed within a square region and the noise floor σ_u^2 at each downlink user is assumed to be modeled as

$$\sigma_u^2 = 10\log_{10}(1000\kappa T) + \mathrm{NF} + 10\log_{10}(W)\,[\mathrm{dBm}], \tag{4.17}$$

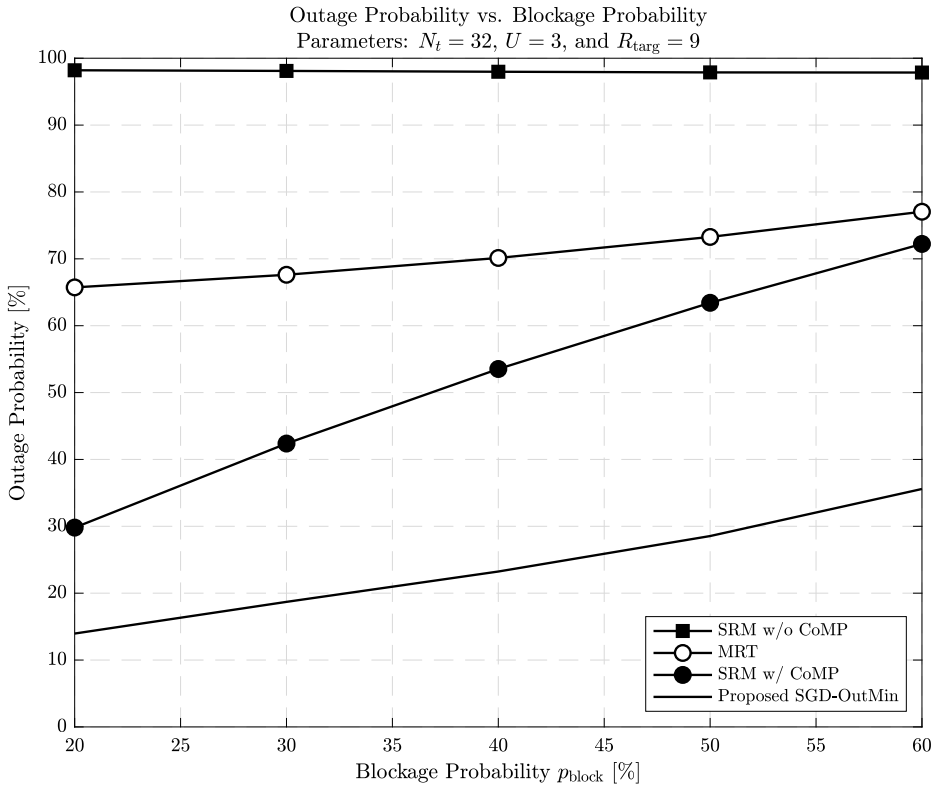
where κ is the Boltzmann's constant, $T = 293.15$ denotes the physical temperature at each user in kelvins, the noise figure at the downlink users NF is assumed to be 5 [dB] and the subcarrier bandwidth W is 20 [MHz].

For the comparisons, we consider a mmWave system operating at a carrier frequency of 28 [GHz], with the associated channel parameters according to [161, Table I]. For the sake of simplicity, it is assumed that blockage events occur with equal probability ($p_{b,u}^m = p_{\mathrm{block}}\ \forall u, b, m$), and that the target SINR is uniformly set to $\gamma_u = \gamma$. Finally, we define $R_{\mathrm{targ}} \triangleq \log_2(1+\gamma)$ for the sake of brevity.

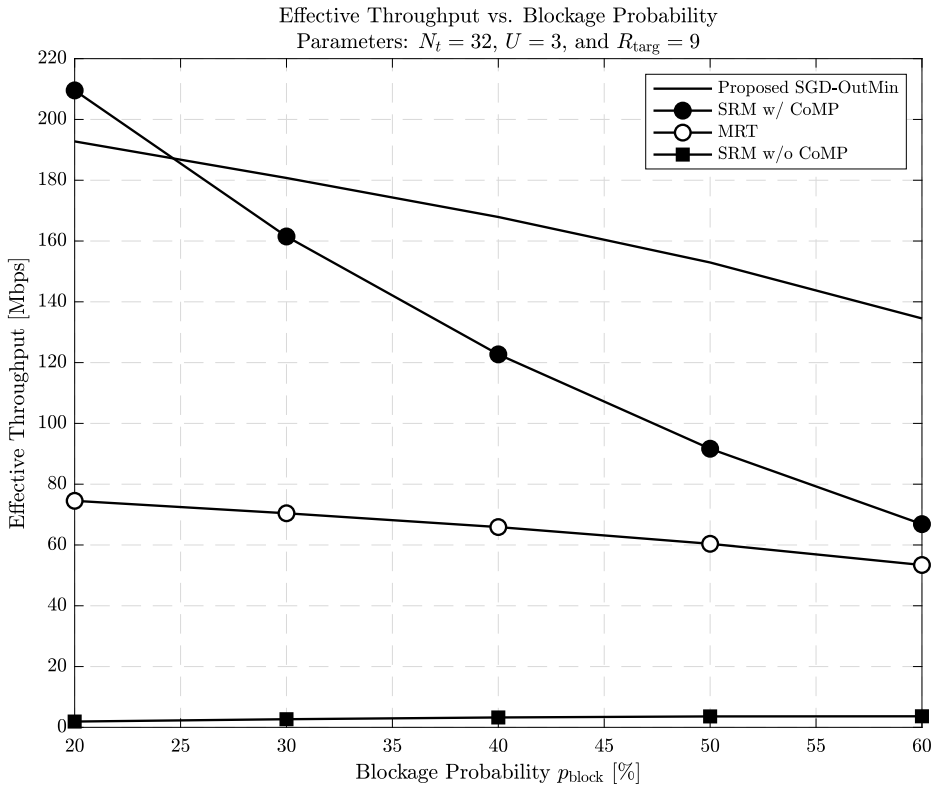
Figure 4.2 compares the outage probability and effective throughput performance of the proposed and SotA schemes as a function of blockage probability p_{block}, with $B = 4$ BSs each equipped with $N_t = 32$ antennas transmitting simultaneously to $U = 3$ single-antenna downlink users while targeting $R_{\mathrm{targ}} = 9$, respectively[4].

It is found non-surprisingly that the SRM scheme without CoMP transmission yields the worst performance, being easily beaten by the relatively simple MRT method. In comparison, the SRM scheme implemented in a CoMP setting improves on the latter by demonstrating the ability to handle inter-user interference. Besides that, it is also found that the proposed beamforming algorithm outperforms all the aforementioned SotA methods in terms of both outage probability and effective throughput over a wide range of blockage probability. Furthermore, we remark that the proposed algorithm can maintain a high data rate even under severe blockage conditions (*i.e.*, $p_{\mathrm{block}} \in [50, 60]\%$).

[4]Rather than arbitrary, this setup is motivated by the conditions of a real-life implementation of the proposed method, currently under pursuit in a collaborative research project in Japan.



(a) Outage probability as a function of blockage probability p_{block}.



(b) Effective throughput as a function of blockage probability p_{block}.

Figure 4.2: Outage probability and effective throughput comparisons with transmit antennas $N_t = 32$, number of users $U = 3$, the target throughput $R_{\text{targ}} = 9$, and the subcarrier bandwidth $W = 20$ [MHz].

4.3 Partially-Connected Hybrid Beamforming

Although the OutMin method proposed above shows its robustness against random path blockages in mmWave systems, it fails to take into consideration possible practical limitations such as a hybrid array structure of the CoMP and/or unavoidable CSI imperfection. To fill in this gap, we propose in this section a hybrid beamforming counterpart suitable to mitigate harmful effects caused by both random path blockages and CSI imperfection while taking into account the virtually-configured partially-connected array structure of the CoMP systems.

4.3.1 System Model: Further Generalization

Despite the similarity of the system model with that of the previous section, we further generalize the latter as follows. First, we consider uniform planar array (UPA) instead of ULA such that the channel model in equatin (4.1) is now rewritten as

$$\hat{\boldsymbol{h}}_{b,u} = \sqrt{\frac{1}{K_{b,u}}} \sum_{k=1}^{K_{b,u}} g_{b,u}^{k} \boldsymbol{a}_{N_t}(\theta_{b,u}^{k}, \phi_{b,u}^{k}), \tag{4.18}$$

where $\theta_{b,u}^{k}$ and $\phi_{b,u}^{k}$ are the elevation and azimuth AoD of the m-th cluster from the b-th AP towards the u-th downlink user, respectively, and $\boldsymbol{a}_T(\theta_{b,u}^{k}, \phi_{b,u}^{k})$ represents the array response vector[5]

Assuming that the UPAs equipping the APs have a regular square shape with $\sqrt{N_t}$ antenna elements on each axis, the array response vector $\boldsymbol{a}_T(\theta_{b,u}^{k}, \phi_{b,u}^{k})$ can be written as

$$\boldsymbol{a}_{N_t}(\theta_{b,u}^{k}, \phi_{b,u}^{k}) = \mathbf{c}_{\sqrt{N_t}}\big(\tfrac{1}{2}\sin(\theta_{b,u}^{k})\cos(\phi_{b,u}^{k})\big) \otimes \mathbf{c}_{\sqrt{N_t}}\big(\tfrac{1}{2}\cos(\theta_{b,u}^{k})\big), \tag{4.19}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{c}_N$ represents the array response vector of a ULA with M antenna elements, which can be expressed as

$$\mathbf{c}_N(x) \triangleq \frac{1}{\sqrt{N}}\left[1, e^{j2\pi x}, \cdots, e^{j2\pi(N-1)x}\right]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}. \tag{4.20}$$

Similar to the previous section, despite the knowledge of $\hat{\boldsymbol{h}}_{b,u}$, during the subsequent actual downlink, the system might be subjected to partial blockage if and when any or some of the LoS and NLoS clusters become temporarily blocked, such that the actual

[5]Although we consider a frequency-independent outdoor urban mmWave channel model, one can consider other mmWave channel models such as indoor scenarios [186], rural macrocell scenarios [187], and frequency-dependent wideband scenarios [188]. In any case, since the proposed algorithm presented later does not assume any particular channel model, the proposed algorithm can be easily extended to such different scenarios.

channel between the b-th AP and the u-th user can be modeled as

$$\boldsymbol{h}_{b,u} = \sqrt{\frac{1}{K_{b,u}}} \sum_{k=1}^{K_{b,u}} \omega_{b,u}^{k}(g_{b,u}^{k}) \boldsymbol{a}_{N_t}(\theta_{b,u}^{k}, \phi_{b,u}^{k}), \tag{4.21}$$

with now $\omega_{b,u}^{k} \sim f(p_{b,u}^{k}; g_{b,u}^{k}, \zeta_{b,u,k}^{2})$, where $f(p_{b,u}^{k}; g_{b,u}^{k}, \zeta_{b,u,k}^{2})$ denotes the Bernoulli-Gaussian distribution given by

$$f(p_{b,u}^{k}; g_{b,u}^{k}, \zeta_{b,u,k}^{2}) = p_{b,u}^{k}\delta(\omega_{b,u}^{k}) + (1 - p_{b,u}^{k})\mathcal{CN}(g_{b,u}^{k}; \zeta_{b,u,k}^{2}), \tag{4.22}$$

with $p_{b,u}^{k}$, $\delta(\cdot)$, and $\zeta_{b,u,k}^{2}$ denoting the corresponding blockage probability, Dirac delta function, and channel gain estimation uncertainty variance, respectively.

It should be noted that the random variable model $\omega_{b,u}^{k}$ adopted here is a generalization of that of the previous section. In particular, while the blockage model used earlier is simple Bernoulli, the model in equation (4.22) is Bernoulli-Gaussian [152], with the first term modeling the mean random blockage that occurs with probability $p_{b,u}^{k}$, while the second term captures variations of the small-scale fading coefficients, which occur due to a summation of effects including CSI imperfection, channel aging, and partial blockage. Setting $\zeta_{b,u,k}^{2} = 0$, the Bernoulli-Gaussian model of equation (4.22) reduces to the usual Bernoulli blockage model considered in the previus section, and the path of the m-th cluster from the b-th AP to the u-th user gets completely blocked when $\omega_{b,u}^{k} = 0$.

Under the assumption that the number of RF chains at each AP is limited to $N_{\mathrm{RF}} \ll N_t$, and considering the channel model detailed above, let $\boldsymbol{f}_{b,u} \in \mathbb{C}^{N_{\mathrm{RF}} \times 1}$ denote the transmit digital baseband beamforming vector from the b-th AP towards the u-th user, and $\boldsymbol{V}_b \in \mathbb{C}^{N_t \times N_{\mathrm{RF}}}$ be the transmit analog beamforming matrix employed by the b-th AP when transmitting to all users, subject to the unit modulus constraint (*i.e.*, $|[\boldsymbol{V}_b]_{i,j}| = 1$) [167].

Then, introducing the aggregate digital baseband beamforming matrix

$$\boldsymbol{F}_b \triangleq [\boldsymbol{f}_{b,1}, \cdots, \boldsymbol{f}_{b,u}], \tag{4.23}$$

employed by the b-th AP, the received signal y_u at the u-th user can be written as

$$\begin{aligned} y_u &= \sum_{b\in\mathcal{B}} \boldsymbol{h}_{b,u}^{\mathrm{H}} \boldsymbol{V}_b \boldsymbol{F}_b \boldsymbol{x} + n_u = \sum_{b\in\mathcal{B}} \boldsymbol{h}_{b,u}^{\mathrm{H}} \boldsymbol{V}_b \boldsymbol{f}_{b,u} x_u + \sum_{u'\in\mathcal{U}\setminus u} \sum_{b\in\mathcal{B}} \boldsymbol{h}_{b,u}^{\mathrm{H}} \boldsymbol{f}_{b,u'} x_{u'} + n_u \\ &= \underbrace{\boldsymbol{h}_u^{\mathrm{H}} \boldsymbol{V} \boldsymbol{f}_u x_u}_{\text{intended signal}} + \underbrace{\sum_{u'\in\mathcal{U}\setminus u} \boldsymbol{h}_u^{\mathrm{H}} \boldsymbol{V} \boldsymbol{f}_{u'} x_{u'}}_{\text{interuser interference}} + n_u, \end{aligned} \tag{4.24}$$

where $\boldsymbol{x} \triangleq [x_1, \cdots, x_U]^{\mathrm{T}}$ is the symbol vector transmitted cooperatively from all APs

to all users, with x_u denoting a symbol targeting at the u-th user; n_u denotes i.i.d. circularly symmetric zero-mean AWGN at the u-th user[6], *i.e.* $n_u \sim \mathcal{CN}(0, \xi_u^2)$; and finally the cooperative analog beamforming matrix, $\boldsymbol{V} \triangleq \text{blkdiag}(\boldsymbol{V}_1, \cdots, \boldsymbol{V}_B) \in \mathbb{C}^{BN_t \times BN_{\text{RF}}}$, the aggregate channel vector, $\boldsymbol{h}_u \triangleq \left[\boldsymbol{h}_{1,u}^{\text{T}}, \cdots, \boldsymbol{h}_{B,u}^{\text{T}}\right]^{\text{T}} \in \mathbb{C}^{BN_t \times 1}$, and digital baseband beamformer, $\boldsymbol{f}_u \triangleq \left[\boldsymbol{f}_{1,u}^{\text{T}}, \cdots, \boldsymbol{f}_{B,u}^{\text{T}}\right]^{\text{T}} \in \mathbb{C}^{BN_{\text{RF}} \times 1}$ from all APs to the u-th user are implicitly defined and introduced in the last equation, for notational simplicity.

We emphasize that the block-diagonal structure of the analog beamforming matrix $\boldsymbol{V}$ implies that the cooperative downlink transmission scheme yielding the received signal described by equation (4.24) is a virtually-configured partially-connected hybrid beamforming architecture, unlike the fully-connected counterpart considered in related works.

4.3.2 Problem Formulation

The task of performing mmWave robust CoMP downlink hybrid beamforming subject to random path blockage can be formulated as a constrained stochastic optimization problem, just as the reformulation given in the previous section shows. To elaborate further, our strategy is to build resilience against random blockages by minimizing the stochastic sum-outage-probability subjected to this phenomenon, which can be formulated as

$$\underset{\boldsymbol{f},\boldsymbol{V}}{\text{minimize}} \quad \sum_{u\in\mathcal{U}} \Pr\left\{\Gamma_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u\right) < \gamma_u\right\} \tag{4.25a}$$

$$\text{subject to} \quad \|\boldsymbol{V}_b\boldsymbol{F}_b\|_{\text{F}}^2 \leq P_{\max,b},\ \forall\, b, \tag{4.25b}$$

$$\boldsymbol{V}_b \in \mathcal{M}_{cc}^{N_{\text{RF}}\times N_t},\ \forall\, b, \tag{4.25c}$$

where γ_u denotes the target SINR for the u-th user, and the effective SINR Γ_u can be given by

$$\Gamma_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u\right) = \frac{|\boldsymbol{h}_u^{\text{H}}\boldsymbol{V}\boldsymbol{f}_u|^2}{\sum_{u'\in\mathcal{U}\setminus u}|\boldsymbol{h}_u^{\text{H}}\boldsymbol{V}\boldsymbol{f}_{u'}|^2 + \xi_u^2}, \tag{4.26}$$

with $\boldsymbol{f} \triangleq \left[\boldsymbol{f}_1^{\text{T}}, \cdots, \boldsymbol{f}_U^{\text{T}}\right]^{\text{T}} \in \mathbb{C}^{BUN_{\text{RF}}\times 1}$.

In equation (4.25), $\mathcal{M}_{cc}^{N_{\text{RF}}\times N_t}$ denotes an N_{RF}-by-N_t sub-Riemannian circle manifold in $\mathbb{C}^{N_{\text{RF}}\times N_t}$, defined as

$$\mathcal{M}_{cc}^{N_{\text{RF}}\times N_t} \triangleq \left\{\boldsymbol{z}\in\mathbb{C}^{N_{\text{RF}}\times N_t}\,\middle|\,|z_i|=1, i=\{1,\cdots,N_{\text{RF}}\times N_t\}\right\}. \tag{4.27}$$

It should be noted that the problem formulated in equation (4.25) differs fundamentally from related robust beamforming methods. The objective of the problem

[6]It should be noted that this method will be proposed later for different types of noise. For the sake of simplicity, the common AWGN model is assumed here.

formulated in equation (4.25) is to enable reliable communication even under harmful random blockages and CSI errors, by minimizing sum-outage, which is more practical from a system point-of-view. In addition, unlike deterministic formulations, which are often based on a max-min sum rate maximization framework followed by convex approximations with complexity that grows rapidly, equation (4.25) is stochastic and has only as many constraints as the number of APs owing to the power constraints on top of the unit modular constraint for analog beamforming.

4.3.3 Hybrid OutMin

Reformulation as Empirical Risk Minimization Problem

In this section, we determine the update rules for the digital and analog beamformers, $\boldsymbol{f}$ and $\boldsymbol{V}$, required to solve equation (4.25). To that end, we employ the Gauss-Seidel-type block-coordinate stochastic gradient technique [189] in conjunction with manifold optimization to decouple the variables $\boldsymbol{f}$ and $\boldsymbol{V}$, so that the intractability imposed jointly by the stochastic objective and the non-convex constraint can be mitigated, providing an iterative alternate solution for the problem formulated in equation (4.25).

To this end, owing to the ambiguity of the objective function, we first introduce the following indicator function $\mathbb{1}_{\Gamma_u(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u)<\gamma_u}$ given by

$$\mathbb{1}_{\Gamma_u(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u)<\gamma_u} = \begin{cases} 1 & \text{if } \Gamma_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u\right) < \gamma_u \\ 0 & \text{otherwise} \end{cases}, \tag{4.28}$$

such that equation (4.25) can be written as a type of expectation minimization problem, namely

$$\begin{aligned} \underset{\boldsymbol{f},\boldsymbol{V}}{\text{minimize}} \quad & \sum_{u\in\mathcal{U}} \mathbb{E}_{\omega_{b,u}^k}\left[\mathbb{1}_{\Gamma_u(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u)<\gamma_u}\right] && \text{(4.29a)} \\ \text{subject to} \quad & \|\boldsymbol{V}_b\boldsymbol{F}_b\|_{\mathrm{F}}^2 \leq P_{\max,b},\ \forall\, b, && \text{(4.29b)} \\ & \boldsymbol{V}_b \in \mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t},\ \forall\, b. && \text{(4.29c)} \end{aligned}$$

Under the assumption that the channel gains and AoDs are obtained from the preceding channel estimation process, while also considering that some of those paths might be under random blockage during the actual data transmission, one may notice that possible combinations of blockage patterns due to $\omega_{b,u}^k$ can be randomly generated and utilized as a training dataset for problem (4.29). This is based on the established fact that the path blockage probabilities of mmWave channels as well as the CSI error variance can be (at least roughly) estimated, as demonstrated in [176–178].

Considering the above, we define $\boldsymbol{h}_u^m$ as the m-th training data batch for the channel

$\boldsymbol{h}_u$ of the u-th user, such that the ERM problem formulation of equation (4.29) can be further rewritten in terms of the summation, *i.e.*

$$\underset{\boldsymbol{f},\boldsymbol{V}}{\text{minimize}} \quad \frac{1}{M}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}} \mathbb{1}_{\Gamma_u(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u^m)<\gamma_u} \tag{4.30a}$$

$$\text{subject to} \quad \|\boldsymbol{V}_b\boldsymbol{F}_b\|_{\mathrm{F}}^2 \leq P_{\max,b}\;\forall\, b, \tag{4.30b}$$

$$\boldsymbol{V}_b \in \mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}\;\forall\, b, \tag{4.30c}$$

where M denotes the size of the training dataset.

To address the intractable non-smoothness of the ERM problem given in equation (4.30), we further introduce a smooth surrogate function $\nu_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u^m\right)$ so that a gradient expression of (4.30a) can be efficiently computed. To this end, we exploit the fact that equation (4.30) can be seen as a classification problem; therefore, the indicator function (4.30a) can be replaced by the hinge surrogate function [185, 190]

$$\nu_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u^m\right) = \begin{cases} 0 & \text{if } 1-\frac{\Gamma_u(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u^m)}{\gamma_u} < 0, \\ 1-\frac{\Gamma_u(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u^m)}{\gamma_u} & \text{otherwise,} \end{cases} \tag{4.31}$$

which in turn enables equation (4.30) to be rewritten as

$$\underset{\boldsymbol{f},\boldsymbol{V}}{\text{minimize}} \quad \frac{1}{M}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}} \nu_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u^m\right) \tag{4.32a}$$

$$\text{subject to} \quad \|\boldsymbol{V}_b\boldsymbol{F}_b\|_{\mathrm{F}}^2 \leq P_{\max,b}\;\forall\, b, \tag{4.32b}$$

$$\boldsymbol{V}_b \in \mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}\;\forall\, b. \tag{4.32c}$$

The latter formulation of the proposed robust cooperative hybrid mmWave beamforming problem can be recognized as a type of differentiable non-convex stochastic optimization problem with manifold constraints, whose solution can be found via a block-coordinate descent algorithm such as that proposed in [189]. To this end, however, the gradients of the objective (4.32a) with respect to the digital baseband matrix $\boldsymbol{f}$ and the analog RF beamforming matrix $\boldsymbol{V}$ must be derived, which is the subject of the following.

Following a standard stochastic coordinate-descent framework [191], this task will be pursued under a hybrid and alternate approach in which the gradients of the objective function (4.32a) with respect to the digital baseband component $\boldsymbol{f}$ with $\boldsymbol{V}$ constant, and with respect to the analog RF component $\boldsymbol{V}$ with $\boldsymbol{f}$ constant will be considered. In addition, we consider a minimization algorithm for variance reduction in terms of gradient direction at each iteration [183] by means of mini-batchs.

Digital Baseband Component

For notation simplicity, we hereafter define the mini-batch size employed at the i-th algorithmic iteration as M_i, with $1 \ll M_i \ll M$.

For a fixed $\boldsymbol{V}$, and over a mini batch of size M_i, the optimization problem (4.32) reduces to the subproblem

$$\underset{\boldsymbol{f}}{\text{minimize}} \qquad \frac{1}{M_i} \sum_{m=1}^{M_i} \sum_{u \in \mathcal{U}} \nu_u \left(\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m \right) \tag{4.33a}$$

$$\text{subject to} \qquad \| \boldsymbol{V}_b \boldsymbol{F}_b \|_{\mathrm{F}}^2 \leq P_{\max,b} \; \forall \, b. \tag{4.33b}$$

Since the entries of the digital baseband beamforming vector $\boldsymbol{f}$ can be any complex value that satisfies the power constraint (4.33b), the gradient of $\nu_u \left(\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m \right)$ with respect to $\boldsymbol{f}$ can be computed by taking its derivative for a fixed $\boldsymbol{V}$, which yields

$$\nabla \nu_u \left(\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m \right) = \begin{cases} 0 & \text{if } 1 - \frac{\Gamma_u (\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m)}{\gamma_u} < 0 \\ - \frac{\nabla \Gamma_u (\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m)}{\gamma_u} & \text{otherwise,} \end{cases} \tag{4.34}$$

where the gradient of the SINR expression Γ_u can be computed by introducing its alternative formula for the stacked baseband beamforming vector $\boldsymbol{f}$

$$\begin{aligned} \nabla \Gamma_u (\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m) &= \nabla \frac{\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f}}{\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2}, \\ &= \frac{\boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f}}{\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2} - \frac{\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f}}{(\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2)^2} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f}, \end{aligned} \tag{4.35}$$

where the second equality follows from the Wirtinger derivative and the auxiliary matrices $\boldsymbol{\Phi}_{\boldsymbol{f}}^m$ and $\boldsymbol{\Psi}_{\boldsymbol{f}}^m$ are respectively defined as

$$\boldsymbol{\Phi}_{\boldsymbol{f}}^m \triangleq \operatorname{diag} (\boldsymbol{e}_u) \otimes \boldsymbol{V}^{\mathrm{H}} \boldsymbol{h}_u^m \boldsymbol{h}_u^{m\mathrm{H}} \boldsymbol{V}, \tag{4.36a}$$

$$\boldsymbol{\Psi}_{\boldsymbol{f}}^m \triangleq \operatorname{diag} (\bar{\boldsymbol{e}}_u) \otimes \boldsymbol{V}^{\mathrm{H}} \boldsymbol{h}_u^m \boldsymbol{h}_u^{m\mathrm{H}} \boldsymbol{V}, \tag{4.36b}$$

where $\boldsymbol{e}_u \in \{0,1\}^{U \times 1}$ denotes the u-th vector of the standard orthonormal basis of dimension U and $\bar{\boldsymbol{e}}_u$ denotes the logical negation of $\boldsymbol{e}_u$ (*i.e.*, $\bar{\boldsymbol{e}}_u = \boldsymbol{1} - \boldsymbol{e}_u$).

From the above, the update of $\boldsymbol{f}$ for a fixed analog RF beamforming matrix $\boldsymbol{V}$ can be written as

$$\boldsymbol{f}^{(i)} = \mathcal{P}_{\| \boldsymbol{V}_b \boldsymbol{F}_b \|_{\mathrm{F}}^2 \leq P_{\max,b}} \Big(\boldsymbol{f}^{(i-1)} - \frac{\alpha_i^f}{M_i} \sum_{m=1}^{M_i} \sum_{u \in \mathcal{U}} \nabla \nu_u \big(\boldsymbol{f} | \boldsymbol{V}, \boldsymbol{h}_u^m \big) \Big), \tag{4.37}$$

where $\boldsymbol{f}^{(i-1)}$ indicates the solution obtained at the $(i-1)$-th iteration, α_i^f is the

corresponding step-size to be tuned later, and $\mathcal{P}_{\|\boldsymbol{V}_b\boldsymbol{F}_b\|_{\mathrm{F}}^2\leq P_{\max,b}}(\cdot)$ denotes the projection onto the feasible convex set defined by the power constraint per-AP given by inequality (4.33b).

Analog RF Component

Analogous to the above, for a fixed $\boldsymbol{f}$, and over a mini batch of size M_i, the optimization problem (4.32) with respect to $\boldsymbol{V}$ reduces to the subproblem

$$\underset{\boldsymbol{V}}{\text{minimize}} \qquad \frac{1}{M_i}\sum_{m=1}^{M_i}\sum_{u\in\mathcal{U}}\nu_u\left(\boldsymbol{V}|\boldsymbol{f},\boldsymbol{h}_u^m\right) \tag{4.38a}$$

$$\text{subject to} \qquad \boldsymbol{V}_b\in\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}\ \forall\, b. \tag{4.38b}$$

Unlike digital baseband beamforming $\boldsymbol{f}$, however, calculating the gradient direction with respect to $\boldsymbol{V}$ is challenging. This is because of the unit modular constraint defined by the complex circle manifold $\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}$; thus, manifold optimization techniques [192] must be employed. To that end, we first define the tangent space at a given $\boldsymbol{z}\in\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}$, which can be written as

$$\mathcal{T}_{\boldsymbol{z}}\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}\triangleq\{\boldsymbol{\tau}\in\mathbb{C}^{N_{\mathrm{RF}}N_t}|\Re\{\boldsymbol{\tau}\circ\boldsymbol{z}^*\}=\boldsymbol{0}\}, \tag{4.39}$$

which contains all tangent vectors to $\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}$ at a certain point $\boldsymbol{z}$.

Since the complex circle manifold $\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}$ is a smooth Riemannian manifold, a positive-definite inner product can be defined on the tangent space given by equation (4.39), such that the Riemannian gradient $\nabla^{\mathcal{M}}\nu_u\left(\boldsymbol{V}|\boldsymbol{f}_u,\boldsymbol{h}_u^m\right)$ at $\boldsymbol{z}$ on $\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}$ can be expressed as the orthogonal projection of the Euclidean gradient $\nabla\nu_u\left(\boldsymbol{V}|\boldsymbol{f}_u,\boldsymbol{h}_u^m\right)$ onto its tangent space. In other words, we may define

$$\begin{aligned}\nabla^{\mathcal{M}}\nu_u\left(\boldsymbol{V}|\boldsymbol{f},\boldsymbol{h}_u^m\right)&\triangleq\mathcal{P}_{\mathcal{T}_{\boldsymbol{z}}\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}}\left(\frac{\sum_{m=1}^{M_i}\nabla\nu_u(\boldsymbol{V}|\boldsymbol{f},\boldsymbol{h}_u^m)}{M_i}\right)\\&=\frac{1}{M_i}\sum_{m=1}^{M_i}\nabla\nu_u\left(\boldsymbol{V}|\boldsymbol{f},\boldsymbol{h}_u^m\right)-\Re\left\{\frac{1}{M_i}\sum_{m=1}^{M_i}\nabla\nu_u\left(\boldsymbol{V}|\boldsymbol{f},\boldsymbol{h}_u^m\right)\circ\boldsymbol{V}^*\right\}\circ\boldsymbol{V}.\end{aligned} \tag{4.40}$$

To facilitate the derivation of the Euclidean gradient $\nabla\nu_u\left(\boldsymbol{V}|\boldsymbol{f}_u,\boldsymbol{h}_u^m\right)$ with respect to $\boldsymbol{V}$, we reformulate the received signal and the SINR expressions so as to form a tractable optimization variable. In particular, we rewrite the received signal originally given in equation (4.24) as

$$y_u=\boldsymbol{h}_u^{\mathrm{H}}\boldsymbol{V}\boldsymbol{f}_u x_u+\sum_{u'\in\mathcal{U}\setminus u}\boldsymbol{h}_u^{\mathrm{H}}\boldsymbol{V}\boldsymbol{f}_{u'}x_{u'}+n_u \tag{4.41a}$$

$$=((x_u\boldsymbol{f}_u)^{\mathrm{T}}\otimes\boldsymbol{h}_u^{\mathrm{H}})\mathrm{vec}(\boldsymbol{V})+\sum_{u'\in\mathcal{U}\setminus u}((x_{u'}\boldsymbol{f}_{u'})^{\mathrm{T}}\otimes\boldsymbol{h}_u^{\mathrm{H}})\overbrace{\mathrm{vec}(\boldsymbol{V})}^{\triangleq\boldsymbol{v}}+n_u, \tag{4.41b}$$

where vec$(\cdot)$ denotes the vectorized (column stacked) representation of a matrix, such that the implicitly defined vectorized representation $\boldsymbol{v} \triangleq \text{vec}(\boldsymbol{V})$ is a sparse vector owing to the partially connected structure of $\boldsymbol{V}$.

It will prove convenient, therefore, to further define the auxiliary matrix

$$\tilde{\boldsymbol{V}} \triangleq \big[\boldsymbol{V}_1, \cdots, \boldsymbol{V}_B\big], \tag{4.41c}$$

such that its vectorized representation $\tilde{\boldsymbol{v}} \triangleq \text{vec}(\tilde{\boldsymbol{V}})$ is a dense vector, and we may write

$$y_u = ((x_u \boldsymbol{f}_u)^{\mathrm{T}} \otimes \boldsymbol{h}_u^{\mathrm{H}}) \boldsymbol{W} \tilde{\boldsymbol{v}} + \sum_{u' \in \mathcal{U} \setminus u} ((x_{u'} \boldsymbol{f}_{u'})^{\mathrm{T}} \otimes \boldsymbol{h}_u^{\mathrm{H}}) \boldsymbol{W} \tilde{\boldsymbol{v}} + n_u, \tag{4.41d}$$

where $\boldsymbol{W}$ denotes the transform matrix mapping the dense vector $\tilde{\boldsymbol{v}}$ onto its sparse representation $\boldsymbol{v}$.

In light of equation (4.41), the SINR originally expressed as in equation (4.26) can be rewritten for a given channel $\boldsymbol{h}_u^m$ as

$$\Gamma_u\left(\tilde{\boldsymbol{v}} | \boldsymbol{f}, \boldsymbol{h}_u^m\right) = \frac{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2}, \tag{4.42}$$

where

$$\boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \triangleq \boldsymbol{W}^{\mathrm{H}} \big(\boldsymbol{f}_u^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\big)^{\mathrm{H}} \big(\boldsymbol{f}_u^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\big) \boldsymbol{W}, \tag{4.43a}$$

$$\boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \triangleq \boldsymbol{W}^{\mathrm{H}} \sum_{u' \in \mathcal{U} \setminus u} \big(\boldsymbol{f}_{u'}^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\big)^{\mathrm{H}} \big(\boldsymbol{f}_{u'}^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\big) \boldsymbol{W}. \tag{4.43b}$$

Given the above, the Euclidean gradient $\nabla \nu_u\left(\tilde{\boldsymbol{v}} | \boldsymbol{f}, \boldsymbol{h}_u^m\right)$ can be represented as

$$\nabla \nu_u\left(\tilde{\boldsymbol{v}} | \boldsymbol{f}, \boldsymbol{h}_u^m\right) = \begin{cases} 0 & \text{if } 1 - \frac{\Gamma_u(\tilde{\boldsymbol{v}} | \boldsymbol{f}, \boldsymbol{h}_u^m)}{\gamma_u} < 0 \\ -\frac{\nabla \Gamma_u(\tilde{\boldsymbol{v}} | \boldsymbol{f}, \boldsymbol{h}_u^m)}{\gamma_u} & \text{otherwise,} \end{cases} \tag{4.44}$$

where, based on the SINR reformulation given in equation (4.42), the Euclidean gradient of the SINR can be expressed similarly to equation (4.35), *i.e.*

$$\nabla \Gamma_u\left(\tilde{\boldsymbol{v}} | \boldsymbol{f}, \boldsymbol{h}_u^m\right) = \frac{\boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2} - \frac{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{\big(\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2\big)^2} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}. \tag{4.45}$$

In light of the above, the Riemannian gradient can be finally rewritten as

$$\nabla^{\mathcal{M}}\nu_u\left(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m\right)=\frac{1}{M_i}\sum_{m=1}^{M_i}\nabla\nu_u\left(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m\right) - \Re\Bigg\{\frac{1}{M_i}\sum_{m=1}^{M_i}\nabla\nu_u\left(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m\right)\circ\tilde{\boldsymbol{v}}^*\Bigg\}\circ\tilde{\boldsymbol{v}}, \tag{4.46}$$

which can be recognized as a variation of equation (4.40) over the vectorized dense representation $\tilde{\boldsymbol{v}}$ of $\boldsymbol{V}$.

Based on the Riemannian gradient of equation (4.46), the updated vector $\tilde{\boldsymbol{v}}$ obtained over the tangent space and mapped onto the complex circle manifold $\mathcal{M}_{cc}^{N_{\mathrm{RF}}N_t}$ can be written as

$$\tilde{\boldsymbol{v}}^{(i)}=\mathrm{Retr}\Big(\tilde{\boldsymbol{v}}^{(i-1)}-\alpha_i^v\sum_{u\in\mathcal{U}}\nabla^{\mathcal{M}}\nu_u\big(\tilde{\boldsymbol{v}}^{(i-1)}|\boldsymbol{f},\boldsymbol{h}_u^m\big)\Big), \tag{4.47a}$$

where $\mathrm{Retr}(\cdot)$ denotes the retraction operator that rescales a given vector to element-wise unit modular entries [192] and α_i^v is a step size to be given later.

After obtaining $\tilde{\boldsymbol{v}}^{(i)}$, the update of $\boldsymbol{V}$ for a fixed digital baseband beamforming matrix $\boldsymbol{f}$ can finally be obtained by unvectorizing its associated sparse version, *i.e.*

$$\boldsymbol{V}^{(i)}=\mathrm{unvec}(\boldsymbol{W}\tilde{\boldsymbol{v}}^{(i)}). \tag{4.47b}$$

Choice of Stepsize Parameters

Having derived gradient updates corresponding to both the digital baseband beamforming $\boldsymbol{f}$ and the virtually configured partially connected analog RF beamforming $\boldsymbol{V}$ constrained by the unit modular manifold $\mathcal{M}_{cc}$, we proceed to propose a new blockage-robust hybrid beamforming algorithm via the block-coordinate stochastic gradient descent (BSGD) framework for CoMP systems operating at mmWave bands. To that end, we first need to determine the learning rates α_i^f and α_i^v employed in equations (4.37) and (4.47a), respectively.

There are two well-known approaches to setup the stepsize for SGD algorithms, namely, the shrinkage criterion [193], which entails a shrinking learning rate $\alpha_1 > \cdots > \alpha_i > \cdots$ satisfying $\sum_{i=1}^{\infty}\alpha_i\to\infty$ and $\sum_{i=1}^{\infty}\alpha_i^2<\infty$; and the Lipschitz-based criterion [189], according to which the learning rate is set to be $\alpha_i\leq 1/L$, $\forall i$, where L is the Lipschitz constant.

Since it is highly challenging to determine the exact Lipschitz constant for the problem at hand, setting $\alpha_i=\rho/(\sqrt{i}\cdot L^*)$, where L^* is a lower bound of the Lipschitz constant L and ρ is a scaling coefficient, has been considered in the literature [189,194]. With that in mind, one can leverage the well-known Taylor theorem to obtain a lower

bound of the Lipschitz constant. In particular, upon momentarily altering our notation and using ν_u to denote the u-th term of either of the surrogate objective functions in equations (4.33a) and (4.38a), the Taylor Theorem yields $L_u \cdot \mathbf{I} \succeq \nabla^2 \nu_u \implies L_u \geq \lambda_{\max}(\nabla^2 \nu_u)$, where $\lambda_{\max}(\cdot)$ is the largest eigenvalue (spectral norm) of a given matrix, such that the lower bound on the Lipschitz constant is given by

$$L^* = \sum_{u=1}^{U} \lambda_{\max}(\nabla^2 \nu_u), \tag{4.48}$$

which is also leveraged in [194].

In light of the above, we offer below the derivation of the Lipschitz constant lower bounds L_f^* and L_v^* corresponding to the stochastic-gradient-based solutions of the subproblems described by equations (4.33) and (4.38), respectively.

The Hessian of $\nu_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right)$ is given by

$$\begin{aligned} H_f(\nu_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right)) &\triangleq \nabla^2 \nu_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right) \\ &= \begin{cases} \mathbf{0} & \text{if } 1 - \frac{\Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m)}{\gamma_u} < 0 \\ \frac{-H_f(\Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m))}{\gamma_u} & \text{otherwise,} \end{cases} \end{aligned} \tag{4.49}$$

where $H_f(\Gamma_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right))$ is the Hessian of the SINR with respect to $\boldsymbol{f}$, which in turn is given by

$$\begin{aligned} H_f\left(\Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m)\right) &\triangleq \nabla^2 \Gamma_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right) \begin{bmatrix} \frac{\partial^2 \Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m)}{\partial \boldsymbol{f}^* \partial \boldsymbol{f}^{\mathrm{T}}} & \frac{\partial^2 \Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m)}{\partial \boldsymbol{f}^* \partial \boldsymbol{f}^{\mathrm{H}}} \\ \frac{\partial^2 \Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m)}{\partial \boldsymbol{f} \partial \boldsymbol{f}^{\mathrm{T}}} & \frac{\partial^2 \Gamma_u(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m)}{\partial \boldsymbol{f} \partial \boldsymbol{f}^{\mathrm{H}}} \end{bmatrix} \\ &= \overbrace{\frac{1}{\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2} \begin{bmatrix} \boldsymbol{\Phi}_{\boldsymbol{f}}^m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{\boldsymbol{f}}^{m\mathrm{T}} \end{bmatrix}}^{\triangleq \boldsymbol{q}_1} - \overbrace{\frac{\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f}}{(\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2)^2} \begin{bmatrix} \boldsymbol{\Psi}_{\boldsymbol{f}}^m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \end{bmatrix}}^{\triangleq \boldsymbol{q}_2} \\ &\quad - \frac{\overbrace{2 \begin{bmatrix} \boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f} \boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m & \boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f} \boldsymbol{f}^{\mathrm{T}} \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \\ \boldsymbol{\Phi}_{\boldsymbol{f}}^{m\mathrm{T}} \boldsymbol{f}^* \boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m & \boldsymbol{\Phi}_{\boldsymbol{f}}^{m\mathrm{T}} \boldsymbol{f}^* \boldsymbol{f}^{\mathrm{T}} \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \end{bmatrix}}^{\triangleq \boldsymbol{q}_3}}{(\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2)^2} \\ &\quad + \underbrace{\frac{2 \boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Phi}_{\boldsymbol{f}}^m \boldsymbol{f}}{(\boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} + \xi_u^2)^3} \begin{bmatrix} \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} \boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m & \boldsymbol{\Psi}_{\boldsymbol{f}}^m \boldsymbol{f} \boldsymbol{f}^{\mathrm{T}} \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \\ \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \boldsymbol{f}^* \boldsymbol{f}^{\mathrm{H}} \boldsymbol{\Psi}_{\boldsymbol{f}}^m & \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \boldsymbol{f}^* \boldsymbol{f}^{\mathrm{T}} \boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}} \end{bmatrix}}_{\triangleq \boldsymbol{q}_4}. \end{aligned} \tag{4.50}$$

Given the relation between $\nabla^2 \nu_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right)$ and $H_f(\Gamma_u\left(\boldsymbol{f}|\boldsymbol{V}, \boldsymbol{h}_u^m\right))$, as per equation

(4.49), it follows from equations (4.48) and (4.50) that, for a given user, we have

$$\begin{aligned} L_{f_u}^* &= \frac{1}{\gamma_u}\lambda_{\max}(-H_f(\Gamma_u(\boldsymbol{f}|\boldsymbol{V},\boldsymbol{h}_u^m))) \\ &\leq \frac{1}{\gamma_u}\big(\lambda_{\max}(-\boldsymbol{q}_1)+\lambda_{\max}(\boldsymbol{q}_2)+\lambda_{\max}(\boldsymbol{q}_3)+\lambda_{\max}(-\boldsymbol{q}_4)\big), \end{aligned} \tag{4.51}$$

where the latter inequality follows straightforwardly from the triangular inequality.

Note that $\boldsymbol{q}_1$ and $\boldsymbol{q}_4$ are both rank-1 positive semi-definite matrices, which implies that $\lambda_{\max}(-\boldsymbol{q}_1)=0$ and $\lambda_{\max}(-\boldsymbol{q}_4)=0$. In addition, for $\boldsymbol{q}_3$, we have

$$\begin{aligned} \lambda_{\max}(\boldsymbol{q}_3) &= \frac{2(\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Phi}_{\boldsymbol{f}}^m\boldsymbol{\Psi}_{\boldsymbol{f}}^m\boldsymbol{f}+\boldsymbol{f}^{\mathrm{T}}\boldsymbol{\Phi}_{\boldsymbol{f}}^{m\mathrm{T}}\boldsymbol{\Psi}_{\boldsymbol{f}}^{m\mathrm{T}}\boldsymbol{f}^*)}{(\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Psi}_{\boldsymbol{f}}^m\boldsymbol{f}+\xi_u^2)^2} \\ &= \frac{4\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Phi}_{\boldsymbol{f}}^m\boldsymbol{\Psi}_{\boldsymbol{f}}^m\boldsymbol{f}}{(\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Psi}_{\boldsymbol{f}}^m\boldsymbol{f}+\xi_u^2)^2}=0, \end{aligned} \tag{4.52}$$

where we employed the fact that $\boldsymbol{\Phi}_{\boldsymbol{f}}^m\boldsymbol{\Psi}_{\boldsymbol{f}}^m=\boldsymbol{0}$.

Using these results in equation (4.51), we obtain

$$L_{f_u}^* \leq \frac{1}{\gamma_u}\lambda_{\max}(\boldsymbol{q}_2), \tag{4.53}$$

with

$$\begin{aligned} \lambda_{\max}(\boldsymbol{q}_2) &= \frac{\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Phi}_{\boldsymbol{f}}^m\boldsymbol{f}}{(\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Psi}_{\boldsymbol{f}}^m\boldsymbol{f}+\xi_u^2)^2}\lambda_{\max}(\boldsymbol{\Psi}_{\boldsymbol{f}}^m) \\ &\leq \frac{\boldsymbol{f}^{\mathrm{H}}\boldsymbol{\Phi}_{\boldsymbol{f}}^m\boldsymbol{f}}{\xi_u^4}\lambda_{\max}(\boldsymbol{\Psi}_{\boldsymbol{f}}^m) \\ &= \frac{\mathrm{Tr}\left(\boldsymbol{V}\boldsymbol{f}_u\boldsymbol{f}_u^{\mathrm{H}}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{h}_u^m\boldsymbol{h}_u^{m\mathrm{H}}\right)\|\boldsymbol{h}_u^{m\mathrm{H}}\boldsymbol{V}\|_2^2}{\xi_u^4} \\ &\leq BN_tN_{\mathrm{RF}}\frac{\|\hat{\boldsymbol{h}}_u\|_2^4}{\xi_u^4}\sum_{b=1}^{B}P_{\max,b}, \end{aligned} \tag{4.54}$$

where the identity $\lambda_{\max}\left(\left[\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}^{\mathrm{T}}\end{smallmatrix}\right]\right)=\lambda_{\max}(\boldsymbol{A})$ and the bound $\mathrm{Tr}\,(\boldsymbol{AB})\leq\lambda_{\max}(\boldsymbol{A})\mathrm{Tr}\,(\boldsymbol{B})$ combined with the trivial result $\mathrm{Tr}\left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{H}}\right)=BN_tN_{\mathrm{RF}}$ were used in the first equation and the last inequality, respectively.

Combining equations (4.48), (4.53), and (4.54), we finally have

$$L_f^* \leq BN_tN_{\mathrm{RF}}\sum_{b=1}^{B}P_{\max,b}\cdot\sum_{u=1}^{U}\frac{\|\hat{\boldsymbol{h}}_u\|_2^4}{\gamma_u\xi_u^4}, \tag{4.55}$$

from which it follows that the learning rates α_i^f to be employed at each iteration of equation (4.37) in order to solve problem (4.33) and obtain the optimal digital

baseband beamformer $\boldsymbol{f}$ is given by

$$\alpha_i^f = \rho \left(\sqrt{i} B N_t N_{\mathrm{RF}} \sum_{b=1}^{B} P_{\max,b} \cdot \sum_{u=1}^{U} \frac{\|\hat{\boldsymbol{h}}_u\|_2^4}{\gamma_u \xi_u^4} \right)^{-1}. \tag{4.56}$$

Following steps similar to the above, the Hessian matrix with respect to $\tilde{\boldsymbol{v}}$ is given by

$$\begin{aligned} H_v\left(\nu_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m)\right) \triangleq \nabla^2 \Gamma_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m) &= \begin{bmatrix} \frac{\partial^2 \Gamma_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m)}{\partial \tilde{\boldsymbol{v}}^* \partial \tilde{\boldsymbol{v}}^{\mathrm{T}}} & \frac{\partial^2 \Gamma_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m)}{\partial \tilde{\boldsymbol{v}}^* \partial \tilde{\boldsymbol{v}}^{\mathrm{H}}} \\ \frac{\partial^2 \Gamma_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m)}{\partial \tilde{\boldsymbol{v}} \partial \tilde{\boldsymbol{v}}^{\mathrm{T}}} & \frac{\partial^2 \Gamma_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m)}{\partial \tilde{\boldsymbol{v}} \partial \tilde{\boldsymbol{v}}^{\mathrm{H}}} \end{bmatrix} \\ &= \overbrace{\frac{1}{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2} \begin{bmatrix} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \end{bmatrix}}^{\triangleq \boldsymbol{\ell}_1} - \overbrace{\frac{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2)^2} \begin{bmatrix} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \end{bmatrix}}^{\triangleq \boldsymbol{\ell}_2} \\ &\quad - \overbrace{\frac{2 \begin{bmatrix} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} \tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m & \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} \tilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \\ \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \tilde{\boldsymbol{v}}^* \tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m & \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \tilde{\boldsymbol{v}}^* \tilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \end{bmatrix}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2)^2}}^{\triangleq \boldsymbol{\ell}_3} \\ &\quad + \underbrace{\frac{2\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2)^3} \begin{bmatrix} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} \tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m & \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} \tilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \\ \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \tilde{\boldsymbol{v}}^* \tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m & \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \tilde{\boldsymbol{v}}^* \tilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m\mathrm{T}} \end{bmatrix}}_{\triangleq \boldsymbol{\ell}_4}, \end{aligned} \tag{4.57}$$

with $\lambda_{\max}(-\boldsymbol{\ell}_1) = \lambda_{\max}(-\boldsymbol{\ell}_4) = 0$, such that

$$L_{v_u}^* = \frac{1}{\gamma_u} \lambda_{\max}(-H_v(\Gamma_u(\tilde{\boldsymbol{v}}|\boldsymbol{f},\boldsymbol{h}_u^m)) \leq \frac{\lambda_{\max}(\boldsymbol{\ell}_2) + \lambda_{\max}(\boldsymbol{\ell}_3)}{\gamma_u}, \tag{4.58}$$

where

$$\lambda_{\max}(\boldsymbol{\ell}_2) = \frac{\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2)^2} \lambda_{\max}(\boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m), \tag{4.59a}$$

$$\lambda_{\max}(\boldsymbol{\ell}_3) = \frac{4\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^m \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}} \boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^m \tilde{\boldsymbol{v}} + \xi_u^2)^2}. \tag{4.59b}$$

For the sake of future convenience, we introduce the following equalities:

$$\boldsymbol{W}^{\mathrm{H}} \left(\boldsymbol{f}_u^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\right)^{\mathrm{H}} \left(\boldsymbol{f}_u^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\right) \boldsymbol{W} = \boldsymbol{W}^{\mathrm{H}} \left(\boldsymbol{f}_u^* \boldsymbol{f}_u^{\mathrm{T}} \otimes \boldsymbol{h}_u^m \boldsymbol{h}_u^{m\mathrm{H}}\right) \boldsymbol{W} \tag{4.60a}$$

$$\boldsymbol{W}^{\mathrm{H}} \sum_{u' \in \mathcal{U} \setminus u} \left(\boldsymbol{f}_{u'}^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\right)^{\mathrm{H}} \left(\boldsymbol{f}_{u'}^{\mathrm{T}} \otimes \boldsymbol{h}_u^{m\mathrm{H}}\right) \boldsymbol{W} = \boldsymbol{W}^{\mathrm{H}} \sum_{u' \in \mathcal{U} \setminus u} \left(\boldsymbol{f}_{u'}^* \boldsymbol{f}_{u'}^{\mathrm{T}} \otimes \boldsymbol{h}_u^m \boldsymbol{h}_u^{m\mathrm{H}}\right) \boldsymbol{W}, \tag{4.60b}$$

where we utilized $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC} \otimes \boldsymbol{BD})$.

Then, we obtain

$$\begin{aligned}\lambda_{\max}(\boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m}) &= \mathrm{Tr}\left(\boldsymbol{W}^{\mathrm{H}}\left(\boldsymbol{f}_u^{*}\boldsymbol{f}_u^{\mathrm{T}}\otimes\boldsymbol{h}_u^{m}\boldsymbol{h}_u^{m\mathrm{H}}\right)\boldsymbol{W}\right),\\ &\leq \sum_{b=1}^{B} P_{\max,b}\|\hat{\boldsymbol{h}}_u\|_2^2,\end{aligned} \tag{4.61}$$

and

$$\frac{\tilde{\boldsymbol{v}}^{\mathrm{H}}\boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^{m}\tilde{\boldsymbol{v}}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}}\boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m}\tilde{\boldsymbol{v}}+\xi_u^2)^2} \leq \frac{BN_tN_{\mathrm{RF}}\sum_{b=1}^{B}P_{\max,b}}{\xi_u^4}\|\hat{\boldsymbol{h}}_u\|_2^2, \tag{4.62a}$$

$$\frac{4\tilde{\boldsymbol{v}}^{\mathrm{H}}\boldsymbol{\Phi}_{\tilde{\boldsymbol{v}}}^{m}\boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m}\tilde{\boldsymbol{v}}}{(\tilde{\boldsymbol{v}}^{\mathrm{H}}\boldsymbol{\Psi}_{\tilde{\boldsymbol{v}}}^{m}\tilde{\boldsymbol{v}}+\xi_u^2)^2} \leq \frac{4BN_tN_{\mathrm{RF}}(\sum_{b=1}^{B}P_{\max,b})^2}{\xi_u^4}\|\hat{\boldsymbol{h}}_u\|_2^4, \tag{4.62b}$$

which, when combined with equations (4.58) and (4.59) and ultimately substituted into equation (4.48), yield

$$L_v^{*} \leq 5BN_tN_{\mathrm{RF}}\left(\sum_{b=1}^{B}P_{\max,b}\right)^2\cdot\sum_{u}\frac{\|\hat{\boldsymbol{h}}_u\|_2^4}{\gamma_u\xi_u^4}. \tag{4.63}$$

Consequently, the learning rates α_i^v to be employed at each iteration of equation (4.47a) in order to solve problem (4.38) and obtain the optimal digital baseband beamformer $\boldsymbol{V}$ according to equation (4.47b) are given by

$$\alpha_i^v = \rho\left(\sqrt{i}5BN_tN_{\mathrm{RF}}\left(\sum_{b=1}^{B}P_{\max,b}\right)^2\cdot\sum_{u}\frac{\|\hat{\boldsymbol{h}}_u\|_2^4}{\gamma_u\xi_u^4}\right)^{-1}. \tag{4.64}$$

Algorithm Description

Combining the results above, a complete scheme to solve the stochastic formulation of the mmWave cooperative hybrid blockage-robust beamforming design described originally in equation (4.4) is achieved, which is summarized in Algorithm 7.

Note that in order to reduce the variance of the gradient direction, we adopt an mini-batch stochastic gradient descent (MSGD)-type variance reduction technique, which updates the solution by taking the sample mean over a small fraction of the training dataset at each algorithmic iteration. With regards to generating the training dataset, we remark that the CSI error level $\zeta_{b,u,k}^2$ is typically several orders of magnitude smaller than its corresponding mean quantity $g_{b,u}^k$, which can also be systematically incorporated into the robust beamforming design in Algorithm 7.

In summary, the proposed algorithm exhibits robustness not only against random blockages but also against CSI errors, which to the best of our knowledge is a novel contribution that has not been presented before. Finally, the efficacy of the method is elucidated in the next section.

Algorithm 7 Partially-Connected Hybrid OutMin

Inputs: Channel gain and angle estimates $g_{b,u}^k$, $\theta_{b,u}^k$ and $\phi_{b,u}^k$; corresponding uncertainty variance $\zeta_{b,u,k}^2$; received signals y_u; maximum transmit power $P_{\max,b}$ and target SINR $\gamma_u \ \forall\, u, b, k$

Outputs: Digital and analog beamformers $\boldsymbol{V}$ and $\boldsymbol{f}$

1: Initialize $\boldsymbol{V}^{(0)}$ by phase matching to $\hat{\boldsymbol{h}}_{b,u}$.
2: Initialize $\boldsymbol{f}^{(0)}$ by phase matching to $\hat{\boldsymbol{h}}_{b,u}$.
3: **repeat**
4: Generate training batchs.
5: Update $\boldsymbol{f}^{(i)}$ via equation (4.37) with learning rate α_i^f as in equation (4.56).
6: Update $\boldsymbol{V}^{(i)}$ via equations (4.47) with learning rate α_i^v as in equation (4.64).
7: **until** $t = t_{\max}$
8: Retain last $\boldsymbol{V} = \boldsymbol{V}^{(i)}$ as the final analog beamformer.
9: Obtain final digital beamformer $\boldsymbol{f}$ via equation (4.37) with the last beamformer $\boldsymbol{V}$ and learning rate α_i^f as in equation (4.56).

4.4 Performance Assessment

In this section, we evaluate the sum-outage-minimizing hybrid robust mmWave beamforming method proposed above, comparing its performance against three fully digital schemes, namely, the full digital OutMin technique, the SRM method [195], and the classic MRT.

Since fully digital approaches assume that transmit beams are optimized without restrictions (employing a $BN_t \times U$ fully digital beamforming matrix to serve U users), the performances of fully digital methods serve as bounds on those of their hybrid counterparts.

In our simulations, a setup similar to that employed in the measurement campaign carried out for the 28 [GHz] band reported in [196] is considered, unless mentioned otherwise. In particular, a mmWave square microcell model, with sides 100 meters wide and with $B = 4$ APs each located at a corner of the square, is assumed, while $U = 2$ UEs are randomly placed within the square being served by the system. We set the heights of the APs and UEs to 15 [m] and 1.6 [m], respectively, and it is assumed that each AP is equipped with $N_t = 16$ transmit antennas but only $N_{\mathrm{RF}} = 2$ digital RF chains, with the maximum transmit power constrained to $P_{\max,b} = 30$ [dBm].

The mmWave channel propagation model proposed in [161, Table I] is utilized to characterize the path loss and the number of clusters, and the AWGN variance ξ_u^2 of each UE is set such that

$$10\log_{10}\left(\xi_u^2\right) = 10\log_{10}\left(1000\kappa T\right) + 10\log_{10}\left(W\right) + \mathrm{NF}, \tag{4.65}$$

where $\kappa \approx 1.38 \times 10^{-23}$ is the Boltzmann constant, $T = 293.15$ [K] is the physical temperature at each user location in Kelvins (*i.e.*, 20 degrees Celsius), $W = 100$ [MHz] is the subcarrier bandwidth, and NF = 5 [dB] is the standard noise figure corresponding to each UE, ultimately yielding $\xi_u^2 \approx -89$ [dBm].

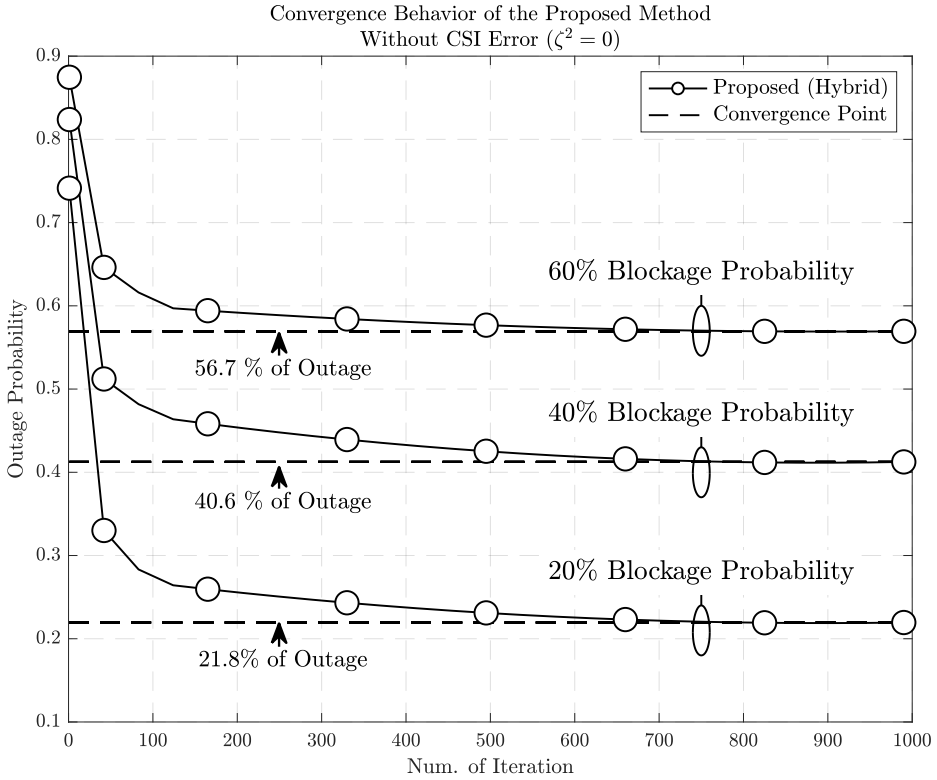
Finally, we set the target data rate to be either 3 or 5 [bits/s/Hz] depending on the user requirements, where γ_u is the desired SINR as defined in equation (4.4), while the mini-batch size $M_i = 16$.

4.4.1 Convergence Behavior

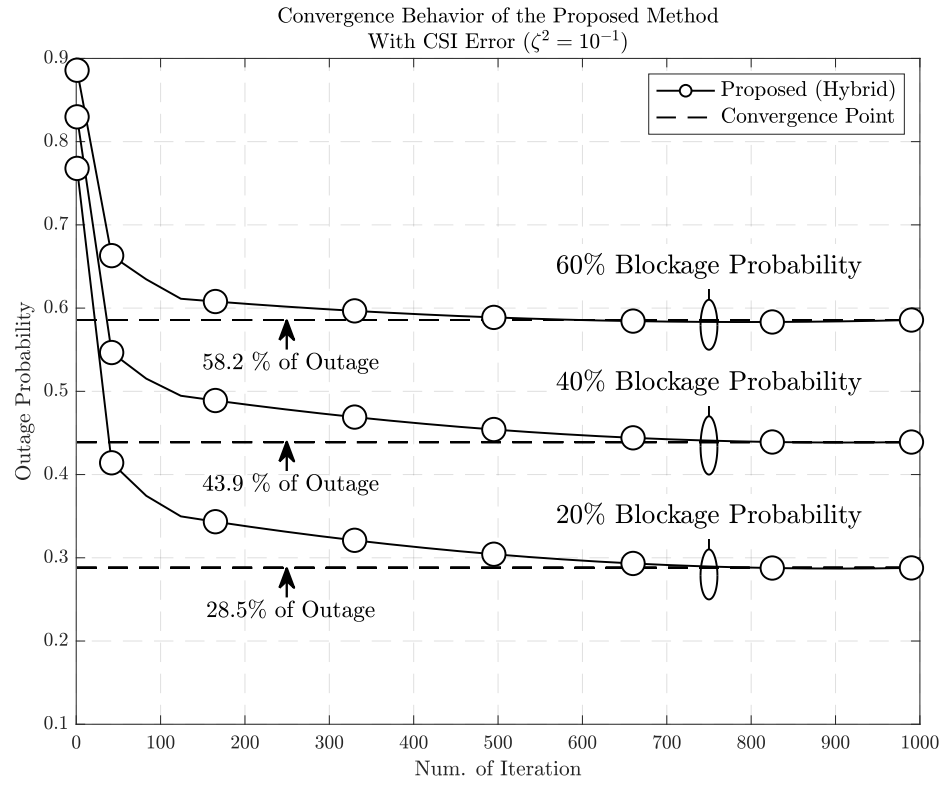
We start our assessment by evaluating the convergence behavior of the proposed hybrid beamforming method in terms of outage probability, which is the objective cost function of interest in this study. For the sake of simplicity and illustration, we assume that the blockage probabilities ($p_{b,u}^k \ \forall \ m, b, u$) are identical, regardless of the superscripts and subscripts, *i.e.*, $p_{b,u}^k = p \ \forall \ m, b, u$, and p is assumed to vary from 20% to 60% as suggested by the findings in [163, 165]. With that in mind, three blockage scenarios are investigated in the simulations, namely, 20%, 40%, and 60%, which are referred to as moderate, severe, and critical blockage conditions, respectively. In turn, the CSI uncertainty $\zeta_{b,u,k}^2$ is considered to be proportional to the corresponding average small-scale fading coefficient, that is, $\zeta_{b,u,k}^2 = \zeta^2 \cdot g_{b,u}^k$, where ζ^2 is a parameter that controls the CSI accuracy level.

The first set of results is presented in Figure 4.3, which depicts the convergence of the proposed iterative hybrid beamforming design for blockage-robust CoMP mmWave systems, as a function of the number of algorithmic iterations, for the three distinct blockage scenarios described above. It is found that regardless of the severity of the scenario in terms of blockage probability, the proposed method quickly learns the crucial blockage patterns and converges to a convergence point, not only under the ideal assumption of perfect CSI typically adopted in related literature, as shown in Figure 4.3a, but also under the more realistic assumption that fluctuation of the small-scale fading coefficients takes place (*i.e.*, imperfect CSI), as shown in Figure 4.3b. In fact, in both cases (perfect or imperfect CSI), and under all blockage scenarios (moderate, severe, and critical), the number of iterations required for the proposed beamformer to converge is approximately the same, which demonstrates the remarkable robustness of the method.

Finally, it is also observed (interestingly but as expected) that the outage performance of the proposed method is dominated by probabilities of random path blockages themselves, rather than by the CSI error variaces. This fact will be elucidated further in the next subsection.



(a) Without CSI Error.



(b) With CSI Error.

Figure 4.3: Convergence behavior of the proposed method for different channel conditions.

4.4.2 Statistical Analysis of Throughput

Next, we numerically analyze, via Monte Carlo simulations, the statistical behavior of the four beamforming methods considered herein—the proposed design, MRT, digital OutMin method, and SRM method—in terms of their achievable throughput and outage probability under different channel conditions.

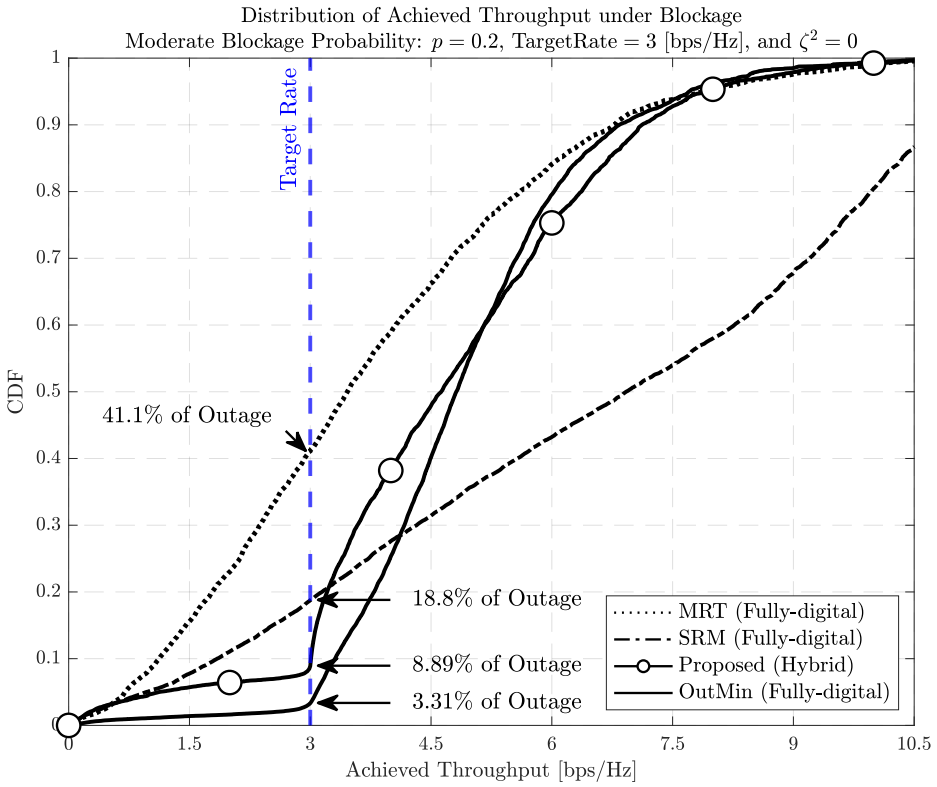
We again emphasize that the fully digital versions of the MRT, OutMin, and SRM schemes are simulated such that the corresponding performances should be considered performance upper-bounds for their hybrid counterparts. Finally, to demonstrate that the advantage of the proposed method extends beyond average performance, all results are offered in the form of throughput cumulative density functions (CDFs).

In Figures 4.4 and 4.5, the CDFs of the achievable throughput obtained by the four beamforming methods, with target rates of 3 and 5 [bps/Hz], respectively, are shown for different blockage scenarios and under perfect CSI, while Figures 4.6 and 4.7 display equivalent results obtained under CSI uncertainty. Since the bandwidth of the system is $W = 100$ [MHz], these target rates imply that in the cases corresponding to Figures 4.4 and 4.6, the system is set to ideally serve each user with at least 300 [Mbps], while in Figures 4.5 and 4.7, the system aims to serve each user with at least 500 [Mbps], respectively.

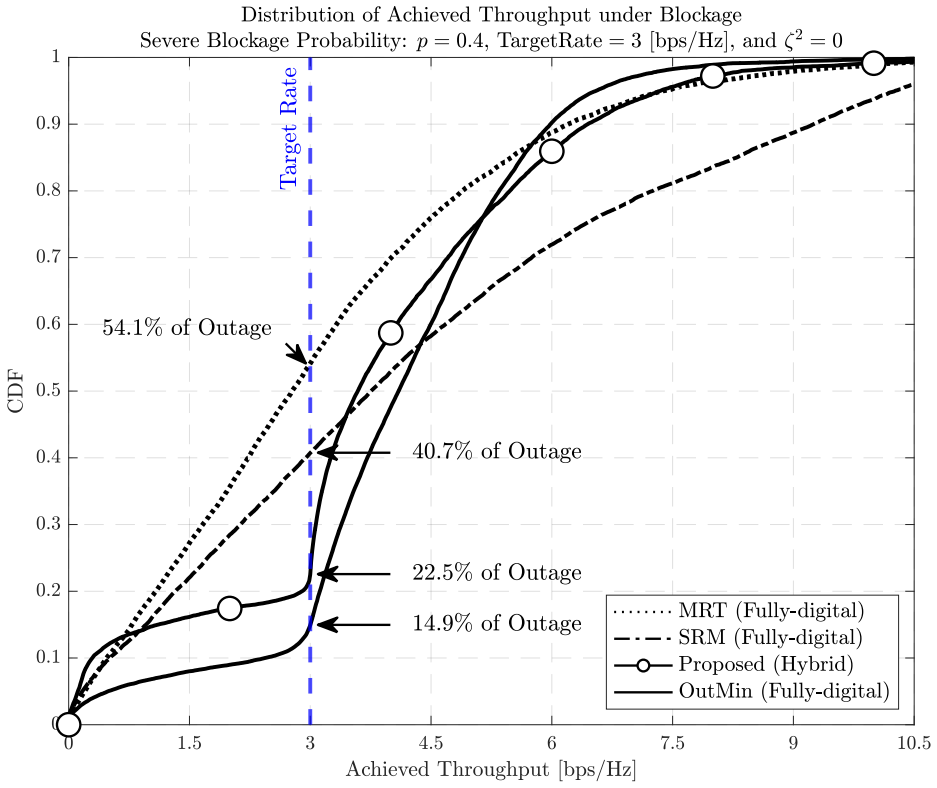
In all figures, lines without markers are used for the SotA methods, while a solid line with a white circular marker indicates the proposed method. For readability, the target throughput is highlighted in the figures by a black vertical line annotated with the text "Target Rate". In addition, the outage probability achieved by each beamforming method — as defined in equation (4.66) — is marked by an arrow also annotated with the corresponding numerical value.

$$\Pr\left\{\log_2(1+\Gamma_u\left(\boldsymbol{f},\boldsymbol{V}|\boldsymbol{h}_u\right)) < \log_2(1+\gamma_u)\right\}. \tag{4.66}$$

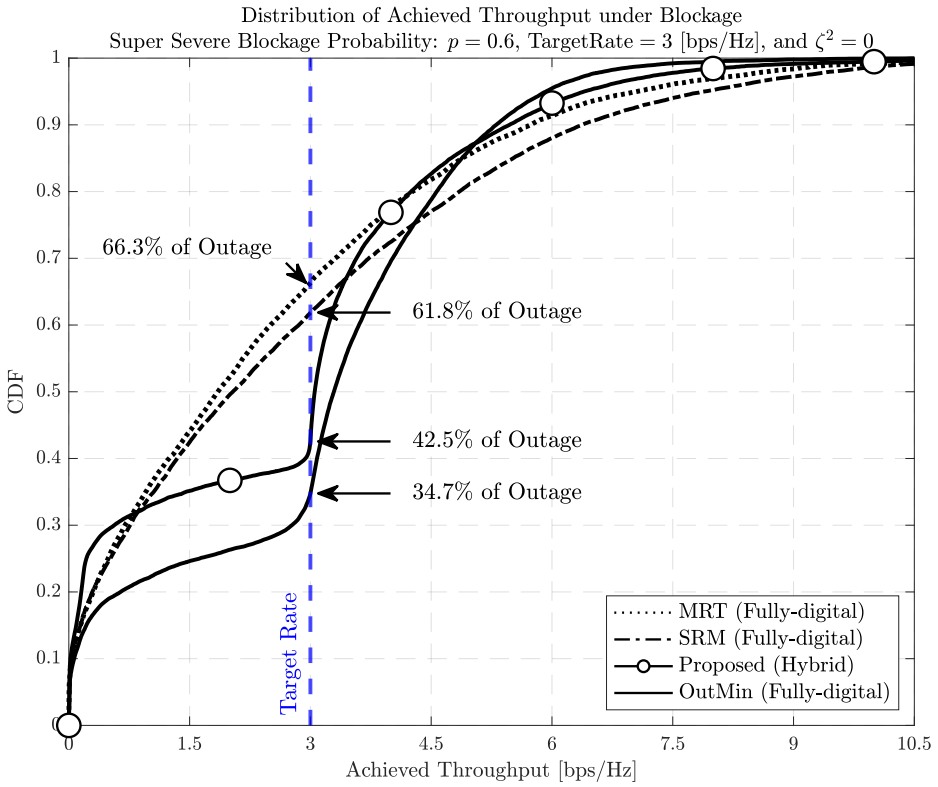
In all the figures, the proposed hybrid method achieves a lower outage probability than the fully digital MRT and SRM beamformers, regardless of CSI quality and blockage conditions. Furthermore, the new technique is found to consistently approach the performance of the also-fully-digital OutMin method. Taking into account that the proposed hybrid scheme makes use of only $N_{\mathrm{RF}} = 2$ RF chains per AP, as opposed to fully digital methods which require all $N_{\mathrm{RF}} = N_t = 16$ RF chains per AP, the results demonstrate the remarkable effectiveness of the proposed hybrid method in combatting path blockage with significant potential to reduce hardware costs (by alleviating RF chain requirements) with reasonable outage performance deterioration.



(a) 20% of Blockage Probability

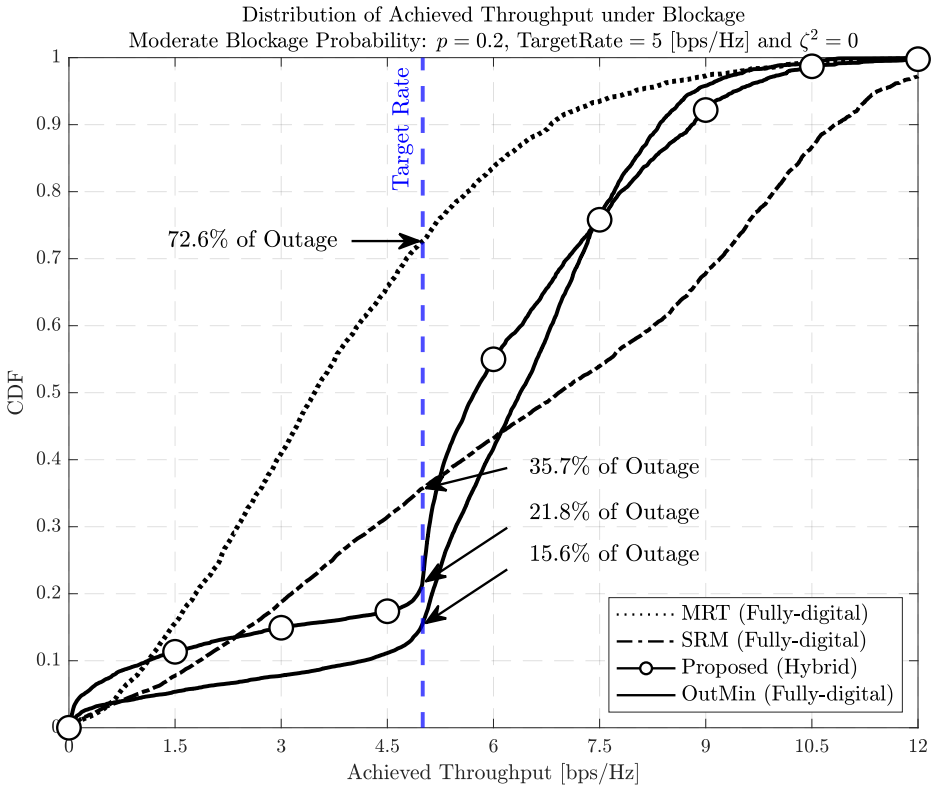


(b) 40% of Blockage Probability

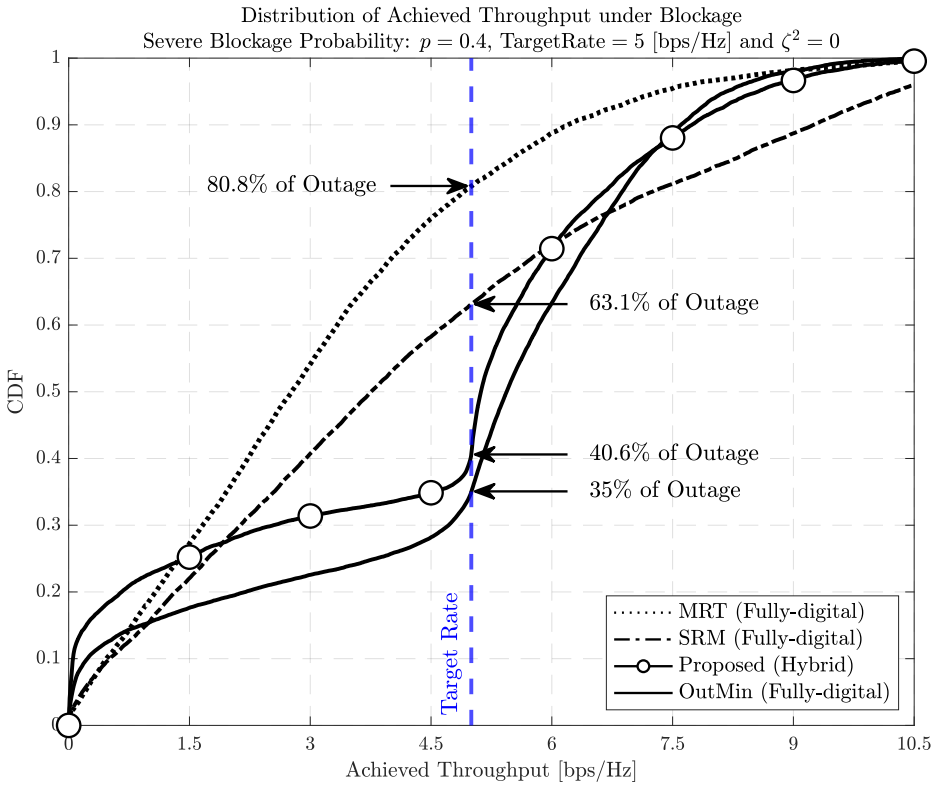


(c) 60% of Blockage Probability

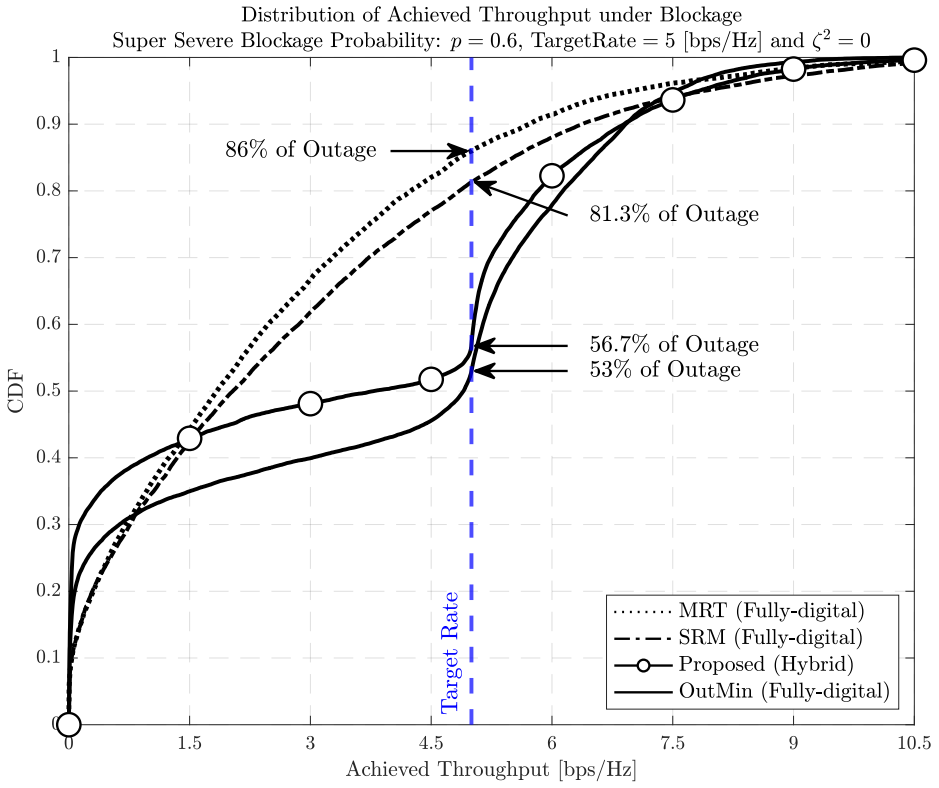
Figure 4.4: CDF of achieved data rates for different blockage probabilities with target rate of 3 [bps/Hz] and perfect CSI.



(a) 20% of Blockage Probability



(b) 40% of Blockage Probability



(c) 60% of Blockage Probability

Figure 4.5: CDF of achieved data rates for different blockage probabilities with target rate of 5 [bps/Hz] and perfect CSI.

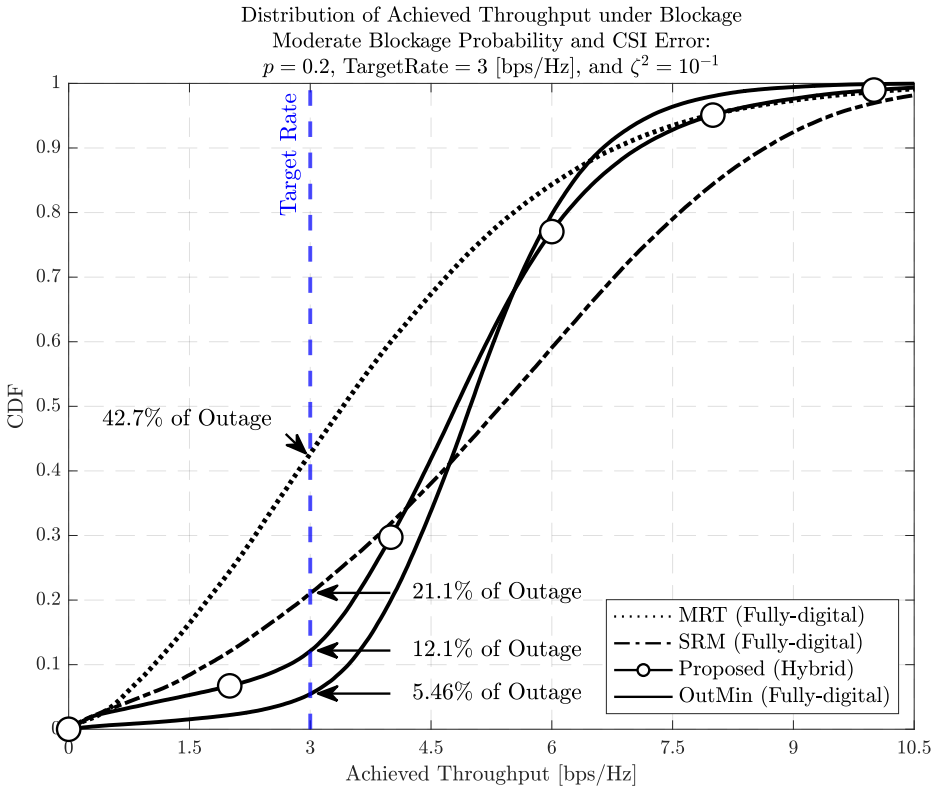
As for the impact of CSI errors, a comparison of Figures 4.4 and 4.5 against Figures 4.6 and 4.7 indicates that the overall impact of CSI errors is to retract some sharpness from the outage minimization approaches in enforcing the prescribed target rate. Even under such an effect, however, both the relative gain over the MRT and SRM beamformers, as well as the proximity in performance of the proposed scheme to the ideal OutMin method, remain mostly unchanged.

To cite a few examples, both under perfect and imperfect CSI, the proposed hybrid CoMP beamformer, under 20% path blockage probability and a target throughput of 3 [bps/Hz], achieves an outage reduction of approximately 30% and 10% over the MRT and SRM beamformers, respectively (see Figures 4.4a and 4.6a). Similarly, at 60% blockage probability and a target rate of 5 [bps/Hz], the new hybrid scheme outperforms the MRT and SRM methods by approximately 30% and 25% in terms of the reduction of outage probability, respectively, both under perfect and imperfect CSI conditions (see Figures 4.5c and 4.7c).

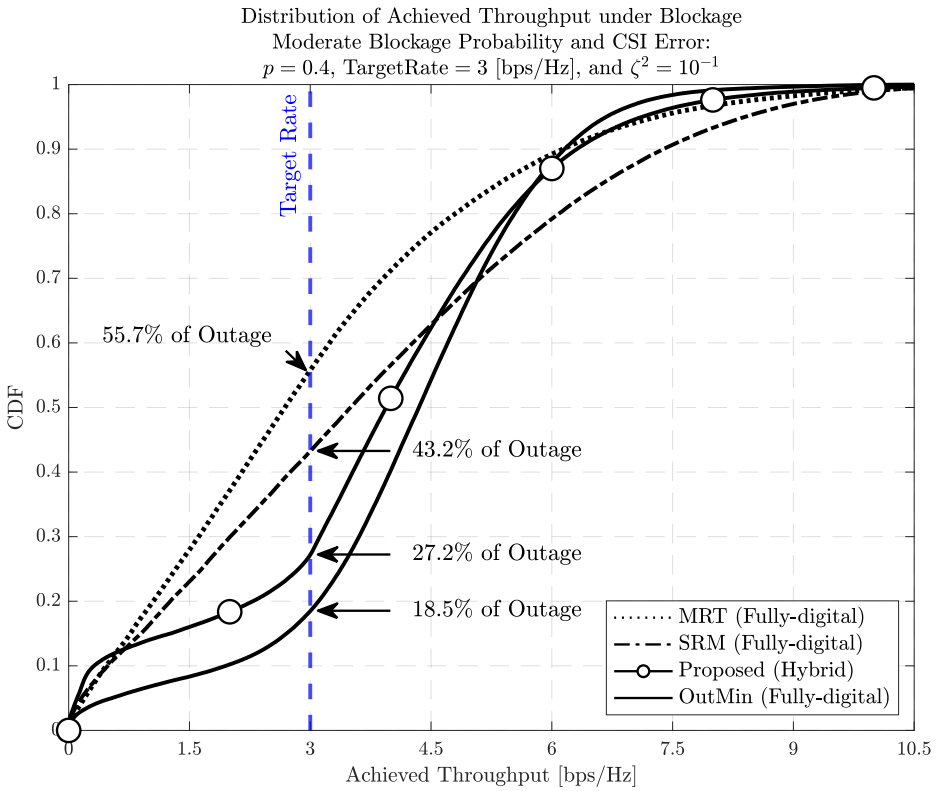
Thus, the results demonstrate that the proposed SGD approach for cooperative hybrid beamforming aimed at minimizing outage in mmWave systems subjected to path blockage and CSI imperfections is highly effective.

An interesting and final conclusion that can be drawn from the comparison of the results shown in Figures 4.4 through 4.7 is that among the considered methods, the MRT, OutMin, and new hybrid techniques are in fact all somewhat robust to CSI imperfections, whereas the SRM schemes seems to be the most sensitive to the quality of channel estimates. This is somewhat unsurprising, since the SRM method is not designed to minimize outage, but rather to maximize the achievable rate, such that one could argue that the aforenoted comparison is "unfair" to that particular approach. Scrutinizing this conclusion is a worthy exercise, nevertheless, as it provides insight on the conceptual question of whether rate maximization is the most suitable figure of merit in the optimization of mmWave systems, which in practice needs to be carefully chosen depending on the system objectives and requirements, as implied by Figures 4.4 through 4.7.

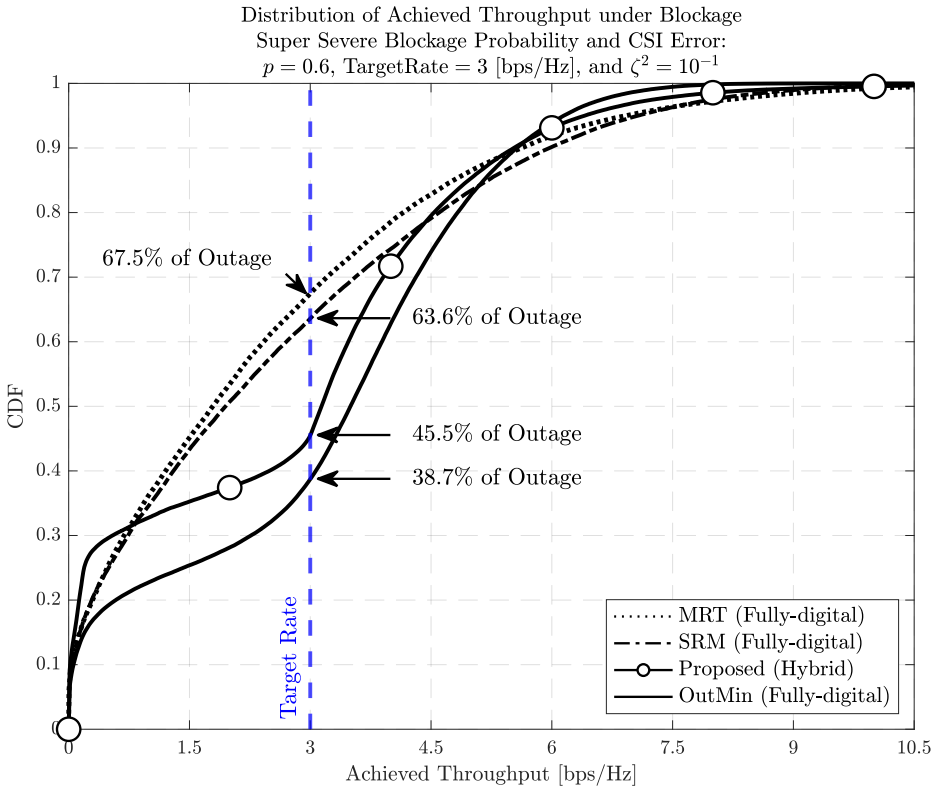
To that end, we compare in Figure 4.8 the effective throughput (*i.e.*, the average throughput at and above the target rate) achieved by each of the considered beamforming schemes, as a function of the path blockage probability and for both evaluated channel estimation conditions, namely, perfect CSI ($\zeta = 0$) and imperfect CSI with $\zeta = 10^{-1}$. Following the related literature, the effective throughput is defined as $\mathbb{E}[\log_2(1+x)]$, where $x = \Gamma_u(\boldsymbol{f}, \boldsymbol{V}|\boldsymbol{h}_u)$ if $\Gamma_u(\boldsymbol{f}, \boldsymbol{V}|\boldsymbol{h}_u) \geq \gamma_u$ and $x = 0$ otherwise.



(a) 20% of Blockage Probability

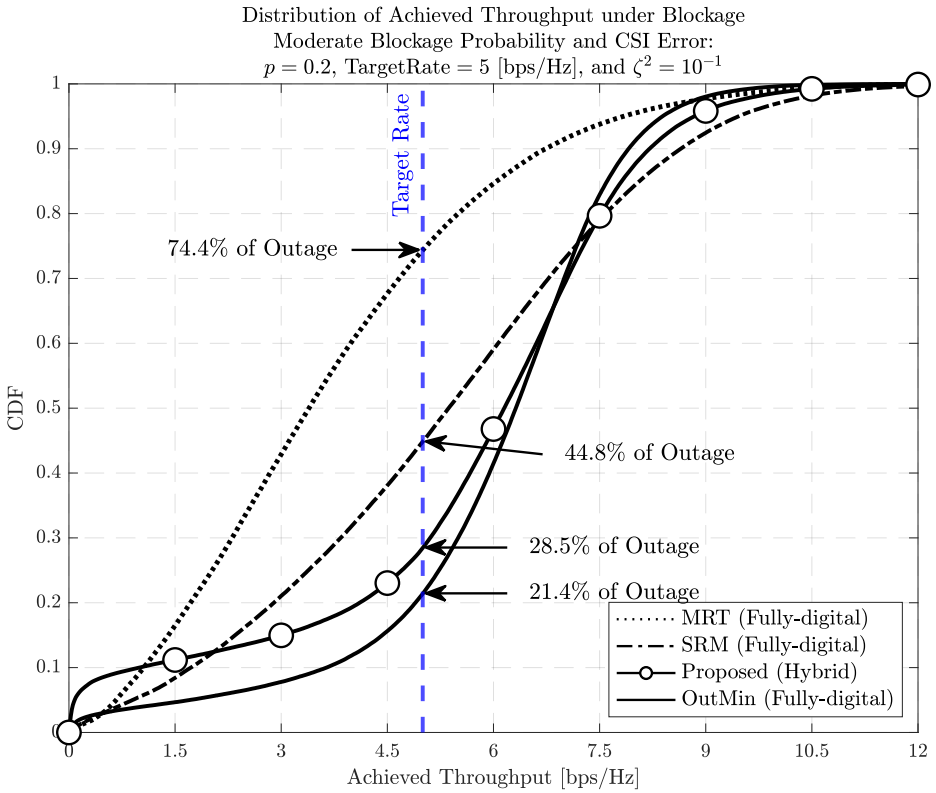


(b) 40% of Blockage Probability

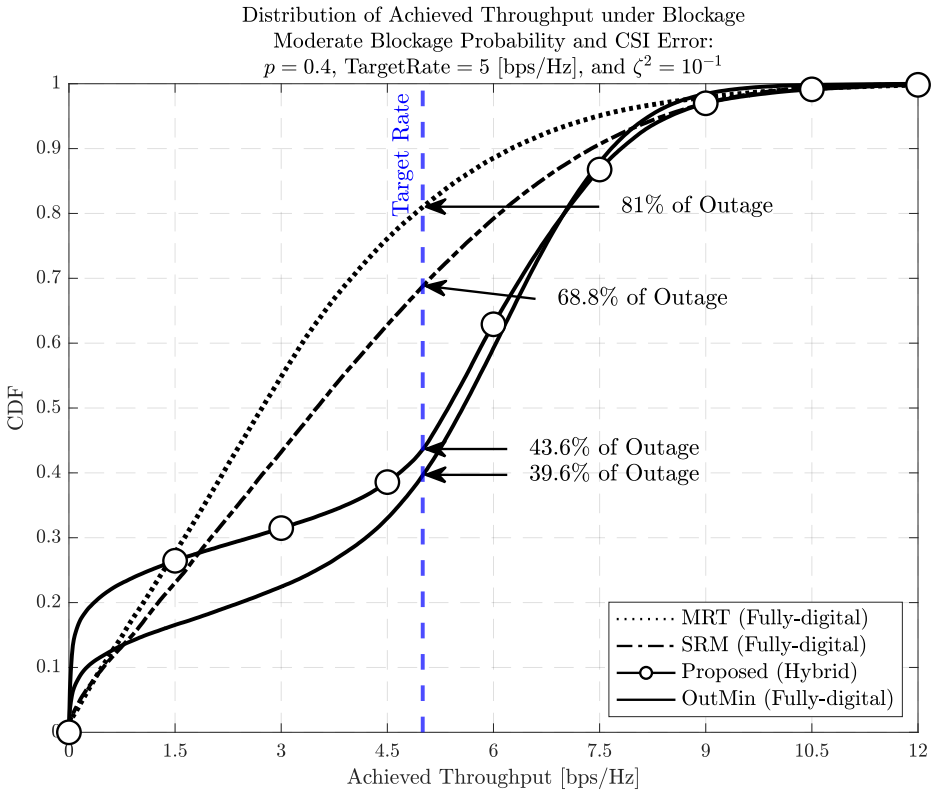


(c) 60% of Blockage Probability

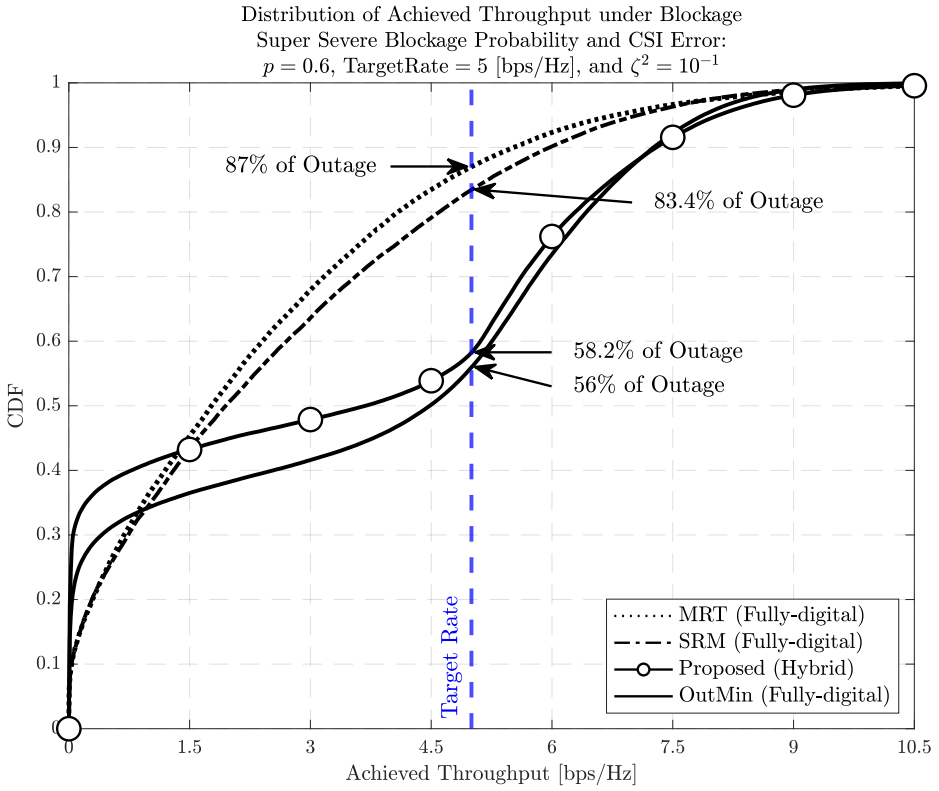
Figure 4.6: CDF of achieved data rates for different blockage probabilities with target rate of 3 [bps/Hz] and CSI errors.



(a) 20% of Blockage Probability

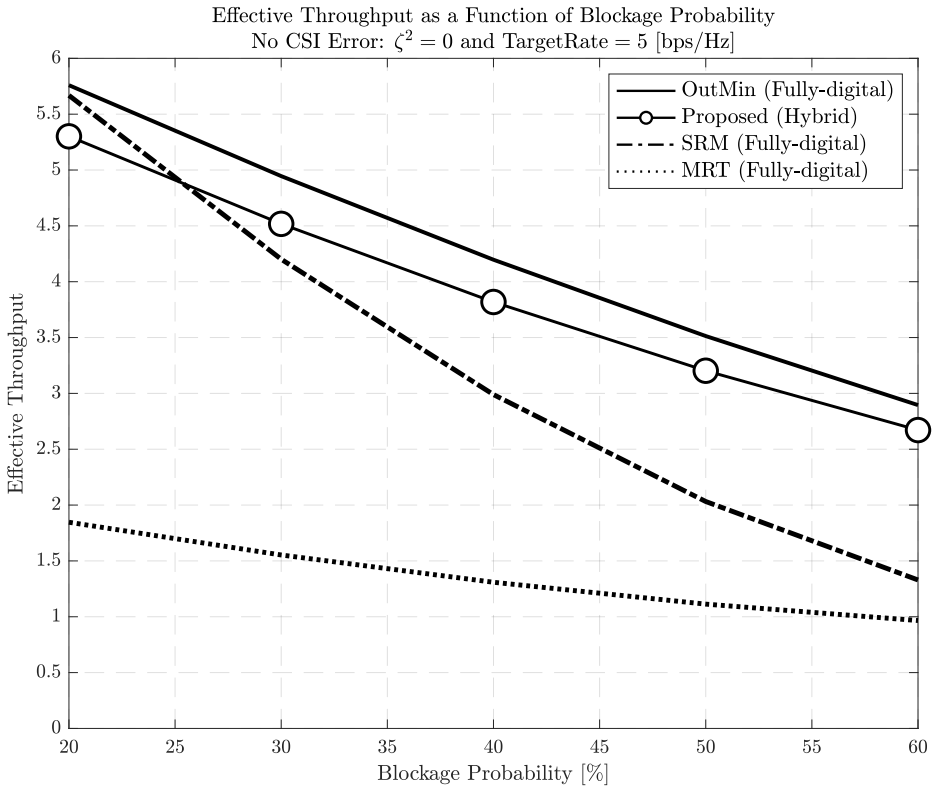


(b) 40% of Blockage Probability

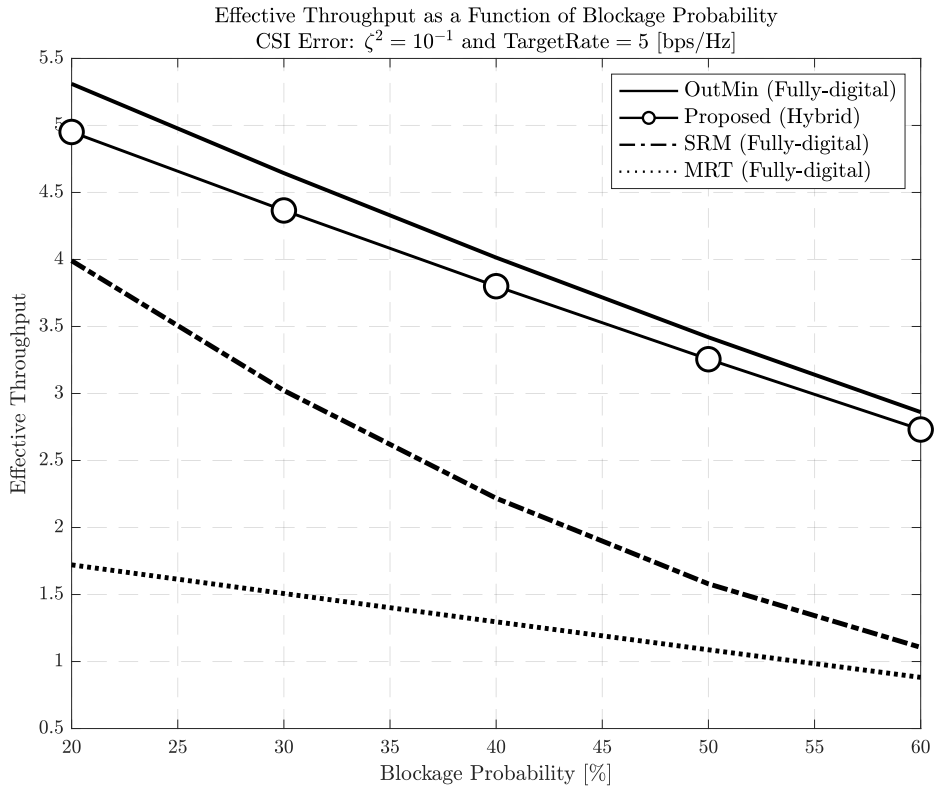


(c) 60% of Blockage Probability

Figure 4.7: CDF of achieved data rates for different blockage probabilities with target rate of 5 [bps/Hz] and CSI errors.



(a) Without CSI Error.



(b) With CSI Error.

Figure 4.8: Effective throughput as a function of the blockage probabilities, with and without CSI uncertainty.

It is found, once again, that the proposed hybrid beamforming scheme performs only slightly worse than its fully digital couterpart, outperforming the other SotA methods, regardless of the CSI conditions and blockage probabilities. The only exception is the case of perfect CSI for blockage probabilities below 25%, where it is found that the SRM scheme outperforms the proposed method. It must be emphasized, however, that: a) the SRM scheme is fully digital, employing a total of $N_t \times B = 16 \times 4 = 64$ RF chains to serve $U = 2$ users, in contrast to the proposed hybrid scheme, which requires only $N_{\text{RF}} \times B = 2 \times 4 = 8$ RF chains; and b) that the SRM scheme is designed to maximize effective throughput, paying to that end the price of sacrificing overall QoS by allowing higher outage probabilities (see *e.g.*, Figure 4.5a where it is shown that the SRM beamformer with perfect CSI under 20% blockage probability leads to outages above 35%).

By contrast, when subjected to CSI errors, even that eventual localized "advantage" of the SRM approach is lost. In fact, as shown in Figure 4.8b, the SRM approach proves the most sensitive to both path blockages and CSI errors, exhibiting a significantly higher degree of degradation in performance than all other methods.

In summary, the results demonstrate collectively that the proposed method offers a competitive approach to mitigate the path blockage challenge in mmWave systems; notably, it also possesses robustness against CSI imperfections and exhibits little sacrifice in performance, compared to the fully digital OutMin approach, for the significant potential reduction in hardware cost due to its hybrid architecture.

4.5 Conclusion

In this chapter, we contributed to the reliability problem in mmWave systems subject to path blockage by proposing a stochastic optimization framework to design a robust beamforming mechanism against such uncertainties. To elaborate, the proposed algorithm aims to minimize the total outage probability characterized by the sum of probabilities that a user SINR quantity is below a certain requirement, unlike the well-studied rate maximization alternatives such as SRM and max-min methods. The motivation behind was that the fundamental goal of robust beamforming design for mmWave systems subject to such path blockages is rather to minimize the QoS violation (*i.e.*, outage probability) in order to run a certain application at each user without any interruption.

Based on this framework, we first proposed a fully-digital beamforming design in order to clarify that the proposed outage minimization approach via stochastic optimization posseses potential to mitigate such random blockage, which is then shown via simulation to be effective for a wide range of blockage probabilities. In order to incorporate practical considerations such as infrastructural limitations and unavoidable

CSI uncertainties, we then proposed a partially-connected hybrid robust beamforming design in conjunction with the Bernoulli-Gaussian blockage model that enables both challenges to be jointly tackled.

The simulation results confirm the robustness of the proposed algorithms in terms of the QoS violation reduction. Also, it is shown via numerical performance assessments that the hybrid robust beamforming method is capable of approaching the full-digital counterpart in terms of outage probability.

Regarding potential future works, one can consider a fast-converging low-complexity alternative of the proposed algorithms by introducing the momentum method and adaptive learning rate tuning as shown in the adaptive moment estimation (ADAM) optimizer well-studied in machine learning literature [197]. In addition to that, a related open problem is to further incorporate other practical aspects such as limited fronthaul links, phase errors at the phase shifters, and frequency-selective beam squint effects in wideband transmissions. Furthermore, since the proposed stochastic optimization framework is generic, one can consider a similar approach to tackle nonlinear effects in hardware. For instance, although most of the hardware impairment aware beamforming designs in full duplex literature (*e.g.*, [74, 198, 199]) assume a linearlized approximation model to capture such nonlinearity caused by nonideal hardware, a similar stochastic approach can directly learn the nonlinearity without resorting to a linear approximation, which is interesting and worthy of further quantitative investigation.

Chapter 5

Final Remarks

In this chapter, we offer in the sequel conclusions and key points of the dissertation. Furthermore, potential future works are listed as derivatives of this research, which are worth pursuing in the future.

5.1 Conclusion

Key potential QoS requirements in modern and future wireless communication systems, that is, *Massive Connetivity*, *Latency*, and *Reliability*, have been addressed by the thesis by means of optimization and Bayesian techniques, proposing several novel algorithms that are shown via simulation to be effective in terms of satisfying the aforementioned key requirements. In summary, main findings offered in each section of the thesis are listed below:

Chapter 2

In this chapter, we developed an efficient and robust receiver framework in non-orthogonal systems with the aim of satisfying the *Massive Connetivity* requirements in future wireless systems. The main findings of this chapter are the following:

- The discreteness-aware regularization approach inspired by CS can be exploited for an inverse problem with discrete inputs, such that the solution is efficiently enforced to be a member of the prescribed discrete constellation set (*e.g.*, QAMs). Note that this constraint of the inputs to be discrete has been ignored by the conventional linear receivers.
- The proposed regularization approach composed of a combination of a newly-introduced asymptotically-tight non-convex ℓ_0-norm approximation and an FP technique leads to a generalization of the conventional ZF and LMMSE estimators with adherence to the prescribed discrete constellation set.

- The proposed framework can then be extended to seemlessly incorporate robustness against practical limitations (*i.e.*, CSI and hardware imperfection), showing its compatibility and flexibility for different system setups.
- A wide range of its applicability is also illustrated by introduing the concept of discreteness-awareness in LRMC problems with discrete entries as is the case with recommnder systems.

Chapter 3

In this chapter, we tackle the overhead reduction in uplink distributed MIMO systems with the aim of reducing the overall latency. We then proposed two Bayesian inference based algorithms with different objectives. The main findings of this chapter are the following:

- In the context of CF-MIMO systems, the goal is to efficiently reduce the overhead due to the use of resource-consuming piloting by a large number of uplink users without sacrificing the per-user throughput. To this end, it is shown that this challenging task can be efficiently addressed by jointly leveraging bilinear inference and the pseudo-orthogonality of the independently-generated data symbols.
- In contrast, in the context of XL-MIMO systems, we turn our attention to one of the peculiarities of XL-MIMO systems (*i.e.*, spatial non-stationarity), which is the fact that the signal from each user is apparent only to distributed portions of the XL-MIMO antenna array, while obeying the low-latency demand in massive uplink channels. This is enabled by exploiting bilinear inference and expectation maximization approaches.
- Notice that although the proposed algorithms are designed for distributed MIMO scenarios, both are also compatible with centralized MIMO systems without any further technical modifications.

Chapter 4

In this chapter, we developed a stochastic optimization framework to design outage-minimum robust beamforming in downlink mmWave systems subject to random path blockage. We start with formulating the outage minimization problem for a full-digital beamforming architecture, extending the latter so as to incorporate practical limitations such as the limited number of RF chains and CSI imperfection. The main findings of this chapter are the following:

- The outage minimization approach is prefered in order to satisfy the motivation that the fundamental goal of robust beamforming in mmWave systems suffering

from random path blockage is to minimize the QoS violation for the users to maintain their running applications.

- A stochastic optimization framework is effective to capture a non-closed-form expression of the objective outage probability.

5.2 Potential Extensions

The results offered in this thesis can be further extended in many different directions. Summarizing the thesis, some possible direct extensions of the work are compactly listed below

- Due to the flexibility of the proposed IDLS framework shown in Chapter 2, the latter can be further extended and applied to a wide range of interference-limited communication scenarios including multi-user multi-cell MIMO, mmWave systems with phase noise, and full-duplex communications. As mentioned earlier, an IDD extension of the IDLS framework is also left for future work.

- Another interesting extension of IDLS is to leverage the deep-unfolding framework [54,200–202] to dynamically control the regularization and tightness parameter by learning, which may lead to better performance and/or faster convergence speed

- As for possible extensions of Chapter 3, one can consider a coded extension of the proposed JACDE framework. An open and debatable question is whether AUD should be performed before or after the channel decoder and how to incorporate such information in the message passing rules. In addition, the design of the detector output is worth considering for coded grant-free systems.

- Another interesting extension of Chapter 3 is to incorporate the data detection capability into Algorithm 5, which is challenging due to the fact that adding the data part introduces a new variable dimension to be jointly estimated; therefore, the resultant estimation problem is no longer bilinear but rather multilinear. A direct approach to this task is to consider generalizing the bilinear inference framework.

- Both Algorithm 3 and 5 assume ideal hardware and channel conditions, providing an upperbound of achievable system performance. To bring robustness against realistic limitations such as carrier frequency offset (CFO) and imperfect clock synchronization, one needs to tailor the message passing rules.

- Regarging extensions of Chapter 4, a low-complexity fast-converging alternative to the proposed robust beamforming designs is essential for real-world implemen-

tations. To that end, the momentum method and adaptive stepsize optimization may be incorporated.

- Thanks to the flexibility of the considered stochastic optimization framework, a similar idea can be taken advantage of to address nonlinearity effects in other communication systems such as full-duplex systems.

Appendix A

Derivations of the Covariance of $\tilde{\boldsymbol{n}}$

Recall from equation (2.56) that the total effective noise, including the contributions due to CSI imperfection and hardware impairment is given by

$$\tilde{\boldsymbol{n}} \triangleq \hat{\boldsymbol{H}}\boldsymbol{w} + \frac{\tau\boldsymbol{E}\boldsymbol{s} + \tau\boldsymbol{E}\boldsymbol{w} + \boldsymbol{n}}{\sqrt{1-\tau^2}}. \tag{A.1}$$

By definition we then have

$$\begin{aligned} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}} &\triangleq \mathbb{E}\big[\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^{\mathrm{H}}\big] \\ &= \mathbb{E}\left[\Big(\hat{\boldsymbol{H}}\boldsymbol{w}+\frac{\tau\boldsymbol{E}\boldsymbol{s}+\tau\boldsymbol{E}\boldsymbol{w}+\boldsymbol{n}}{\sqrt{1-\tau^2}}\Big)\Big(\hat{\boldsymbol{H}}\boldsymbol{w}+\frac{\tau\boldsymbol{E}\boldsymbol{s}+\tau\boldsymbol{E}\boldsymbol{w}+\boldsymbol{n}}{\sqrt{1-\tau^2}}\Big)^{\mathrm{H}}\right]. \end{aligned} \tag{A.2}$$

Distributing the terms in parenthesis and expanding yields

$$\begin{aligned} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}} &= \mathbb{E}\left[\hat{\boldsymbol{H}}\boldsymbol{w}\Big(\hat{\boldsymbol{H}}\boldsymbol{w}+\frac{\tau\boldsymbol{E}\boldsymbol{s}+\tau\boldsymbol{E}\boldsymbol{w}+\boldsymbol{n}}{\sqrt{1-\tau^2}}\Big)^{\mathrm{H}}\right] \\ &\quad+\frac{\tau}{\sqrt{1-\tau^2}}\mathbb{E}\left[\boldsymbol{E}\boldsymbol{s}\Big(\hat{\boldsymbol{H}}\boldsymbol{w}+\frac{\tau\boldsymbol{E}\boldsymbol{s}+\tau\boldsymbol{E}\boldsymbol{w}+\boldsymbol{n}}{\sqrt{1-\tau^2}}\Big)^{\mathrm{H}}\right] \\ &\quad+\frac{\tau}{\sqrt{1-\tau^2}}\mathbb{E}\left[\boldsymbol{E}\boldsymbol{w}\Big(\hat{\boldsymbol{H}}\boldsymbol{w}+\frac{\tau\boldsymbol{E}\boldsymbol{s}+\tau\boldsymbol{E}\boldsymbol{w}+\boldsymbol{n}}{\sqrt{1-\tau^2}}\Big)^{\mathrm{H}}\right] \\ &\quad+\frac{1}{\sqrt{1-\tau^2}}\mathbb{E}\left[\boldsymbol{n}\Big(\hat{\boldsymbol{H}}\boldsymbol{w}+\frac{\tau\boldsymbol{E}\boldsymbol{s}+\tau\boldsymbol{E}\boldsymbol{w}+\boldsymbol{n}}{\sqrt{1-\tau^2}}\Big)^{\mathrm{H}}\right] \\ &= \hat{\boldsymbol{H}}\overbrace{\mathbb{E}\big[\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\big]}^{=\eta\mathbf{I}_{N_t}}\hat{\boldsymbol{H}}^{\mathrm{H}}+\frac{\tau^2\big(\overbrace{\mathbb{E}\big[\boldsymbol{E}\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big]+\mathbb{E}\big[\boldsymbol{E}\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big]}^{=(1+\eta)\mathrm{Tr}(\boldsymbol{\Phi}_t)\boldsymbol{\Phi}_r}\big)}{1-\tau^2} \\ &\quad+\frac{\sigma_{\boldsymbol{n}}^2}{1-\tau^2}\mathbf{I}_{N_r} \\ &= \eta\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^{\mathrm{H}}+\frac{\tau^2}{1-\tau^2}(1+\eta)\mathrm{Tr}\left(\boldsymbol{\Phi}_t\right)\boldsymbol{\Phi}_r+\frac{\sigma_{\boldsymbol{n}}^2}{1-\tau^2}\mathbf{I}_{N_r}, \end{aligned} \tag{A.3}$$

where, in the second-to-last equation we have substituted the expectations $\mathbb{E}\big[\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\big]$, $\mathbb{E}\big[\boldsymbol{E}\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big]$, $\mathbb{E}\big[\boldsymbol{E}\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big]$ and $\mathbb{E}\big[\boldsymbol{n}\boldsymbol{n}^{\mathrm{H}}\big]$, taken over noise realizations and utilizing the identity $\mathbb{E}\big[\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{H}}\big] = \sigma^2\mathrm{Tr}\left(\boldsymbol{A}\right)\mathbf{I}$, valid when the elements of $\boldsymbol{X}$ are zero-mean complex Gaussian variables with variance σ^2.

For the terms $\mathbb{E}\big[\boldsymbol{E}\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big]$ and $\mathbb{E}\big[\boldsymbol{E}\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big]$ in particular, which are subjected to correlation due to the relation $\boldsymbol{E} = \boldsymbol{\Phi}_r^{\frac{1}{2}}\boldsymbol{E}_{\mathrm{i.i.d.}}\boldsymbol{\Phi}_t^{\frac{1}{2}}$, we have used

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{E}\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big] &= \mathbb{E}_{\boldsymbol{E}_{\mathrm{i.i.d.}}}\big[\mathbb{E}_{\boldsymbol{s}}\big[\boldsymbol{E}\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}} \mid \boldsymbol{E}_{\mathrm{i.i.d.}}\big]\big] \\ &= \mathbb{E}_{\boldsymbol{E}_{\mathrm{i.i.d.}}}\left[\boldsymbol{\Phi}_r^{\frac{1}{2}}\boldsymbol{E}_{\mathrm{i.i.d.}}\boldsymbol{\Phi}_t\boldsymbol{E}_{\mathrm{i.i.d.}}^{\mathrm{H}}{\boldsymbol{\Phi}^{\frac{1}{2}}}_r^{\mathrm{H}}\right] = \mathrm{Tr}\left(\boldsymbol{\Phi}_t\right)\boldsymbol{\Phi}_r,\end{aligned} \tag{A.4a}$$

and

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{E}\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}}\big] &= \mathbb{E}_{\boldsymbol{E}_{\mathrm{i.i.d.}}}\big[\mathbb{E}_{\boldsymbol{w}}\big[\boldsymbol{E}\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}\boldsymbol{E}^{\mathrm{H}} \mid \boldsymbol{E}_{\mathrm{i.i.d.}}\big]\big] \\ &= \eta\cdot\mathbb{E}_{\boldsymbol{E}_{\mathrm{i.i.d.}}}\left[\boldsymbol{\Phi}_r^{\frac{1}{2}}\boldsymbol{E}_{\mathrm{i.i.d.}}\boldsymbol{\Phi}_t^{\frac{1}{2}}{\boldsymbol{\Phi}^{\frac{1}{2}}}_t^{\mathrm{H}}\boldsymbol{E}_{\mathrm{i.i.d.}}^{\mathrm{H}}{\boldsymbol{\Phi}^{\frac{1}{2}}}_r^{\mathrm{H}}\right] = \eta\mathrm{Tr}\left(\boldsymbol{\Phi}_t\right)\boldsymbol{\Phi}_r,\end{aligned} \tag{A.4b}$$

where we assume that the correlation matrices are Hermitian and that $\mathbf{C}_{\boldsymbol{s}} = \mathbf{I}_{N_t}$.

We also remark that in case of $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}_r = \mathbf{I}$, the covariance matrix of the effective noise becomes a function of the dimensionality of the transmit symbols, thus directly affected by the overloading factor γ.

Appendix B

Derivations of QCSIDCO

Leveraging the slack variables $t_{\ell,R}$ and $t_{\ell,I}$, the real and imaginary parts of $\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell$ can be respectively bounded as

$$\left|\Re\left\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell\right\}\right| \leq t_{\ell,R}\cdot\mathbf{1}_{L-1} \quad \text{and} \quad \left|\Im\left\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell\right\}\right| \leq t_{\ell,I}\cdot\mathbf{1}_{L-1}, \tag{B.1}$$

where the inequality is applied in an element-by-element manner.

From equation (B.1), one can readily obtain $\|\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell\|_\infty \leq \sqrt{t_{\ell,R}^2+t_{\ell,I}^2}$. Furthermore, equation (B.1) can also be rewritten as

$$\underbrace{\begin{cases}\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\}+\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\}-t_{\ell,R}\cdot\mathbf{1}_{L-1}\leq\mathbf{0}\\ -\left(\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\}+\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\}\right)-t_{\ell,R}\cdot\mathbf{1}_{L-1}\leq\mathbf{0}\end{cases}}_{\Leftrightarrow\left|\Re\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell\}\right|-t_{\ell,R}\cdot\mathbf{1}_{L-1}\leq\mathbf{0}} \tag{B.2a}$$

$$\underbrace{\begin{cases}\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\}-\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\}-t_{\ell,I}\cdot\mathbf{1}_{L-1}\leq\mathbf{0}\\ -\left(\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\}-\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\}\right)-t_{\ell,I}\cdot\mathbf{1}_{L-1}\leq\mathbf{0}\end{cases}}_{\Leftrightarrow\left|\Im\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell\}\right|-t_{\ell,I}\cdot\mathbf{1}_{L-1}\leq\mathbf{0}} \tag{B.2b}$$

leading to

$$\underbrace{\left[\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ -\mathbf{1}_{(L-1)\times1}\ \mathbf{0}_{(L-1)\times1}\right]}_{\triangleq\boldsymbol{A}_{\ell,R,1}}\boldsymbol{x}_\ell \leq \mathbf{0} \tag{B.3a}$$

$$\underbrace{\left[-\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ -\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ -\mathbf{1}_{(L-1)\times1}\ \mathbf{0}_{(L-1)\times1}\right]}_{\triangleq\boldsymbol{A}_{\ell,R,2}}\boldsymbol{x}_\ell \leq \mathbf{0} \tag{B.3b}$$

$$\underbrace{\left[-\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ \Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ \mathbf{0}_{(L-1)\times 1}\ -\mathbf{1}_{(L-1)\times 1}\right]}_{\triangleq \boldsymbol{A}_{\ell,I,1}} \boldsymbol{x}_\ell \leq \ \mathbf{0} \tag{B.3c}$$

$$\underbrace{\left[\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ -\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\ \mathbf{0}_{(L-1)\times 1}\ -\mathbf{1}_{(L-1)\times 1}\right]}_{\triangleq \boldsymbol{A}_{\ell,I,2}} \boldsymbol{x}_\ell \leq \ \mathbf{0} \tag{B.3d}$$

where $\boldsymbol{x}_\ell$ is defined in Theorem 1, equation (3.7f) can be readily obtained from equation (3.5b), and this completes the proof.

Appendix C

Derivation of equation (3.24) and (3.25)

Given equation (3.18) and (3.21b), the effective PDF can be readily expressed as

$$
\begin{aligned}
&p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)\\
&= p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)\mathcal{CN}_N\left(\boldsymbol{\mu}^h_{k,m},\boldsymbol{\Sigma}^h_{k,m}\right)\\
&= \left[\lambda\,\mathcal{CN}_N\left(0,\boldsymbol{\Gamma}_m\right)+(1-\lambda)\delta(\boldsymbol{h}_m)\right]\mathcal{CN}_N\left(\boldsymbol{\mu}^h_{k,m},\boldsymbol{\Sigma}^h_{k,m}\right)\\
&= \Bigg[\frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\mathcal{CN}_N\left(\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}+\boldsymbol{\Gamma}^{-1}_m\right)^{-1}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\mu}^h_{k,m},\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}+\boldsymbol{\Gamma}^{-1}_m\right)^{-1}\right)\\
&\quad+\frac{(1-\lambda)\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Sigma}^h_{k,m}|}\delta(\boldsymbol{h}_m)\Bigg]
\end{aligned}
\tag{C.1}
$$

where by the Woodbury inverse lemma we obtained $\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}-\boldsymbol{\Sigma}^{h\,-1}_{k,m}\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}+\boldsymbol{\Gamma}^{-1}_m\right)^{-1}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\right)=\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}$. Recalling that $\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}=\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}$ for inversible $\boldsymbol{A}$ and $\boldsymbol{B}$, one may readily obtain equation (3.24) from (C.1). This completes the derivation of (3.24).

Similarly, the normalizing factor $C_{k,m}$ is given by

$$
\begin{aligned}
C_{k,m}&\triangleq\int_{\boldsymbol{h}'_m}p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}'_m) \qquad\qquad (\text{C.2})\\
&=\frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}+\frac{(1-\lambda)\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Sigma}^h_{k,m}|}\delta(\boldsymbol{h}_m)\\
&=\frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\cdot\left(1+\frac{1-\lambda}{\lambda}\left|\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\Gamma}_m+\mathbf{I}_N\right|\exp\left(-\pi^h_{k,m}\right)\right),
\end{aligned}
$$

which completes the derivation of equation (3.25).

Appendix D

Own Publications

First-Authored Journal Papers:

[J1] **Hiroki Iimori**, Takumi Takahashi, Koji Ishibashi, Giuseppe Thadeu Freitas de Abreu, and Wei Yu: "Grant-Free Access via Bilinear Inference for Cell-Free MIMO with Low-Coherence Pilots," to appear in *IEEE Trans. Wireless Commun.*, 2021.

[J2] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, David González G, and Osvaldo Gonsa: "Mitigating Channel Aging and Phase Noise in Millimeter Wave MIMO Systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 7237–7242, Jul. 2021.

[J3] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, Omid Taghizadeh, Razvan-Andrei Stoica, Takanori Hara, and Koji Ishibashi: "A Stochastic Gradient Descent Approach for Hybrid MmWave Beamforming with Blockage and CSI-Error Robustness," *IEEE Access*, vol. 9, pp. 74471–74487, May 2021.

[J4] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, Takanori Hara, Koji Ishibashi, Razvan-Andrei Stoica, David González G, and Osvaldo Gonsa: "Robust Symbol Detection in Large-Scale Overloaded NOMA Systems," *IEEE Open J. Commun. Society*, vol. 2 pp. 512–533, Mar. 2021.

[J5] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, Omid Taghizadeh, Razvan-Andrei Stoica, Takanori Hara and Koji Ishibashi,: "Stochastic Learning Robust Beamforming for Millimeter-Wave Systems with Path Blockage," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1557–1561, Sept. 2020.

[J6] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu and George Alexandropoulos: "MIMO Beamforming Schemes for Hybrid SIC FD Radios with Imperfect HW and

CSI," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4816–4830, Oct. 2019.

Co-Authored Journal Papers:

[J7] Kengo Ando, **Hiroki Iimori**, Takumi Takahashi, Koji Ishibashi, and Giuseppe Thadeu Freitas de Abreu: "Uplink Signal Detection for Scalable Cell-Free Massive MIMO Systems with Robustness to Rate-Limited Fronthaul," *IEEE Access*, vol. 9, pp. 102770–102782, Jul. 2021.

[J8] Omid Taghizadeh, Slawomir Stanczak, **Hiroki Iimori**, and Giuseppe Thadeu Freitas de Abreu: "Full-Duplex Amplify-and-Forward MIMO Relaying: Impairments Aware Design and Performance Analysis," *IEEE Open J. Commun. Society*, vol. 2, pp. 1249–1266, Jun. 2021.

[J9] Takanori Hara, **Hiroki Iimori**, and Koji Ishibashi,: "Hyperparameter-Free Receiver for Grant-Free NOMA Systems With MIMO-OFDM," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 810–814, Apr. 2021.

[J10] Naoya Hirosawa, **Hiroki Iimori**, Koji Ishibashi, and Giuseppe Thadeu Freitas de Abreu: "Minimizing Age of Information in Energy Harvesting Wireless Sensor Networks," *IEEE Access*, vol. 8, pp. 219934-219945, Nov. 2020.

[J11] Razvan-Andrei Stoica, **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, Koji Ishibashi: "Frame Theory and Fractional Programming for Sparse Recovery-based mmWave Channel Estimation," *IEEE Access*, vol. 7, pp. 150757–150774, Oct. 2019.

First-Authored Conference Papers:

[C1] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, and Koji Ishibashi: "Full-Duplex MIMO Systems with Hardware Limitations and Imperfect Channel Estimation," *Proc. IEEE Global Communications Conference (GLOBECOM)*, Taipei, Taiwan, Dec. 2020.

[C2] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, Omid Taghizadeh, and Koji Ishibashi: "Discrete-Aware Matrix Completion via Proximal Gradient," *Proc. Asilomar Conference on Signals, Systems, and Computers*, Nov. 2020.

[C3] **Hiroki Iimori**, R.-A. Stoica, K. Ishibashi, and G. T. F. de Abreu: "Robust sparse reconstruction of mmwave channel estimates via fractional programming," *Proc. ICOIN*, Barcelona, Spain, Jan. 2020.

[C4] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, David Gonzales and Osvaldo Gonsa: "Joint Detection in Massive Overloaded Wireless Systems via Mixed-Norm Discrete Vector Decoding", *Proc. Asilomar Conference on Signals, Systems, and Computers*, Nov. 2019.

[C5] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, and Koji Ishibashi: "Fractional Programming for Robust TX BF Design in Multi-User/Single-Carrier PD-NOMA", *Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks* (WiOpt 2019), Jun. 2019.

[C6] **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, George Alexandropoulos and Koji Ishibashi: "Transmission Strategies in Imperfect Bi-directional Full-Duplex MIMO Systems", *Proc. IEEE Wireless Communications and Networking Conference*, (WCNC 2019), Apr. 2019.

Co-Authored Conference Papers:

[C7] Ryo Okabe, **Hiroki Iimori**, and Koji Ishibashi: "Low-Complexity Robust Beamforming with Blockage Prediction for Millimeter-Wave Communications," *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, Auckland, New Zealand, Dec. 2020.

[C8] Omid Taghizadeh, Slawomir Stanczak, **Hiroki Iimori**, and Giuseppe Thadeu Freitas de Abreu: "Full-Duplex AF MIMO Relaying: Impairments Aware Design and Performance Analysis," *Proc. IEEE Global Communications Conference (GLOBECOM)*, Taipei, Taiwan, Dec. 2020.

[C9] Takanori Hara, **Hiroki Iimori**, and Koji Ishibashi,: "Activity Detection for Uplink Grant-Free NOMA in the Presence of Carrier Frequency Offsets," *Proc. IEEE Int. Conf. Commun.*, Dublin, Ireland, Jun. 2020.

[C10] Razvan-Andrei Stoica, **Hiroki Iimori** and Giuseppe Thadeu Freitas de Abreu: "Multiuser detection for large massively concurrent NOMA systems via fractional programming," *Proc. IEEE International Workshop on Computational Advances in*

Multi-Sensor Adaptive Processing (CAMSAP), Guadeloupe, France, 2019.

[C11] Naoya Hirosawa, **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, and Koji Ishibashi: "Age-of-Information Minimization in Two-User Multiple Access Channel with Energy Harvesting," *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Guadeloupe, France, 2019.

[C12] Razvan-Andrei Stoica, **Hiroki Iimori** and Giuseppe Thadeu Freitas de Abreu: "Sparsely-structured Multiuser Detection for Large Massively Concurrent NOMA Systems", *Proc. Asilomar Conference on Signals, Systems, and Computers*, 3–6 Nov., 2019.

Unpublished Articles:

[A1] **Hiroki Iimori**, Takumi Takahashi, Koji Ishibashi, Giuseppe Thadeu Freitas de Abreu, David González G., and Osvaldo Gonsa: "Joint Activity and Channel Estimation for Extra-Large MIMO Systems Subject to Non-Stationarities," submitted to the IEEE for possible publication.

[A2] Hyeon Seok Rou, **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, David González G., and Osvaldo Gonsa: "Scalable Quadrature Spatial Modulation," submitted to the IEEE for possible publication.

[A3] Omid Taghizadeh, **Hiroki Iimori**, and Giuseppe Thadeu Freitas de Abreu: "Quantization-Aided Secrecy: FD C-RAN Communications with Untrusted Radios," submitted to the IEEE for possible publication.

[A4] Andre S. Guerreiro, **Hiroki Iimori**, and Koji Ishibashi: "Low Latency Beam-Sweeping for Millimeter Wave Systems via Pessimistic Optimization," submitted to the IEEE for possible publication.

[A5] Kengo Ando, **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, and Koji Ishibashi: "User-Heterogeneous Cell-Free Massive MIMO Beamforming via Tensor Decomposition," submitted to the IEEE for possible publication.

[A6] Shuto Fukue, **Hiroki Iimori**, Giuseppe Thadeu Freitas de Abreu, and Koji Ishibashi: "Dynamic TDD in Cell-Free MIMO: Tradeoff between Throughput and User-Fairness," in preparation for submission to the IEEE for possible publication.

[A7] Takumi Takahashi, **Hiroki Iimori**, Kengo Ando, Koji Ishibashi, Shinsuke Ibi, and Giuseppe Thadeu Freitas de Abreu: "Bayesian JCDE for Scalable Cell-Free MIMO Systems with Low-Resolution ADCs," in preparation for submission to the IEEE for possible publication.

Bibliography

[1] D. Seo, *Evolution and Standardization of Mobile Communications Technology*. Information Science Reference, 2013.

[2] G. Punz, *Evolution of 3G Networks*. Springer-Verlag Wien, 2010.

[3] "IMT vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," ITU-R Recommendation M.2083-0., Sep. 2015.

[4] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 2, pp. 905–929, Secondquarter 2020.

[5] 3GPP, TS 38.104, "NR; Base Station (BS) radio transmission and reception," Release 16.

[6] Ericsson, "Mobile data traffic outlook," Jun. 2020.

[7] ITU-R, "IMT traffic estimates for the years 2020 to 2030," Jul. 2015.

[8] Ericsson, "Mobile network traffic update Q1 2021," 2021.

[9] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Society*, vol. 2, pp. 334–366, Feb. 2021.

[10] N. Rajatheva *et al.*, "Scoring the terabit/s goal: Broadband connectivity in 6G," ArXiv, Feb. 2021.

[11] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[12] L. Mucchi, S. Jayousi, S. Caputo, E. Panayirci, S. Shahabuddin, J. Bechtold, I. Morales, R.-A. Stoica, G. Abreu, and H. Haas, "Physical-layer security in 6G networks," *IEEE Open J. Commun. Society*, vol. 2, pp. 1901–1914, Aug. 2021.

[13] E. C. Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," ArXiv, 2021.

[14] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Orlando, USA, Mar. 2012.

[15] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. Hendler, "Towards a theory of semantic communication," in *IEEE Network Science Workshop*, 2011, pp. 110–117.

[16] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1426–1438, Aug. 2006.

[17] G. Scutari, D. P. Palomar, F. Facchinei, and J. shi Pang, "Convex optimization, game theory, and variational inequality theory," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 35–49, May 2010.

[18] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multicell systems," *Foundations Trends Commun. Inf. Theory*, vol. 9, no. 2–3, pp. 113–381, Jan. 2013.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, 1991.

[20] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations Trends Commun. Inf. Theory*, vol. 11, no. 3, pp. 185–396, Jun. 2015.

[21] K. Shen and W. Yu, "Fractional programming for communication systems – Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[22] M. D ur, R. Horst, and N. V. Thoai, "Solving sum-of-ratios fractional programs using efficient points," *Optimization*, vol. 49, no. 5–6, pp. 447–466, 2001.

[23] P. Shen, W. Li, and X. Bai, "Maximizing for the sum of ratios of two convex functions over a convex set," *Computers & Operations Research*, vol. 40, no. 10, pp. 2301–2307, Oct. 2013.

[24] T. Kuno, "A branch-and-bound algorithm for maximizing the sum of several linear ratios," *J. Global Optimization*, vol. 22, pp. 155–174, 2002.

[25] A. Ajagekar, T. Humble, and F. You, "Quantum computing based hybrid solution strategies for large-scale discrete-continuous optimization problems," *Computers & Chemical Engineering*, vol. 132, Jan. 2020.

[26] A. Hjorungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.

[27] K. Kreutz-Delgado. The complex gradient operator and the $\mathbb{CR}$–calculus. [Online]. Available: https://arxiv.org/abs/0906.4835

[28] L. V. Ahlfors, *Complex Analysis.* McGraw-Hill, 1979.

[29] L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, 1991.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, 2006.

[31] L. Hanzo, T. Liew, B. Yeap, and R. Tee, *Turbo Coding, Turbo Equalisation and Space-Time Coding: EXIT-Chart Aided Near-Capacity Designs for Wireless Channels.* John Wiley & Sons, 2010.

[32] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009. [Online]. Available: https://www.pnas.org/content/106/45/18914

[33] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. Uncertainty in Artificial Intelligence*, Aug. 2001, pp. 362–369.

[34] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

[35] W. Yuan, F. Liu, C. Masouros, J. Yuan, D. W. K. Ng, and N. González-Prelcic, "Bayesian predictive beamforming for vehicular networks: A low-overhead joint radar-communication approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1442–1456, Mar. 2021.

[36] L. Wang, T. Takahashi, S. Ibi, and S. Sampei, "Information-optimum approximate message passing for quantized massive MIMO detection," *IEEE Access*, vol. 8, pp. 200 383–200 394, 2020.

[37] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[38] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sept. 2018.

[39] C. Qian, J. Wu, Y. R. Zheng, and Z. Wang, "Two-stage list sphere decoding for under-determined multiple-input multiple-output systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, pp. 6476–6487, 2013.

[40] R. Hayakawa and K. Hayashi, "Convex optimization-based signal detection for massive overloaded MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7080–7091, Nov. 2017.

[41] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing for overloaded massive MIMO-NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, Jan. 2018.

[42] H. Iimori, G. Abreu, D. Gonzales, and O. Gonsa, "Joint detection in massive overloaded wireless systems via mixed-norm discrete vector decoding," in *Proc. Asilomar CSSC*, Pacific Grove, USA, 2019.

[43] R.-A. Stoica, H. Iimori, and G. Abreu, "Sparsely-structured multiuser detection for large massively concurrent NOMA systems," in *Proc. Asilomar CSSC*, Pacific Grove, USA, 2019.

[44] R.-A. Stoica, G. Abreu, T. Hara, and K. Ishibashi, "Massively concurrent non-orthogonal multiple access for 5G networks and beyond," *IEEE Access*, vol. 7, pp. 82 080–82 100, Jun. 2019.

[45] R. Hayakawa and K. Hayashi, "Reconstruction of complex discrete-valued vector via convex optimization with sparse regularizers," *IEEE Access*, vol. 6, pp. 66 499–66 512, Oct. 2018.

[46] L. Liu, Y. Chi, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving MIMO-NOMA: Iterative LMMSE detection," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1758–1773, Apr. 2019.

[47] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-free massive random access with a massive MIMO receiver," in *Proc. Asilomar CSSC*, Pacific Grove, USA, 2019.

[48] A. Aïssa-El-Bey, D. Pastor, S. M. A. Sbaï, and Y. Fadlallah, "Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2008–2018, Apr. 2015.

[49] T. Wo and P. A. Hoeher, "A simple iterative gaussian detector for severely delay-spread MIMO channels," in *Proc. IEEE ICC*, Glasgow, UK, 2007.

[50] Y. Fadlallah, A. Aïssa-El-Bey, K. Amis, D. Pastor, and R. Pyndiah, "New iterative detector of MIMO transmission using sparse decomposition," *IEEE Trans. Veh. Technol.*, vol. 64, no. 8, pp. 3458–3464, Aug. 2015.

[51] T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Low-complexity near-optimal signal detection in underdetermined large-MIMO systems," in *Proc. National Conf. Commun.*, Kharagpur, India 2012.

[52] Z. Hajji, A. Aissa-El-Bey, and K. A. Cavalec, "Simplicity-based recovery of finite-alphabet signals for large-scale MIMO systems," *Digital Signal Process.*, vol. 80, pp. 70–82, 2018.

[53] F. Wen, L. Chu, P. Liu, and R. C. Qiu, "A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning," *IEEE Access*, vol. 6, pp. 69 883–69 906, 2018.

[54] D. Ito, S. Takabe, and T. Wadayama, "Trainable ista for sparse signal recovery," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3113–3125, Jun. 2019.

[55] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing.* Springer Verlag, Aug. 2012.

[56] A. Hjørungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.

[57] F. Rosário, F. A. Monteiro, and A. Rodrigues, "Fast matrix inversion updates for massive MIMO detection and precoding," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 75–79, 2016.

[58] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[59] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized m-estimators in high dimensions," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5592–5628, Aug. 2018.

[60] C. F. Mecklenbräuker, P. Gerstoft, and E. Zöchmann, "c–LASSO and its dual for sparse signal estimation from array data," *Signal Processing*, vol. 130, pp. 204–216, 2017.

[61] L. Luzzi, D. Stehlé, and C. Ling, "Decoding by embedding: Correct decoding radius and DMT optimality," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2960–2973, May 2013.

[62] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, *5G Physical Layer: principles, models and technology components.* Academic Press, 2018.

[63] J. J. Moré, "Generalizations of the trust region problem," *Optim. Methods Softw.*, vol. 2, no. 3–4, pp. 189–209, 1993.

[64] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. New York, NY: Johns Hopkins Univ. Press, Nov. 1996.

[65] S. Adachi and Y. Nakatsukasa, "Eigenvalue-based algorithm and analysis for nonconvex QCQP with one constraint," *Math. Program.*, vol. 173, no. 1–2, pp. 79–116, Jan. 2019.

[66] H. Cai, M. F. Kaloorazi, J. Chen, W. Chen, and C. Richard, "Online dominant generalized eigenvectors extraction via a randomized method," in *Proc. EUSIPCO*, 2020.

[67] Z. Xu and P. Li, "A practical riemannian algorithm for computing dominant generalized eigenspace," in *Proc. Conf. Uncertainty in Artificial Intell. (UAI)*, 2020.

[68] J. Rommes, "Arnoldi and Jacobi-Davidson methods for generalized eigenvalue problems $Ax = \lambda Bx$ with singular B," *Mathematics of Computation*, vol. 77, no. 262, pp. 995–1015, Apr. 2007.

[69] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[70] 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.

[71] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. USA: Cambridge University Press, 2007.

[72] B. Nosrat-Makouei, J. G. Andrews, and R. W. Heath, "MIMO interference alignment over correlated channels with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2783–2794, Jun. 2011.

[73] H. Suzuki, T. V. A. Tran, I. B. Collings, G. Daniels, and M. Hedley, "Transmitter noise effect on the performance of a MIMO-OFDM hardware implementation achieving improved coverage," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 867–876, Aug. 2008.

[74] O. Taghizadeh, A. C. Cirik, and R. Mathar, "Hardware impairments aware transceiver design for full-duplex amplify-and-forward MIMO relaying," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1644–1659, Mar. 2018.

[75] H. Iimori, G. T. F. de Abreu, and G. C. Alexandropoulos, "MIMO beamforming schemes for hybrid SIC FD radios with imperfect hardware and CSI," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4816–4830, Oct. 2019.

[76] S. V. Huffel, *Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms*, G. H. Golub and P. V. Dooren, Eds. Springer, 1991, vol. 70.

[77] L. N. Trefethen and D. B. III, *Numerical linear algebra (1st ed.)*. SIAM, 1997.

[78] S. Hashima and O. Muta, "Fast matrix inversion methods based on Chebyshev and newton iterations for zero forcing precoding in massive MIMO systems," *EURASIP J. Wireless Com. Network.*, vol. 34, pp. 1–12, Feb. 2020.

[79] A. W. Harrow, A. Hassidim, and S. Lloyd, "Quantum algorithm for linear systems of equations," *Phys. Rev. Lett.*, vol. 103, no. 15, pp. 1–4, Oct. 2009.

[80] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5239–5269, Oct. 2019.

[81] L. T. Nguyen, J. Kim, and B. Shim, "Low-rank matrix completion: A contemporary survey," *CoRR*, vol. abs/1907.11705, 2019. [Online]. Available: https://arxiv.org/abs/1907.11705

[82] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[83] L. Vandenberghe and S. P. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, Mar. 1996.

[84] J. F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Mar. 2010.

[85] Q. Yao and J. T. Kwok, "Accelerated inexact soft-impute for fast large-scale matrix completion," in *Proc. IJCAI*, Jul. 2015, pp. 4002–4008.

[86] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," *Fixed–Point Alg. Inv. Prob. Sci. Eng.*, pp. 185–212, May 2011.

[87] N. Antonello, L. Stella, P. Patrinos, and T. van Waterschoot, "Proximal gradient algorithms: Applications in signal processing," ArXiv e-prints, Mar. 2018.

[88] Q. Yao, J. T. Kwok, F. Gao, W. Chen, and T.-Y. Liu, "Efficient inexact proximal gradient algorithm for nonconvex problems," in *Proc. IJCAI*, 2017, pp. 3308–3314.

[89] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Machine Learning Res.*, vol. 11, no. 80, pp. 2287–2322, Aug. 2010.

[90] C. Liu, X. Wang, T. Lu, W. Zhu, J. Sun, and S. C. Hoi, "Discrete social recommendation," in *Proc. AAAI*, 2019, pp. 208–215.

[91] D. Lian, R. Liu, Y. Ge, K. Zheng, X. Xie, and L. Cao, "Discrete content-aware matrix factorization," in *Proc. ACM SIGKDD*, 2017, pp. 325–334.

[92] Z. Huo, J. Liu, and H. Huang, "Optimal discrete matrix completion," in *Proc. AAAI*, Phoenix, USA, 2016, pp. 1–7.

[93] J. Huang, F. Nie, and H. Huang, "Robust discrete matrix completion," in *Proc. AAAI*, Bellevue, USA, 2013, pp. 1–7.

[94] D. M. Nguyen, E. Tsiligianni, and N. Deligiannis, "Learning discrete matrix factorization models," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 720–724, May 2018.

[95] Q. Yao, J. T. Kwok, T. Wang, and T.-Y. Liu, "Large-scale low-rank matrix learning with nonconvex regularizers," *IEEE Trans. Pattern Ana. Machine Intel.*, vol. 41, no. 11, pp. 2628–2643, Nov. 2019.

[96] M. Burger, A. Sawatzky, and G. Steidl, "First order algorithms in variational image processing," *Splitting Met. Commun. Imag. Sci. Eng.*, pp. 345–407, Jan. 2017.

[97] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, 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.

[98] A. Decurninge, L. G. Ordóñez, and M. Guillaud, "Covariance-aided CSI acquisition with non-orthogonal pilots in massive MIMO: A large-system performance analysis," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4489–4512, Jul. 2020.

[99] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing–Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

[100] ——, "Bilinear generalized approximate message passing—Part II: Applications," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5854–5867, 2014.

[101] Y. Kabashima, F. Krzakala, M. Mezard, A. Sakata, and L. Zdeborova, "Phase transitions and sample complexity in Bayes-optimal matrix factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4228–4265, Jul. 2016. [Online]. Available: http://dx.doi.org/10.1109/TIT.2016.2556702

[102] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *Proc. IEEE ICASSP*, Brisbane, Australia, 2015.

[103] K. Ito, T. Takahashi, S. Ibi, and S. Sampei, "Bilinear gaussian belief propagation for large MIMO channel and data estimation," in *Proc. IEEE GLOBECOM*, Taipei, Taiwan 2020.

[104] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A, Math. Gen.*, vol. 36, no. 43, pp. 11 111–11 121, Oct. 2003.

[105] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Singapore, Dec. 2017.

[106] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sept. 2018.

[107] Y. Du, B. Dong, W. Zhu, P. Gao, Z. Chen, X. Wang, and J. Fang, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 682–685, Aug. 2018.

[108] T. Hara and K. Ishibashi, "Grant-free non-orthogonal multiple access with multiple-antenna base station and its efficient receiver design," *IEEE Access*, vol. 7, pp. 175 717–175 726, Nov. 2019.

[109] S. Jiang, X. Yuan, X. Wang, C. Xu, and W. Yu, "Joint user identification, channel estimation, and signal detection for grant-free NOMA," *IEEE Trans. Wireless Commun. Early Access*, 2020.

[110] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan, and L. Hanzo, "Iterative joint channel estimation, user activity tracking, and data detection for FTN-NOMA systems supporting random access," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2963–2977, May 2020.

[111] Y. Zhang, Z. Yuan, Q. Guo, Z. Wang, J. Xi, and Y. Li, "Bayesian receiver design for grant-free NOMA with message passing based structured signal estimation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8643–8656, Aug. 2020.

[112] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925–2951, 2021.

[113] 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.

[114] Z. Chen, F. Sohrabi, and W. Yu, "Multi-cell sparse activity detection for massive random access: Massive MIMO versus cooperative MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4060–4074, Aug. 2019.

[115] L. Liu and W. Yu, "Massive connectivity with massive MIMO–Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

[116] X. Shao, X. Chen, and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420–435, 2020.

[117] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[118] E. D. Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, "Non-stationarities in extra-large-scale massive MIMO," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 74–80, Aug. 2020.

[119] H. Masoumi and M. J. Emadi, "Performance analysis of cell-free massive mimo system with limited fronthaul capacity and hardware impairments," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1038–1053, 2020.

[120] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[121] L. N. Ribeiro, S. Schwarz, and M. Haardt. (2021) Low-complexity zero-forcing precoding for XL-MIMO transmissions. [Online]. Available: https://arxiv.org/abs/2103.00971

[122] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation propagation detector for extra-large scale massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2036–2051, Mar. 2020.

[123] J. C. Marinello, T. Abrão, A. Amiri, E. de Carvalho, and P. Popovski, "Antenna selection for improving energy efficiency in XL-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13 305–13 318, Nov. 2020.

[124] O. S. Nishimura, J. C. Marinello, and T. Abrão, "A grant-based random access protocol in extra-large massive MIMO system," *IEEE Commun. Letters*, vol. 24, no. 11, pp. 2478–2482, Nov. 2020.

[125] X. Yang, F. Cao, M. Matthaiou, and S. Jin, "On the uplink transmission of extra-large scale massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15 229–15 243, Dec. 2020.

[126] Á. O. Martínez, E. D. Carvalho, and J. O. Nielsen, "Towards very large aperture massive MIMO: A measurement based study," in *Proc. IEEE GC Wkshps*, Austin, USA, Dec. 2014.

[127] Y. Han, M. Li, S. Jin, C.-K. Wen, and X. Ma, "Deep learning-based FDD non-stationary massive MIMO downlink channel reconstruction," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 1980–1993, Sept. 2020.

[128] Y. Han, S. Jin, C.-K. Wen, and X. Ma, "Channel estimation for extremely large-scale massive MIMO systems," *IEEE Wireless Commun. Lett*, vol. 9, no. 5, pp. 633–637, May 2020.

[129] J. Zhang, Y. Wei, E. Björnson, Y. Han, and S. Jin, "Performance analysis and power control of cell-free massive MIMO systems with hardware impairments," *IEEE Access*, vol. 6, pp. 55 302–55 314, Sep. 2018.

[130] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1086–1100, Apr. 2021.

[131] C. Rusu and N. González-Prelcic, "Designing incoherent frames through convex techniques for optimized compressed sensing," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2334–2344, May 2016.

[132] C. Rusu, N. González-Prelcic, and R. W. Heath Jr., "Algorithms for the construction of incoherent frames under various design constraints," *Signal Process.*, vol. 152, pp. 363–372, 2018.

[133] R.-A. Stoica, G. T. F. de Abreu, and H. Iimori, "A frame-theoretic scheme for robust millimeter wave channel estimation," in *Proc. IEEE VTC Fall*, Kansas City, USA, Aug. 2018.

[134] T. Strohmer and R. W. H. Jr., "Grassmannian frames with applications to coding and communication," *Applied Comp. Harm. Analysis*, vol. 14, pp. 257–275, 2003.

[135] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured grassmannian constellation for non-coherent SIMO communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1948–1964, Mar. 2020.

[136] M. Thill and B. Hassibi, "Group frames with few distinct inner products and low coherence," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5222–5237, Oct. 2015.

[137] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 50–61, May 2010.

[138] J. A. Tropp, I. S. Dhillon, R. W. Heath, and T. Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 188–209, Jan. 2005.

[139] I. Kammoun, A. M. Cipriano, and J. Belfiore, "Non-coherent codes over the grassmannian," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3657–3667, Oct. 2007.

[140] R. H. Gohary and T. N. Davidson, "Noncoherent MIMO communication: Grassmannian constellations and efficient detection," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1176–1205, Mar. 2009.

[141] I. S. Dhillon, J. R. Heath, T. Strohmer, and J. A. Tropp, "Constructing packings in grassmannian manifolds via alternating projection," *Experimental mathematics*, vol. 17, no. 1, pp. 9–35, 2008.

[142] 3GPP, TS 38.101–1, V15.3.0, "NR; User Equipment (UE) radio transmission and reception," Sep. 2018.

[143] B. K. Jeong, B. Shim, and K. B. Lee, "MAP-based active user and data detection for massive machine-type communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8481–8494, Sep. 2018.

[144] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569–3582, 2019.

[145] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020.

[146] A. Chockalingam and B. S. Rajan, *Large MIMO Systems*. Cambridge University Press, 2014.

[147] T. Takahashi, S. Ibi, and S. Sampei, "Design of adaptively scaled belief in multi-dimensional signal detection for higher-order modulation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1986–2001, Mar. 2019.

[148] Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu, "Covariance based joint activity and data detection for massive random access with massive MIMO," in *Proc. IEEE ICC*, Shanghai, China, 2019.

[149] P. Joshi, D. Colombi, B. Thors, L.-E. Larsson, and C. Törnevik, "Output power levels of 4g user equipment and implications on realistic RF EMF, pages = 4545–4550, number = 5, exposure assessments," *IEEE Access*, 2017.

[150] D. Shen, Z. Pan, K.-K. Wong, and V. O. K. Li, "Effective throughput: A unified benchmark for pilot-aided ofdm/sdma wireless communication systems," in *Proc. IEEE INFOCOM*, San Francisco, USA, 2003.

[151] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, Feb. 2013.

[152] T. Hara, H. Iimori, and K. Ishibashi, "Hyperparameter-free receiver for grant-free NOMA systems with MIMO-OFDM," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 810–814, Apr. 2020.

[153] H. Iimori, T. Takahashi, K. Ishibashi, G. T. F. de Abreu, and W. Yu, "Grant-free access via bilinear inference for cell-free MIMO with low-coherence pilots," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[154] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, 2012.

[155] S. M. Azimi-Abarghouyi, B. Makki, M. Haenggi, M. Nasiri-Kenari, and T. Svensson, "Stochastic geometry modeling and analysis of single- and multi-cluster

wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4981–4996, Oct. 2018.

[156] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks–with a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[157] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62 367–62 414, Apr. 2020.

[158] R.-A. Stoica, H. Iimori, G. T. F. de Abreu, and K. Ishibashi, "Frame theory and fractional programming for sparse recovery-based mmwave channel estimation," *IEEE Access*, vol. 7, pp. 150 757–150 774, Oct. 2019.

[159] A. Dogra, R. K. Jha, and S. Jain, "A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies," *IEEE Access, Early Access*, Oct. 2020.

[160] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 10, pp. 436–453, Apr. 2016.

[161] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[162] G. R. MacCartney, T. S. Rappaport, and S. Rangan, "Rapid fading due to human blockage in pedestrian crowds at 5G millimeter-wave frequencies," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017.

[163] S. Ju, O. Kanhere, Y. Xing, and T. S. Rappaport, "A millimeter-wave channel simulator NYUSIM with spatial consistency and human blockage," in *Proc. IEEE GLOBECOM*, Hawaii, USA, Dec. 2019.

[164] H. Iimori, G. T. F. de Abreu, T. Hara, K. Ishibashi, R.-A. Stoica, D. González G., and O. Gonsa, "Robust symbol detection in large-scale overloaded NOMA systems," *IEEE Open J. Commun. Society*, vol. 2, pp. 512–533, Mar. 2021.

[165] V. Raghavan, L. Akhoondzadeh-Asl, V. Podshivalov, J. Hulten, M. A. Tassoudji, O. H. Koymen, A. Sampath, and J. Li, "Statistical blockage modeling and robustness of beamforming in millimeter-wave systems," *IEEE Trans. Micro. Theory Tech.*, vol. 67, no. 7, pp. 3010–3024, Jul. 2019.

[166] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and J. R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[167] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Sig. Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[168] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmwave OFDM large-scale antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1432–1443, Jul. 2017.

[169] R. Magueta, D. Castanheira, A. Silva, R. Dinis, and A. Gameiro, "Hybrid multi-user equalizer for massive MIMO millimeter-wave dynamic subconnected architecture," *IEEE Access*, vol. 7, pp. 79 017–79 029, Jun. 2019.

[170] A. M. Elbir and K. V. Mishra, "Joint antenna selection and hybrid beamformer design using unquantized and quantized deep learning networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1677–1688, Mar. 2020.

[171] Z. Luo, H. Liu, Y. Li, H. Wang, and L. Zhang, "Robust hybrid transceiver design for AF relaying in millimeter wave systems under imperfect CSI," *IEEE Access*, vol. 6, pp. 29 739–29 746, May 2018.

[172] A. M. Elbir and A. K. Papazafeiropoulos, "Hybrid precoding for multiuser millimeter wave massive MIMO systems: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 552–563, Jan. 2020.

[173] G. R. MacCartney and T. S. Rappaport, "Millimeter-wave base station diversity for 5G coordinated multipoint (CoMP) applications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3395–3410, Jul. 2019.

[174] B. Maham and P. Popovski, "Capacity analysis of coordinated multipoint reception for mmwave uplink with blockages," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 299–16 303, Dec. 2020.

[175] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1952, Feb. 2018.

[176] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmwave beam and blockage prediction using sub-6GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, Sept. 2020.

[177] T. Nishio, H. Okamoto, K. Nakashima, Y. Koda, K. Yamamoto, M. Morikura, Y. Asai, and R. Miyatake, "Proactive received power prediction using machine learning and depth images for mmwave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2413–2427, Nov. 2019.

[178] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, M. Morikura, and K. Nakashima, "Communication-efficient multimodal split learning for mmwave received power prediction," *IEEE Commun. Letters*, vol. 24, no. 6, pp. 1284–1288, Jun. 2020.

[179] S. Mihara, S. Ito, T. Murakami, and H. Shinbo, "Positioning for user equipment of mmwave system using RSSI and stereo camera images," in *Proc. IEEE WCNC*, Nanjing, China, 2021.

[180] M. Gao, B. Ai, Y. Niu, W. Wu, P. Yang, F. Lyu, and X. Shen, "Efficient hybrid beamforming with anti-blockage design for high-speed railway communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9643–9655, Sept. 2020.

[181] D. Kumar, J. Kaleva, and A. Tölli, "Blockage-aware reliable mmwave access via coordinated multi-point connectivity," *IEEE Trans. Wireless Commun., early access*, Feb. 2021.

[182] V. N. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Int. Conf. Neural Info. Process. (NIPS)*, Dec. 1991, pp. 831–838.

[183] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Int. Conf. Neural Info. Process. (NIPS)*, 2013, pp. 315–323.

[184] L. Bottou, *Online Learning and Stochastic Approximations*. Cambridge University Press, 1998.

[185] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss," *J. Machine Learning Research*, vol. 9, pp. 1823–1840, Jun. 2008.

[186] S. Sun, G. R. MacCartney, and T. S. Rappaport, "Millimeter-wave distance-dependent large-scale propagation measurements and path loss models for outdoor and indoor 5G systems," in *Proc. EuCAP*, Davos, Switzerland, Apr. 2016, pp. 1–5.

[187] G. R. MacCartney and T. S. Rappaport, "Rural macrocell path loss models for millimeter wave wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1663–1677, Jul. 2017.

[188] B. Wang, M. Jian, F. Gao, G. Y. Li, and H. Lin, "Beam squint and channel estimation for wideband mmwave massive MIMO-OFDM systems," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5893–5908, Dec. 2019.

[189] Y. Xu and W. Yin, "Block stochastic gradient iteration for convex and nonconvex optimization," *SLAM J. Optim.*, vol. 25, no. 3, pp. 1686–1716, Aug. 2015.

[190] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: theory, robustness to outliers, and savageboost," in *Proc. Int. Conf. Neural Info. Process. (NIPS)*, 2008, pp. 1049–1056.

[191] E. Bottou, "Stochastic gradient learning in neural networks," in *Proc. Neuro-Nimes*, 1991.

[192] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[193] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.

[194] G. Zhou, C. Pan, H. Ren, K. Wang, M. Elkashlan, and M. D. Renzo, "Stochastic learning-based robust beamforming design for RIS-aided millimeter-wave systems in the presence of random blockages," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 1057–1061, Jan. 2021.

[195] O. Taghizadeh, V. Radhakrishnan, A. C. Cirik, R. Mathar, and L. Lampe, "Hardware impairments aware transceiver design for bidirectional full-duplex MIMO OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7450–7464, May 2018.

[196] C. U. Bas, R. Wang, S. Sangodoyin, S. Hur, K. Whang, J. Park, J. Zhang, and A. F. Molisch, "28 GHz microcell measurement campaign for residential environment," in *Proc. IEEE GLOBECOM*, Singapore, Singapore, 2017.

[197] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[198] B. C. Nguyen, X. N. Tran, D. T. Tran, X. N. Pham, and L. T. Dung, "Impact of hardware impairments on the outage probability and ergodic capacity of one-way and two-way full-duplex relaying systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8555–8567, Aug. 2020.

[199] S. Dey, E. Sharma, and R. Budhiraja, "Hardware-impaired rician-faded massive MIMO FD relay: Analysis and optimization," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5209–5227, Aug. 2021.

[200] S. Takabe, T. Wadayama, and Y. Eldar, "Complex trainable ista for linear and nonlinear inverse problems," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 5020–5024.

[201] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," *in Proc. 27th Int. Conf. Mach. Learn.*, pp. 399–406, 2010.

[202] J. Zhang, H. He, C. Wen, S. Jin, and G. Y. Li, "Deep learning based on orthogonal approximate message passing for CP-Free OFDM," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8414–8418.