DESIGN AND OPTIMIZATION OF CODE-DOMAIN NOMA SYSTEMS

A Thesis

Submitted in Partial Fulfilment

of the Requirements for the Degree of

Doctor of Philosophy

by

Vinjamoori Vikas

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039, Assam, India

November 2024

To

My wife and daughter

Certificate

This is to certify that the thesis entitled "**Design and Optimization of Code-domain NOMA Systems** ", submitted by **Vinjamoori Vikas**, a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has met all the requirements as per the institute regulations and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Date:
Place: Guwahati

Dr. A. Rajesh and Dr. Kuntal Deka
Dept. of Electronics and Electrical Engg.,
Indian Institute of Technology Guwahati,
Guwahati - 781 039, Assam, India.

Acknowledgements

First and foremost, I would like to express my deepest and most sincere gratitude to my supervisors Dr. A. Rajesh and Dr. Kuntal Deka for their excellent guidance throughout my study. Their kindness, dedication, hard work and attention to detail have been a source of great inspiration to me. My heartfelt thanks to them for the unlimited support and patience shown to me. I sincerely thank them for the pain they undertook in scrutinizing every work I presented to them and offering critical comments to refine my work.

I am very thankful to my doctoral committee members Prof. Prabin Kumar Bora, Dr. Tony Jacob, and Dr. Smarajit Das for sparing their precious time out of their busy schedule to evaluate my progress and enrich this work with their invaluable suggestions and feedback. I thank Dr. Sanjeev Sharma for his feedback and comments on the work done during our collaboration. I am grateful to Dr. Arghadip Roy for providing me with the opportunity to serve as a JRF in his project during a challenging period in my research.

I would also like to thank the Head of the Department and other faculty members for their kind help in carrying out this work. Thanks are also due to the Science and Engineering Research Board (SERB), Govt. of India for funding my research studies during the later half of my PhD.

I sincerely thank my parents, brother and other family members for their limitless efforts, support and blessings. I express a deep sense of gratitude to my wife and daughter for their tremendous support, patience and love.

(Vinjamoori Vikas)

Abstract

This work focuses on modeling and optimization of transmitter and receiver of code-domain non-orthogonal multiple access (CD-NOMA) systems. CD-NOMA is one of the promising candidates in 6G wireless networks that addresses the challenges namely, ultra giga bandwidth efficiency, ultra low latency, massive connectivity, and high reliability *etc*. NOMA has emerged as a promising technology for transmitting multi-user data over a single resource element (RE) simultaneously. It enables grant-free access, effectively reducing the additional control signaling overhead typically associated with orthogonal multiple access (OMA) schemes. As a result, NOMA plays a crucial role in advancing machine-type communications (MTC), the Internet of Things (IoT), and smart cities, among other applications.

The CD-NOMA can be classified broadly into two categories: sparse CD-NOMA systems and dense CD-NOMA systems. The popular codebook (CB) based sparse and dense CD-NOMA systems are, sparse code multiple access (SCMA) and dense code multiple access (DCMA), respectively. In this thesis, we investigate the design of CD-NOMA systems from three main aspects: codebook design for SCMA, efficient detection schemes for CD-NOMA systems and grant-free CD-NOMA systems. The codebook design and optimization plays an important role in improving the error rate performance of the system.

We present a comprehensive approach for designing near-optimal SCMA codebooks with overloading factors of 150% and 200% across various channel environments. This approach is built upon a novel diversity analysis framework introduced in this thesis. Based on this analysis for the Rayleigh fading channel, we

define a new key performance indicator (KPI) for the average symbol error rate (SER). Since the SER performance of an SCMA codebook is highly influenced by KPIs, carefully selecting these KPIs can enhance overall performance. Our proposed codebook design method leverages constellation rotation and dimensional permutation to optimize the KPIs. Notably, the resulting SCMA codebooks perform equally well in both uplink (UL) and downlink (DL) Rayleigh fading channels, eliminating the need for separate designs for UL and DL. Extensive Monte Carlo simulations are conducted to evaluate the SER of the proposed codebooks across different overloading factors and alphabet sizes. The results demonstrate that our proposed codebooks outperform state-of-the-art SCMA codebooks.

In SCMA, sparse codewords are typically detected using the near-optimal message-passing algorithm (MPA). However, MPA's complexity grows exponentially with the number of overlapped user equipments (UEs) on each RE and the codebook size used at the encoder. MPA's complexity further escalates with an increase in transmit antennas. Additionally, MPA is not practical for DCMA systems due to the absence of sparsity. To overcome these challenges, we propose an iterative linear detector for CD-NOMA systems based on the alternating direction method of multipliers (ADMM). In this approach, the standard Maximum Likelihood (ML) detection problem for CD-NOMA is reformulated as a sharing optimization problem, which is then solved in a distributed manner using ADMM. This reformulation enables resource-wise processing in CD-NOMA systems, making the ADMM algorithm well-suited for both sparse and dense coded NOMA systems. By reformulating the system model, single input multi-output (SIMO) and spatially modulated (SM) CD-NOMA systems become special cases of spatial multiplexing CD-NOMA (SMX-CD-NOMA) for detection purposes. The ADMM algorithm's resource-wise processing framework also provides a solution for detection in various multiple input multiple output (MIMO) CD-NOMA systems. In summary, the ADMM-based detector offers a flexible and effective approach in improving the detection per-

formance of CD-NOMA systems, especially in the context of future 6G wireless networks.

The proposed ADMM detector efficiently handles data detection. However, in grant-free NOMA scenarios, the detector must be capable of joint activity and data detection (JADD). Extending the ADMM-based detection framework to support such scenarios is crucial. This work focuses on designing a detector for CB-based SIMO grant-free NOMA systems. One of the key challenges in grant-free NOMA is accurately detecting both user activity and transmitted data, especially when user activity fluctuates dynamically across the transmission frame. In this study, CB-based grant-free NOMA systems are modeled using a block sparsity signal structure. The JADD problem is formulated as a group LASSO and sparse group LASSO-based block compressive sensing (BCS) problem for DCMA and SCMA systems, respectively. The ADMM approach is then applied to solve these BCS problems efficiently. Moreover, the ADMM algorithm is adapted to address the JADD problem under the most practical grant-free scenario, i.e., a frame-wise dynamic block sparsity model. A robust prior-information-aided ADMM algorithm is proposed, leveraging the temporal correlation of active user support sets between consecutive time slots. To further enhance the detector's performance, a two-step procedure is introduced, where prior information—in terms of both active user support sets and signal data—is incorporated into the ADMM iterations. This approach not only improves detection performance but also maintains low complexity compared to existing benchmark detectors. Extensive numerical experiments and theoretical analysis demonstrate the efficiency of the proposed algorithm, making it a suitable solution for massive machine type communication (mMTC) networks in 6G.

Contents

List of Figures

List of Tables

Important List of Abbreviations

ADMM	Alternating Direction Method of Multipliers
AMP	Approximate Message Passing
BCR	Box Constraint Relaxation
BCS	Block Compressive Sensing
BS	Base Station
CB	Codebook
CS	Compressive Sensing
CD-NOMA	Code-domain NOMA
CSI	Channel State Information
DCMA	Dense Code Multiple Access
DCS	Dynamic Compressive Sensing
DL	Downlink
EPA	Expectation Propagation Algorithm
FG	Factor Graph
GAMP	Generalized AMP
GSD	Generalized Sphere Decoder
IoT	Internet-of-Things
IUI	Inter-user Interference
JADD	Joint Activity and Data Detection
KPIs	Key Performance Indicators
LASSO	Least Absolute Shrinkage and Selection Operator
LDPC	Low Density Parity Check Codes
LDS	Low-density spreading
LMMSE	Linear Minimum Mean Square Error
MAP	Maximum aposteriori Probability
MC	Mother Constellation

MDC	Multi-dimensional Constellation
MED	Minimum Euclidean Distance
MED_{MC}	MED of MC
MED_{sup}	MED of Superimposed Constellation
MIMO	Multiple Input Multiple Output
mMTC	Massive machine-type Communications
MPA	Message Passing Algorithm
MPD	Minimum Product Distance
MPD_{MC}	MPD of MC
MPD_{sup}	MPD of Superimposed Constellation
MUD	Multi User Detection
MUI	Multi User Interference
MUSA	Multi User Shared Access
NGMA	Next generation multiple access
NOMA	Non-orthogonal Multiple Access
OAMP	Orthogonal AMP
PD-NOMA	Power-domain NOMA
RE	Resource Element
SCMA	Sparse Code Multiple Access
SIC	Successive Interference Cancellation
SINR	Signal to Interference Noise Ratio
SNR	Signal to Noise Ratio
SMX	Spatial Multiplexing
SM	Spatial Modulation
UE	User Element
UL	Uplink
URLLC	Ultra-reliable Low Latency Communications
5G	Fifth Generation
6G	Sixth Generation

List of Symbols

$\mathbb{R}^n$	n-dimensional real vector space	$\mathbb{C}^n$	n-dimensional complex vector space
J	Number of UEs	K	Number of REs
ϕ_k	Set of UEs connected to kth RE	ψ_j	Set of REs connected to jth UE
d_f	Cardinality of ϕ_k	d_v	Cardinality of ψ_j
$\mathbf{C}_j$	jth UE codebook	M	Codebook size
X	SCMA block	$\mathbf{S}$	Superimposed codeword
λ	Overloading factor	$\mathbf{F}$	Factor graph matrix
$(\cdot)^T$	Transpose	$(\cdot)^H$	Hermitian transpose
$\lvert(\cdot)\rvert$	Absolute value	$\lVert(\cdot)\rVert_2$	Euclidean norm of
$\prod_{k\in S}$	product of elements $(\cdot)$ that correspond to the set S	$\mathbb{Z}^n$	n-dimensional integer Lattice
$\mathbb{K}$	Number field	$\mathbb{Q}$	Set of rational numbers
$\mathbf{M}$	Lattice generator matrix	Λ	Lattice
$\mathcal{CN}(\mu, \sigma^2)$	Complex Gaussian distribution with mean μ and variance σ^2	$\mathbf{I}_K$	Identity matrix of size $K \times K$
$\mathbf{G}$	Codebook generator matrix	$\mathbf{s}_i$	ith SC
$\prod_{[-\alpha,\alpha]}$	Euclidean projection on to the interval $[-\alpha, \alpha]$	$\langle\cdot,\cdot\rangle$	inner product
N_t	Number of transmit antennas	N_r	Number of receive antennas
ρ	Augmented Lagrangian penalty parameter	$\mathbf{u}_i$	ith dual variable in ADMM
d_k	k-th dimension-wise distance	$d_{k,\min}$	kth dimension-wise minimum distance

diag(**h**)	Diagonal matrix with elements of **h** in its diagonal	**H**	Channel matrix
∇	Gradient operator	∂	Sub-gradient operator

List of Publications

Journal Papers

- **Vinjamoori Vikas**, A. Rajesh, Kuntal Deka and Sanjeev Sharma, "A Comprehensive Technique to Design SCMA Codebooks," in **IEEE Communications Letters**, vol. 26, no. 8, pp. 1735-1739, Aug. 2022, doi: 10.1109/LCOMM.2022.3177542.
- **Vinjamoori Vikas**, Kuntal Deka, Sanjeev Sharma and A. Rajesh, "ADMM-based Detector for Large-scale MIMO Dense Code-domain NOMA Systems," in **IEEE Transactions on Vehicular Technology**, doi: 10.1109/TVT.2024.3424690.
- **Vinjamoori Vikas**, Kuntal Deka, Sanjeev Sharma and A. Rajesh, "Robust Prior Information-Aided ADMM for Multi-User Detection in Codebook-Based Grant-Free NOMA Under Dynamic Scenarios," in **arXiv, Oct. 2024**.

Conference Paper :

- **Vinjamoori Vikas**, Kuntal Deka, Sanjeev Sharma and A. Rajesh, "Block Compressive Sensing Based Multi-User Detection for Uplink Grant-Free Codebook-Based NOMA," accepted in **National Conference on Communications (NCC)**.

1

Introduction

Contents

1.1 Background

Over the past few decades, wireless communication has evolved significantly, advancing from 1G to 5G cellular standards. The multiple access plays a crucial role in each cellular standard and that enables multi-users to share the common communication channel. Cellular standards have witnessed a revolution in terms of their MA schemes. The MA schemes evolved form 1G to 5G as follows: frequency division multiple access (FDMA) in 1G, time division multiple access (TDMA) in 2G, code division multiple access (CDMA) in 3G, and orthogonal frequency division multiple access (OFDMA) in 4G as well as 5G New Radio (NR). These MA schemes fall under orthogonal multiple access (OMA), where resources, such as frequency, time, or code, or a combination thereof, are allocated to users in an orthogonal manner [1]. In FDMA, the available band width is divided into multiple frequency bands and each user is assigned a unique frequency band. In TDMA, each user is assigned a unique time slot, ensuring no overlap between the time slots, and different users allowed to share the same frequency band. In CDMA, a new dimension in the code domain was used as a resource for transmission of data. Each user is assigned a specific spread sequence (pseudo-noise (PN)-sequence) and ensuring the PN-sequences of any two users are orthogonal to each other. In OFDMA, resource refers to the time and frequency (sub-carrier) element used for the data transmission. The orthogonality between two subcarriers is the key property of OFDMA and which ensures the efficient extraction of the different user data modulated over different subcarriers. The orthogonality in all of the above approaches strictly limits the number of users able to access the network simultaneously to the number of available resources. However, OMA effectively eliminates multi-user interference (MUI) within the channel by assigning separate, non-overlapping resource units to each user. Its strength lies in the orthogonal allocation of wireless resources, ensuring robust user separation with low-cost equalizers. On the other hand, this orthogonality can be disrupted by time or frequency dispersion in fading channels, leading to the MUI in the channel. This necessitates the use of complex multi-user equalizers at the receiver to effectively separate overlapping signals. In addition, OMA schemes typically rely on grant-based access,

which introduces significant overhead. While OMA efficiently manages access for a limited number of users in traditional human-type communications (HTC) through channel scheduling, connection requests, and grants, it becomes a bottleneck in next-generation wireless networks. This is due to its limited spectral efficiency (SE), high latency, substantial overhead, and constrained capacity to support a large number of devices or users [2].

The next generation wireless networks demands massive connectivity, low latency and high throughput to support massive machine-type communications (mMTC), ultra-reliable low-latency communications (URLLC), and various other applications [2]. Critical performance targets have been established for 6G networks [3]: a) A peak data rate of at least one terabit per second; b) latency for high-mobility users should range from 0.01 to 0.1 milliseconds; c) massive connectivity must be ten times greater than that of 5G systems; d) spectral efficiency (SE) and energy efficiency (EE) must be 5-10 times and 10-100 times higher than those of 5G, respectively; and e) reliability must exceed 99.999999%. In order to circumvent the limitations of OMA and achieve aforementioned performance targets, a sophisticated multiple access scheme termed next generation multiple access (NGMA) needs to be designed.

1.1.1 Main considerations for NGMA design

The critical performance targets have been outlined earlier in the text. However, the key considerations for designing NGMA [3], which form the primary motivation and focus of this thesis, are as follows:

(i) **Massive Access:** With the proliferation of cellular connectivity into the Internet of Things (IoT) and mMTC networks, the sheer number of connected devices—commonly referred to as massive access—is expected to grow significantly in 6G systems. To manage this demand within the constraints of existing bandwidth, innovative approaches to multiple access design are essential.

(ii) **Ultra Low Latency:** In addition to massive access, ultra low latency is one of the main important considerations for NGMA. IoT or mMTC network users generate the data in the form of short packets and demand low-latency communication [1]. Conventional grant-based access methods, typically used in HTC, are unsuitable for such scenarios [4].

Therefore, new techniques to efficiently support short-packet transmissions must be developed and integrated into multiple access schemes as part of NGMA.

1.1.2 Possible Candidates of NGMA

This section discuss the new approaches and possible candidates considered for NGMA as given below:

(i) **Power-Domain Non-orthogonal Multiple Access (PD-NOMA):** NOMA is an innovative approach to address the challenges and achieve the performance objectives of 6G wireless networks. The core principle of NOMA lies in the non-orthogonal allocation of resources among users, enabling the accommodation of a larger number of users with a limited pool of resources [5]. In PD-NOMA, multiple users share the same resource (e.g., time, frequency, or spatial dimensions) with distinct transmit power levels based on their channel conditions [5,6]. Utilizing superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, PD-NOMA enables simultaneous multi-user data transmission over shared resources. PD-NOMA can also be combined with OFDMA by allocating multiple users to a single OFDMA subcarrier and differentiating them using varying power levels [5]. However, the performance of the SIC decoder is highly susceptible to error propagation, which becomes more pronounced as the number of users increases [7]. This issue often results in significant re-transmission overhead, leading to higher transmission latency [8]. Additionally, PD-NOMA requires dynamic power allocation and user pairing, and as the number of users grows, these processes demand complex optimization techniques [9]. These challenges, including performance degradation and increased system overhead, make PD-NOMA less effective in scenarios involving a large number of users.

(ii) **Space Division Multiple Access (SDMA):** SDMA enables multiple users to transmit data simultaneously over shared resources (e.g., time, frequency, or code) by differentiating them in the spatial domain [10]. It relies on multi-antenna technologies with transmit and receive beamforming. With the advancement of multiple-input multiple-output (MIMO) techniques in 5G, SDMA has garnered significant attention for 6G applications.

At the transmitter, SDMA employs a linear precoder (LP) to mitigate interference and enhance detection performance. However, this approach requires accurate channel state information (CSI) at the transmitter, making its performance highly sensitive to CSI accuracy. Additionally, the presence of multiple RF chains at the transmitter increases the cost, particularly in massive MIMO systems. Moreover, SDMA is best suited for underloaded and critically loaded scenarios, as the number of users it can support is constrained by the available spatial degrees of freedom (DoF).

(iii) **Rate Splitting Multiple Access (RSMA):** RSMA is an emerging and promising non-orthogonal multi-antenna, multi-user transmission scheme that effectively manages interference [11]. It combines linear precoding at the transmitter, similar to SDMA, with SIC at the receiver, as used in PD-NOMA, to enhance system performance. The core concept of RSMA lies in splitting each user's message into submessages at the transmitter, which are further categorized into private and common submessages. Private submessages contain information unique to individual users, while the common submessage carries information shared among all users.

(iv) **Code-Domain NOMA (CD-NOMA):** CDMA, one of the earliest code-domain multiple access techniques, employs orthogonal spread sequences to spread user data at the transmitter end [12]. CD-NOMA can be seen as an extension of CDMA that incorporates non-orthogonality. This non-orthogonal feature allows CD-NOMA to support the simultaneous transmission of multiple users' data over shared resources (e.g., time, frequency, or subcarriers (scs)) by assigning unique code sequences or codewords to each user. The fundamental principle is to maximize resource sharing by allocating low-correlation code sequences or codebooks to users. At the receiver, multi-user detector is employed to separate and decode individual users' data. Depending on the code structure, CD-NOMA has two extensively studied variants, spread sequence (SS)-based NOMA and CB-based NOMA as depicted in Fig. 1.1.

- **SS-based NOMA:** SS-based NOMA employs separate mapping and spreading operations during the encoding process, which is similar to CDMA. Two prominent

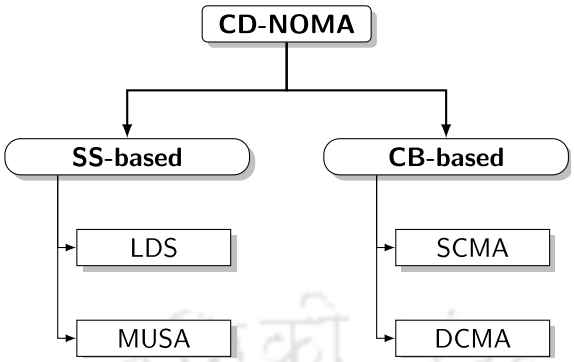


Fig. 1.1: Classification of CD-NOMA

SS-based NOMA schemes are low-density spreading (LDS) [13, 14] and multi-user shared access (MUSA) [15]. LDS utilizes sparse spreading sequences, while MUSA employs dense spreading sequences. LDS/CDMA [13, 14] and LDS/OFDM [16] are two prominent variants of LDS. MUSA utilizes the short length, low cross-correlation complex sequences for spreading. LDS repeats the QAM symbols over pre-allocated resources, while MUSA repeats the QAM symbols over all resources. The sparsity in LDS enables the near-optimal MPA for MUD at the receiver. The inherent power control mechanism presents in MUSA allows SIC for MUD at the receiver. These spreading sequences are characterized by their simplicity and ease of design. However, they are limited by their relatively low coding gain.

- **CB-based NOMA:** CB-based NOMA integrates the mapping and spreading operations found in SS-based NOMA into a single step [17]. In this approach, data is directly mapped to codewords during encoding, with these codewords being selected from pre-designed codebooks. Unlike simple or complex sequences, these codebooks are constructed using multi-dimensional constellations (MDCs), providing enhanced shaping and coding gains [18, 19]. As a result, CB-based NOMA systems achieve significantly higher performance compared to SS-based NOMA systems. Two types of codebook structures are widely explored in the literature. Sparse CB-based NOMA, commonly known as sparse code multiple access (SCMA) [20], and dense CB-based NOMA, referred to as dense code multiple access (DCMA) [21]. At the receiver, advanced multi-user detection techniques are employed to separate the user data multiplexed over shared resources. The sparsity of SCMA codebooks

can be represented using factor graphs (FGs), similar to the representation used for low-density parity-check (LDPC) codes. Consequently, SCMA employs a near-optimal yet computationally intensive MPA for detection [17]. SCMA demonstrates superior error rate performance compared to SS-based NOMA systems [17]. However, the sparse nature of SCMA codewords limits the achievable diversity order. There remains potential for performance improvement by leveraging dense codewords to exploit full diversity. DCMA, another key CB-based NOMA technique, utilizes densely structured codebooks [21], thereby achieving full diversity gains. At the receiver, DCMA employs a generalized sphere decoder (GSD) for multi-user detection [21]. However, the tree search process inherent to GSD is computationally expensive, making it impractical for large-scale DCMA systems.

1.2 Motivation and Focus of the Thesis

Over the past decade, NOMA has been extensively studied, leading to various implementations. NOMA allows multiple user data streams to be overloaded onto a single resource element (RE) in a non-orthogonal manner. This capability enables NOMA to support a large volume of user data using a limited number of REs. The non-orthogonal nature of NOMA introduces an MUI in the channel, NOMA necessitates advanced signal processing techniques at the receiver to mitigate MUI. NOMA offers three critical features that align with the primary considerations for NGMA design, as outlined below:

(i) **Enhanced Spectral Efficiency:** The non-orthogonal allocation of resources in NOMA enables multiple users to share the same time-frequency resources. This overloading nature of NOMA results in significant improvement in the spectral efficiency. In this way, large number of user data can be transmitted over limited number of resources. Therefore, the enhanced spectral efficiency of NOMA supports massive connectivity in 6G wireless networks.

(ii) **Grant-Free Access:** Beyond massive connectivity, another key advantage of NOMA is the ultra low latency through grant-free access. In IoT or mMTC networks, traffic is

often sporadic, with data transmitted in the form of short packets. Traditional grant-based access methods in existing wireless networks (e.g., LTE/LTE-A), which rely on random access channel (RACH) techniques, are unsuitable for future networks due to excessive signaling overhead and high latency. These methods are also inefficient for handling sporadic traffic scenarios. Grant-free NOMA, by contrast, offers a viable solution for such use cases, enabling ultra-low latency communication in mMTC networks.

Two prominent variants of NOMA that have been extensively studied in the literature, introduced in Section 1.1.2 are power-domain NOMA (PD-NOMA) [5, 6] and code-domain NOMA (CD-NOMA) [13, 15, 22].

1.2.1 Focus of the Thesis

This thesis concentrates on the design, analysis, and optimization of CD-NOMA systems, with a specific focus on CB-based NOMA systems, aiming to tackle the following challenges:

1.2.2 Problem 1

Motivation: The construction of codebooks for CD-NOMA involves MDCs [22]. Carefully designed MDCs can simultaneously enhance both coding and shaping gains [18]. Over the past decade, substantial research has been conducted on codebook design for small-scale SCMA and DCMA systems [23–29], among others. However, for a given overloading factor, the design of optimal codebooks for CD-NOMA systems remains an open problem, and extending this design to large-scale CD-NOMA systems with high overloading factors presents an even greater challenge.

Problem Statement: Designing codebooks to enhance performance in both uplink (UL) and downlink (DL) SCMA systems.

1.2.3 Problem 2

Motivation: The overloading characteristic of NOMA introduces significant MUI within the channel, with DCMA systems being more affected than SCMA systems. Effective mitigation of MUI requires advanced receiver processing. Although the message passing algorithm (MPA) is commonly used and effective for SCMA detection [30], it suffers from exponential

computational complexity. Moreover, to enhance the throughput and spatial diversity of CD-NOMA systems, it is essential to integrate MIMO with CD-NOMA. However, MPA becomes impractical, particularly for MIMO-aided spatial multiplexed CD-NOMA systems. Furthermore, MPA is not suitable for DCMA systems [21]. Therefore, there is a need for the development of a new detection algorithm tailored to MIMO-CD-NOMA systems. Designing a detector that strikes a balance between computational complexity and performance presents a significant challenge. Moreover, an efficient detection algorithm that can be adapted for both DCMA and SCMA systems is highly sought after.

Problem Statement: Develop an unified detection algorithm for both DCMA and SCMA systems.

1.2.4 Problem 3

Motivation: Grant-free NOMA allows users to transmit data on-demand without requiring connection grants or channel scheduling from the base station (BS), enabling devices to access the network randomly [4]. In mMTC or IoT networks, where traffic is sporadic, only about 10% of devices are typically active even during peak usage periods [31]. As a result, the network primarily consists of sparsely active users. Therefore, the receiver in grant-free NOMA must first identify the active users before proceeding with data detection. A major challenge in this context is achieving joint activity and data detection (JADD). Developing a low-complexity MUD algorithm that ensures reliable performance for JADD remains a critical area of ongoing research.

Problem Statement: Design an efficient multi-user detector for JADD in grant-free NOMA systems.

1.3 Contributions of the Thesis

Following are the contributions of the thesis:

1.3.1 SCMA codebook design

Following are the contributions in the proposed SCMA codebook design:

(i) Various SCMA codebooks with sizes $M = 4$ and $M = 8$ are designed for overloading factors of 150% and 200%, targeting both UL and DL Rayleigh fading channel scenarios. These codebooks demonstrate significant performance improvements compared to benchmark codebooks.

(ii) An enhanced systematic algorithm is proposed to optimize the codebooks, focusing on newly improved key performance indicators (KPIs) for codebook design.

1.3.2 Multi-user detection for MIMO-CD-NOMA systems

(i) A novel and efficient detection algorithm is proposed within the framework of the alternating direction method of multipliers (ADMM) optimization, utilizing resource-wise processing through the ADMM approach.

(ii) The computationally intensive maximum likelihood (ML) problem in MIMO-CD-NOMA is reformulated into a sharing optimization problem.

(iii) A numerical experimental results and comprehensive comparison of the proposed ADMM detector with various existing detectors are provided.

1.3.3 JADD for grant-free NOMA systems

(i) Block-sparsity models are proposed for CB-based grant-free NOMA systems, incorporating practical use cases commonly found in grant-free scenarios.

(ii) The JADD problems in grant-free DCMA and SCMA systems are formulated as group least absolute shrinkage and selection operator (LASSO) and sparse group LASSO problems, respectively.

(iii) A robust ADMM algorithm incorporating prior information is introduced to address the JADD problem in grant-free NOMA systems under practical conditions.

1.4 Organization of thesis

The rest of the thesis is organized into four chapters. A brief outline of each of these chapters is as follows:

Chapter 2

This chapter focuses on the design of codebooks for SCMA systems. It explores several aspects of codebook design, including diversity analysis, KPIs, and the sequence of steps involved in codebook optimization. A new algorithm is proposed to design comprehensive codebooks tailored for various channel scenarios. The proposed codebooks are thoroughly evaluated and compared with numerous existing codebooks in the literature.

Chapter 3

This chapter proposes a novel multi-user detector based on the ADMM framework for large-scale MIMO CD-NOMA systems. The computationally intensive and optimum ML detection problem is reformulated as a shared optimization problem. The ADMM iterations via resource-wise processing, making it suitable and low-complexity for various MIMO CD-NOMA systems.

Chapter 4

This chapter extends the concept of CB-based NOMA to grant-free scenarios, introducing it as CB-based grant-free NOMA systems. The ADMM algorithm proposed in Chapter 3 is further developed to address the JADD problem in grant-free NOMA systems. A prior information-aided ADMM detection algorithm is designed to enhance JADD performance in dynamic scenarios.

Chapter 5

This chapter provides a summary of the conclusions drawn from the thesis and outlines potential future directions and scope for extending the research.

2

A Comprehensive Technique to Design SCMA Codebooks

Contents

2.1 Introduction

As discussed in Chapter 1, NOMA enables multiple users to transmit data simultaneously over the same resource element (RE). This allows a large number of user equipments (UEs) to be served despite limited REs and constrained bandwidth, a concept referred to as overloading.

This thesis focuses on CD-NOMA, a key NOMA technique. In CD-NOMA, significant data overloading is achieved using code-domain multiplexing. There are two well-known types of CD-NOMA: sequence-based (sparse or dense) and CB-based (sparse or dense). In this chapter, we discuss sparse CB-based NOMA, commonly called SCMA, with a focus on codebook design. SCMA uses sparse codebooks at the encoder, which play a vital role in determining the error rate performance of the system.

The sparse structure of SCMA codebooks is represented by a factor graph (FG) [22], shown in Fig. 2.1. In SCMA, the FG has function nodes (FNs) and variable nodes (VNs) as neighboring nodes. Here, FNs represent REs, and VNs represent UEs. The FG shows the relationships between users and resources. It serves as a foundation for codebook design, multiple access, and MUD [22]. A detailed discussion of FG representation for SCMA systems is provided in Section 2.2.1.

The CD-NOMA system codebook is primarily based on MDCs [22]. This is a key feature that distinguishes CB-based NOMA systems from other NOMA systems. MDCs offer better CG and SG compared to traditional quadrature amplitude modulation (QAM) constellations [18]. As a result, codebook-based NOMA systems achieve better signal-to-noise ratio (SNR) efficiency than SS-based NOMA systems [22].

As noted in Chapter 1, DCMA uses dense codebooks, while SCMA uses sparse codebooks. This chapter focuses on SCMA systems with a special emphasis on designing near-optimal codebooks. The performance of an SCMA codebook depends on the CG and SG of the underlying MDC. Therefore, the design of the MDC has a strong impact on the performance of the SCMA system. CG and SG are characterized by the minimum Euclidean distance (MED) and minimum product distance (MPD). These are widely recognized as KPIs

for MDC design [19, 32] and, subsequently, SCMA codebook design [22, 25], among others. These KPIs play a crucial role in minimizing the pairwise error probability (PEP) of an SCMA system [25, 33].

Thus, MDC design takes these KPIs into account under different channel conditions [34]. Existing near-optimal SCMA codebook designs in the literature mainly focus on optimizing these KPIs [22, 25, 26, 29, 33], among others. The steps for optimizing codebooks to enhance SCMA system performance are described in Section 2.4.

The detection of SCMA codewords is done using the MPA. MPA is a sub-optimal decoding technique commonly used in LDPC decoding. A detailed discussion on SCMA detection is provided in Section 2.2.3.

A comprehensive literature review on SCMA codebook design and optimization is presented below.

2.1.1 Related literature

In this section, we review the state-of-the-art advancements in SCMA codebook design. A systematic sequence of steps for designing SCMA codebooks based on the MDC was initially introduced in [22]. It serves as foundation for many subsequent designs in the literature. This methodology is further detailed in [35]. In this approach, the MDC serves as the base or "mother constellation" (MC) for generating user-specific codebooks. Essentially, multiple user codebooks are derived from a single MC. The process begins with designing an MC that has a favorable Euclidean distance profile. Subsequently, individual user codebooks are obtained from the MDC through user-specific constellation operations, such as rotations and permutations, to enhance the MPD. While this approach primarily focuses on codebook design for UL SCMA systems, it exhibits performance limitations in DL SCMA systems. In response, significant efforts have been made to address codebook design for both UL and DL SCMA systems. Examples of such works include UL-focused designs in [36–39] and DL-specific designs in [23, 26, 40–46].

In [36], a novel approach to SCMA codebook design is presented from a capacity perspective, aiming to enhance sum-rate performance. However, KPIs such as MED and MPD, which

are crucial for PEP, are not considered as design metrics in this work, resulting in relatively inferior error rate performance for the proposed codebooks. In [37], a dimension-distance-based codebook design method is introduced, where dimension-wise Euclidean distance and product distance are utilized as KPIs. This work employs random constellation rotations to improve the product distance between constellation points in the MDC. However, it leaves room for further improvement by optimizing the selection of rotation angles. The authors of [38] propose a new method for designing multi-user codebooks for UL SCMA systems, focusing on the cut-off rate from an equivalent MIMO system perspective. While this approach achieves significant performance gains compared to existing codebooks, it overlooks the importance of MED and MPD in MDC, suggesting potential for further enhancement. In [39], UL SCMA codebooks are designed using spherical codes as outlined in [47]. However, the performance of these codebooks is limited to scenarios with up to 150% overloading.

The multi-dimensional (MD)-SCMA codebook proposed in [23] was developed for DL Rayleigh fading channels, using the MED of the mother constellation (MED_{MC}) and the MPD of the mother constellation (MPD_{MC}) as KPIs. In [40], constellation rotation was introduced to enhance the dimension-wise Euclidean distance of the superimposed constellation ($d_{k,\min}$). However, these codebooks exhibit suboptimal performance, highlighting the need for further optimization. In [42], a novel MC was designed based on energy diversity and normalized MED, followed by the construction of multi-user codebooks. Simulation comparisons demonstrate significant performance improvements over existing MCs. However, this work does not account for MPD during codebook optimization, leading to performance degradation, particularly in Rayleigh fading channels. The work in [44] employs star-QAM as the MC for SCMA codebooks, leveraging its large MED to achieve efficient performance in a AWGN channels. Inspired by [13], a high-dimensional SCMA codebook was proposed to enhance SNR efficiency. Its straightforward construction methodology, based on a Latin rectangular matrix, provides valuable insights into designing high-dimensional codebooks for practical applications.

In [26], SCMA codebook design based on uniquely decomposable constellation groups (UDCG) was introduced, targeting uniquely decodable constellations (UDC). The UDC prop-

erty minimizes the complexity of MUD by facilitating the separation of individual users' data from the superimposed constellation. Additionally, power-imbalance codebooks integrating power-domain and code-domain NOMA concepts were proposed for DL scenarios in [46], achieving significant SNR gains compared to existing DL SCMA codebooks. All the aforementioned approaches outline systematic steps for SCMA codebook design, typically focusing on separate designs for UL and DL Rayleigh fading channels.

Codebooks for both UL and DL SCMA systems have been proposed in [24, 25]. The golden angle modulation (GAM)-based SCMA codebook introduced in [24] targets UL and DL Rayleigh fading channels, with a specific emphasis on minimizing the peak-to-average power ratio and maximizing the normalized MED. Similarly, a near-optimal SCMA codebook proposed in [25] achieves 150% overloading for both UL and DL scenarios, utilizing the MED and minimum product distance (MPD) of the superimposed constellation (MED_{sup} and MPD_{sup}) as KPIs. These codebooks demonstrate superior error performance under various channel conditions. However, the method becomes computationally demanding for overloading factors of 200% or higher. The extensive literature survey reveals that different codebook designs prioritize different KPIs, leading to inconsistency in performance metrics across studies. This inconsistency motivates the development of new multi-user codebooks that unify all the critical KPIs identified in the literature. Despite the progress achieved so far, there is still room for improvement in SCMA system performance by carefully selecting appropriate KPIs and devising a systematic sequence of steps for optimized codebook design.

The focus and contributions of this work are outlined as follows:

2.1.2 Focus and contributions

We consider the codebook design for AWGN , UL Rayleigh fading and DL Rayleigh fading channels for $\lambda = 150\%, 200\%$ overloading scenarios. The contributions are

- This work proposes a comprehensive treatment of probability of error of SCMA system in various channel scenarios. Its results in improved diversity analysis. The improved diversity analysis leads to new KPIs for Rayleigh fading channels related to the MPD of each diversity order up to a specific order. The proposed SCMA codebooks enjoys

additional SNR gains for a series of SER values.

- An improved systematic sequence of steps are proposed for optimization of codebook. Further, various codebooks are designed based on the proposed steps for $\lambda = 150\%, 200\%$ and $M = 4, 8$ over different channels.

Notations: Lowercase, bold lowercase, and bold uppercase letters represent scalars, vectors, and matrices, respectively.

2.2 Preliminaries on SCMA

This section introduces the preliminaries of the SCMA system. It includes FG representation, SCMA codebook mapping, multiple access procedure and SCMA MUD in the following.

2.2.1 FG representation of SCMA

The SCMA system consists of J UEs and K REs. It can be efficiently represented by a factor graph (FG), as shown in Fig. 2.1. The FG contains J VNs representing the UEs and K FNs representing the REs. When $J > K$, the system is overloaded. The overloading factor is defined as $\lambda = \frac{J}{K}$.

The FG can also be represented by a $K \times J$ binary matrix $\mathbf{F} = [f_{k,j}]$. In this matrix, $f_{k,j} = 1$ if the jth UE is connected to the kth RE, and $f_{k,j} = 0$ otherwise. The rows of $\mathbf{F}$ correspond to the REs, while the columns represent the UEs. The sparse structure of the SCMA system results into a sparse $\mathbf{F}$. In each row of $\mathbf{F}$, the entries with 1 indicate the users that overlap on a particular resource. Similarly, in each column of $\mathbf{F}$, the entries with 1 represent the active REs for a specific user. The set of indices of the active UEs connected to kth (where $k = 1, \ldots, K$) RE denoted as ϕ_k and active REs connected to jth (where $j = 1, \ldots, J$) UE denoted as φ_j on each row and each column, respectively, defined as follows:

$$\phi_k = \{j : 1 \leq j \leq J \mid f_{k,j} = 1\} \tag{2.1}$$

$$\varphi_j = \{k : 1 \leq k \leq K \mid f_{k,j} = 1\} \tag{2.2}$$

For a regular factor graph matrix, $|\phi_1| = \ldots = |\phi_K|$ and $|\varphi_1| = \ldots = |\varphi_J|$, and let $d_f = |\phi_k|$ and $d_v = |\varphi_j|$. The following example illustrates the FG representation for SCMA system.

Example: Consider an 150% overloading SCMA system with $J = 6$ and $K = 4$. The representation of the FG and matrix F for this system are given as,

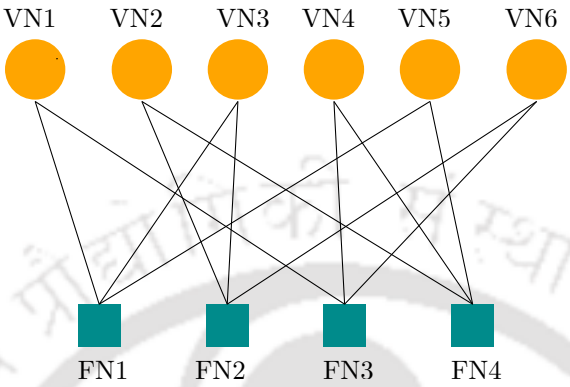


Fig. 2.1: FG representation of SCMA system

$$\boldsymbol{F} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

where d_f=3 and d_v=2. The first column of **F** corresponds to UE-1, which has access to RE-1 and RE-3, while the first row of **F** represents RE-1, which is shared by UE-1, UE-3, and UE-5. This pattern applies to all columns and rows of **F**. Additionally, the jth column of **F** is used to construct the mapping matrix $\mathbf{V}_j$, given by $\mathbf{V}_j = \tilde{\mathbf{V}}_j \Phi_j$. The user-specific rotation matrix is defined as

$$\Phi_j = \text{diag}(e^{i\phi_{j,1}}, \ldots, e^{i\phi_{j,d_v}}). \tag{2.3}$$

For example, the matrix $\tilde{\mathbf{V}}_1$ and $\tilde{\mathbf{V}}_2$ for the UE-1 and UE-2, respectively are defined as

$$\tilde{\mathbf{V}}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \tilde{\mathbf{V}}_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Similarly, for $j = 3, 4, 5, 6$, $\tilde{\mathbf{V}}_j$ can be obtained from the corresponding columns of F. The utilization of $\mathbf{V}_j$ in the construction of the jth UE codebook is elaborated in Section 2.4.1. The

optimization of the rotation angles introduced in Φ_j is presented in Section 2.4.2.

2.2.2 SCMA encoding and multiple access procedure

In this section, we detailed the SCMA encoding and multiple access procedure.

2.2.2.1 SCMA encoding

As discussed in Chapter 1, unlike SS-based NOMA, SCMA integrates the modulation and spreading operations into a single process, referred to as SCMA encoding [17]. Figure 2.2 illustrates the distinction in the encoding processes between SS-based NOMA (LDS-CDMA/LDS-OFDM) and SCMA systems. In SS-based NOMA, spreading is performed after modulation, whereas in SCMA, the encoder merges these two steps into a unified SCMA encoding process. The SCMA encoding method is further explained using an example with a codebook structure for an SCMA system of size $J = 6$ and $K = 4$, as shown in Fig. 2.3.

In SS-based NOMA, identical QAM symbols are repeated across pre-allocated REs. In [22], SCMA codebooks were designed using an MDC structure [18] with rotated QAM constellations. The MDC structure ensures that distinct symbols are assigned to the pre-allocated REs. Subsequent studies [38, 41], among others, have leveraged enhanced MDC structures [44, 48] to improve SCMA codebook design and enhance SCMA performance. In this work, we have designed a novel MDC based SCMA codebook to achieve improved error rate performance.

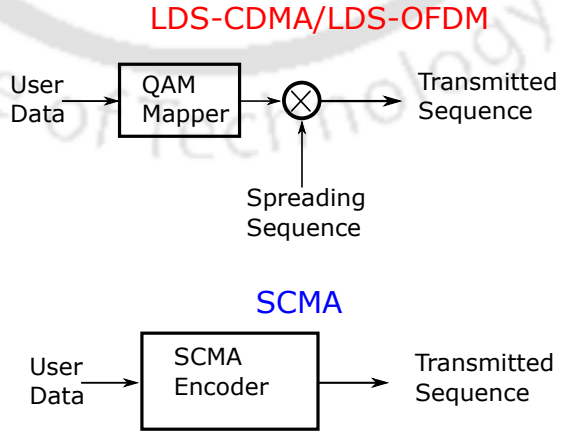


Fig. 2.2: SS-based NOMA and SCMA encoding procedure.

An SCMA encoder maps input data in the form of a binary string to an SCMA codeword, i.e., from $\mathbb{B}^{\log_2 M}$ to $\mathcal{X}_j$, where $\mathcal{X}_j$ represents the pre-designed codebook of the jth user with a

size of M. This can be expressed as a mapping rule

$$f_j : \mathbb{B}^{\log_2 M} \rightarrow \mathcal{X}_j$$
$$f_j(\mathbf{b}_j) = \mathbf{x}_j \in \mathcal{X}_j, \tag{2.4}$$

where $\mathbf{b}_j \in \mathbb{B}^{\log_2 M}$, $\mathcal{X}_j \subset \mathbb{C}^K$, $\mathbf{x}_j$ represents the codeword of $\mathcal{X}_j$. The index set φ_j of (2.2) determines the locations of the non-zero entries in the codebook for user j. Then subsequently, the codewords J-UEs form a SCMA block as given below

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J]. \tag{2.5}$$

This is illustrated in Fig. 2.3 by considering an SCMA system with $J = 6$, $K = 4$, and $M = 4$ whose FG is shown in Fig. 2.1 . Here, $J = 6$ represents six UEs, each associated with a pre-designed codebook, where the codebook for the jth UE is denoted as $\mathbf{C}_j$. For instance, in the case of UE-1, the binary input string "01" is converted to its decimal equivalent '1' and used to select the codeword indexed as 1 in $\mathbf{C}_1$, and so on. Here, $K = 4$ specifies an SCMA system with sparse 4-dimensional complex codewords with a cardinality of $M = 4$ codewords.

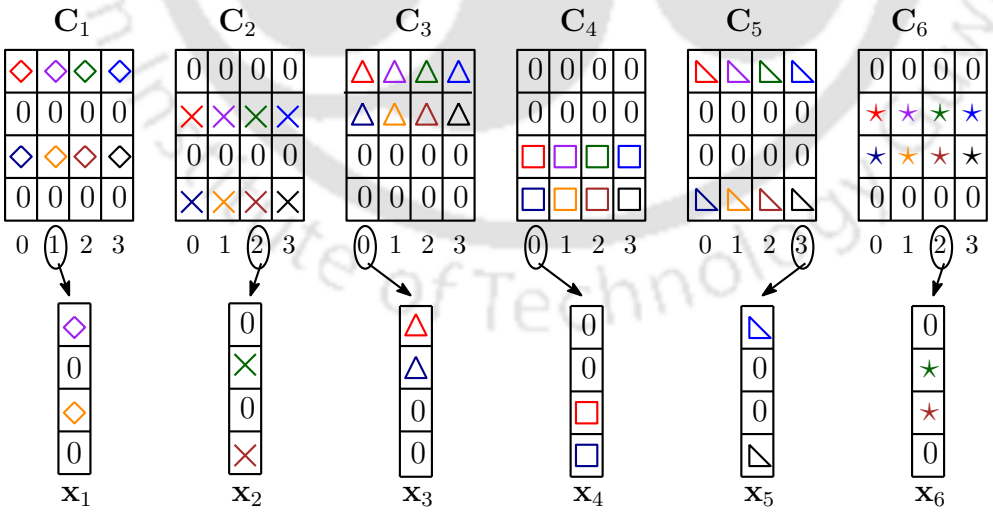


Fig. 2.3: Illustration of SCMA codebooks and bit to codeword mapping

2.2.2.2 Multiple Access Procedure

This section elaborates on the multiple access procedure in SCMA under various channel conditions. As discussed in Section 2.2.1, for an SCMA system with J UEs and K REs,

SCMA encoding is described in (2.4). Following SCMA encoding, the resulting codewords of J UEs form a single SCMA block as given in (2.5). The codewords in SCMA block are then multiplexed to produce a superimposed codeword. As stated earlier, the superimposition of codewords differs between UL and DL SCMA systems. In UL, the codewords in each SCMA block are superimposed over the channel, whereas in DL, the codewords are first superimposed and then transmitted over the channel.

The received signal vector $\mathbf{y}$ across various channel conditions is given as

$$\mathbf{y} = \sum_{j=1}^{J} \mathbf{x}_j + \mathbf{n}, \qquad \text{(AWGN channel)} \tag{2.6}$$

$$\mathbf{y} = \sum_{j=1}^{J} \text{diag}(\mathbf{h}_j)\mathbf{x}_j + \mathbf{n}, \qquad \text{(UL channel)} \tag{2.7}$$

$$\mathbf{y}_j = \text{diag}(\mathbf{h}_j) \sum_{j=1}^{J} \mathbf{x}_j + \mathbf{n}, \qquad \text{(DL channel for the user-}j\text{)} \tag{2.8}$$

where

- $\mathbf{x}_j = [\mathbf{x}_j[1], \mathbf{x}_j[2], \ldots, \mathbf{x}_j[K]]^T$, where $\mathbf{x}_j[k]$ is the kth element of $\mathbf{x}_j$.
- $\mathbf{h}_j \sim \mathcal{CN}(0, \mathbf{I}_K)$ denotes the Rayleigh channel fading vector between the BS and user-j is defined as $\mathbf{h}_j = \left[\mathbf{h}_j[1], \ldots, \mathbf{h}_j[k], \ldots, \mathbf{h}_j[K]\right]^T$.
- Diagonalisation of $\mathbf{h}_j$ results in $\text{diag}(\mathbf{h}_j)$, which is a diagonal matrix with $\mathbf{h}_j[k]$ being the kth diagonal element as given by

$$\text{diag}(\mathbf{h}_j) = \begin{bmatrix} \mathbf{h}_j[1] & & \\ & \ddots & \\ & & \mathbf{h}_j[K] \end{bmatrix}$$

- $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_K)$ denotes the additive white Gaussian noise (AWGN) at the corresponding receiver and $\mathbf{n} = [\mathbf{n}[1], \mathbf{n}[2], \ldots, \mathbf{n}[K]]^T$.

2.2.3 SCMA MUD

In this section, we discuss the MUD process in SCMA. As previously mentioned, overloading in SCMA introduces significant MUI. Therefore, it is essential to implement an efficient MUD scheme to accurately detect each user's codeword in the presence of MUI. The perfor-

mance of the SCMA system is strongly influenced by the choice of the MUD method.

In MUD , perfect channel state information (CSI) is assumed to perform a MAP detection. However, MAP detection requires an exhaustive search, which significantly increases computational complexity. To address this, the sparse nature of the codewords is exploited to develop an MPA similar to that used for LDPC codes. In [14], an MPA is proposed for MUD in LDS-CDMA systems by utilizing their sparse structure. A similar technique is applied for SCMA MUD in [17], utilizing its sparse design. The MPA described in [14] is implemented using a SP algorithm detailed in [49]. The SP algorithm iteratively processes a FG, exchanging messages between neighboring nodes. These messages contain only extrinsic information, meaning that a message received via one edge cannot be used to update the message sent on the same edge. The SP algorithm iteratively finds locally optimal solutions by using the factored form of the global solution and combining these results to approximate the overall solution. Based on this foundation, we now present a detailed explanation of the MPA used for SCMA detection.

As discussed in 2.2.1, SCMA possesses a FG structure similar to that of LDS-CDMA described in [14]. In SCMA, FNs and VNs correspond to REs and UEs, respectively. The K REs receive the channel observation vector, with each FN carrying the observation for its associated RE as shown in Fig. 2.4. Each VN, on the other hand, holds the computed posterior probabilities for the M codewords of its corresponding UE. Computations are performed at both VNs and FNs, and the resulting messages are exchanged between neighboring nodes along the edges of the FG. It is to be noted that the computations at the FNs are particularly expensive because of the exponential complexity of their operations. During each iteration, the messages at the nodes are updated, and this iterative process continues until either convergence is achieved or the maximum number of iterations is reached.

In SCMA, as stated earlier, J UEs share K REs, where $J > K$. Multiple UEs overlap on each RE, and each UE spreads its data across multiple REs. For example, as shown in Fig. 2.4, users UE-1, UE-3, and UE-5 share RE-1, with a similar overlapping pattern observed for all REs. Additionally, UE-1 spreads its data across RE-1 and RE-2 and so forth.

The primary objective of the SCMA detector is to accurately separate each user's data on

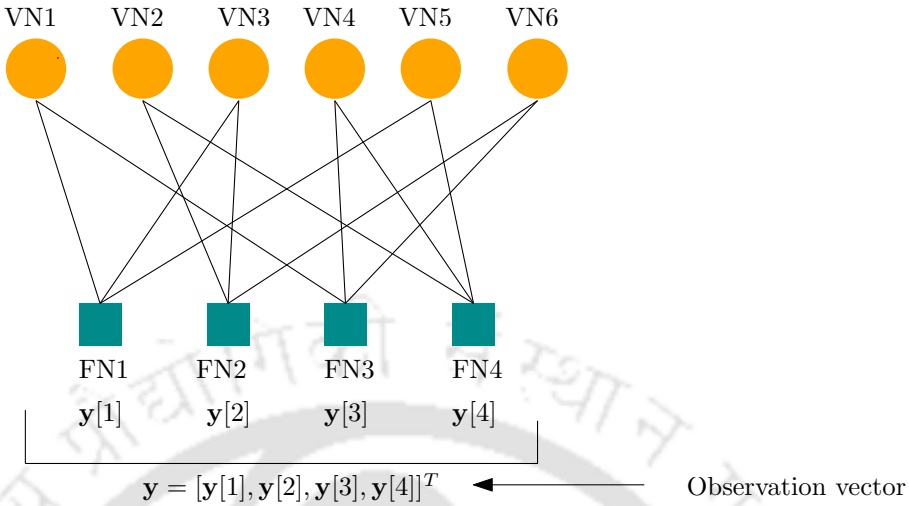


Fig. 2.4: FG aided MPA detection for SCMA

each RE. To achieve this, in the MPA, each FN performs local computations as defined in (2.9). These local computations are then aggregated to approximate the global MAP solution.

Updating message:

$V_{j\to k}^{(t)}$ and $U_{k\to j}^{(t)}$ represent the messages passed from jth VN to kth FN and kth FN to jth VN in FG in tth iteration, respectively. $V_{j\to k}^{(t)}$ and $U_{k\to j}^{(t)}$ get update in each iteration. The iterations of the MPA progress as follows:

Initialization: The M codewords of each user are equally likely to be transmitted and all the VNs are initialized with uniform probabilities, i.e., $V_{j\to k}^{(0)}(\mathbf{x}_j) = \frac{1}{M}$.

Step 1: FN update

The received signal on kth RE can be represented as $\mathbf{y}[k]$. Compute the probability distribution of mth codeword of the jth overlapped user on kth RE and pass it to the jth VN

$$U_{k\to j}^{(t)}(\mathbf{x}_{jm}) = \sum_{\mathbf{x}_i \in \mathbb{X}_i, \forall i \in \phi_k \setminus j} P[\mathbf{y}[k]|\mathbf{X}] \prod_{i \in \phi_k \setminus j} V_{i\to k}^{(t-1)}(\mathbf{x}_i) \tag{2.9}$$

where $\mathcal{X}_i$ is the ith user codebook , $\mathbf{X}$ is th multi user or joint codeword and $P[\mathbf{y}[k]|\mathbf{X}]$ is given by

$$P[\mathbf{y}[k]|\mathbf{X}] = \frac{1}{\pi N_0} \exp\left(-\frac{1}{N_0} \left| \mathbf{y}[k] - \sum_{j \in \phi_k} h_j[k] x_j[k] \right|^2 \right)$$

Step 2: VN update

Each VN use the messages received from the neighbouring FNs to estimate the probability

distribution of the all codewords $\mathbf{x}_j$, for $j = 1, 2, \ldots, J$.

$$V_{j\to k}^{(t)}(\mathbf{x}_j) = \prod_{l\in\varphi_j\setminus k} U_{l\to j}^{(t)}(\mathbf{x}_{jm})$$

Normalization for numerical stability given as

$$V_{j\to k}^{(t)}(\mathbf{x}_j) = \frac{V_{j\to k}^{(t)}(\mathbf{x}_{jm})}{\sum_{m'=1}^{M} V_{j\to k}^{(t)}(\mathbf{x}_{jm'})}$$

MPA performs **Step 1** and **Step 2** iteratively until it converges and the final estimated probabilities available at the VNs are as follows:

Step 3: Final estimation

The MPA algorithm terminates when the messages converge or the maximum number of iterations (T) is reached. Each VN then computes the final probability distribution of the codeword as given by

$$V_{j\to k}^{(T)}(\mathbf{x}_j) = \prod_{k\in\phi_j} U_{k\to j}^{(T)}(\mathbf{x}_j), \quad \mathbf{x}_j \in \mathcal{X}_j, \quad j = 1, 2, \ldots, J$$

Finally, the codeword with maximum $V_{j\to k}^{(T)}(\mathbf{x}_j)$ regarded as the transmitted codeword of the jth user.

2.3 KPIs of the Proposed SCMA Codebook Design Method

This section presents the definitions of the KPIs used to design the proposed SCMA codebooks. Let $\mathbf{S} = \sum_{j=1}^{J} \mathbf{x}_j$ be the superimposed codeword over AWGN and DL Rayleigh fading channel. Suppose $\mathbf{S}_a$ and $\mathbf{S}_b$ are two different superimposed SCMA codewords. Let $\mathbf{S}_a[k]$ and $\mathbf{S}_b[k]$ be the k-dimensional points corresponding to $\mathbf{S}_a$ and $\mathbf{S}_b$ respectively. The pair-wise error probability (PEP) over the AWGN channel, based on the system model in (2.6), is expressed as

$$P(\mathbf{S}_a \to \mathbf{S}_b) = Q\left(\sqrt{\frac{\|\mathbf{S}_a - \mathbf{S}_b\|^2}{2N_0}}\right) \tag{2.10}$$

where, N_0 is the one-sided power spectral density of AGWN. The MED of the superimposed constellation is defined as

$$\text{MED}_{\text{sup}} = \min_{a,b} \|\mathbf{S}_a - \mathbf{S}_b\|^2. \tag{2.11}$$

The proposed design method considers MED_{sup} as the KPI over AWGN channel.

The conditional PEP over DL channel based on the system model in (2.7) is expressed as

$$P(\mathbf{S}_a \rightarrow \mathbf{S}_b|\mathbf{h}) = Q\left(\sqrt{\frac{\|\text{diag}(\mathbf{h})(\mathbf{S}_a - \mathbf{S}_b)\|^2}{2N_0}}\right).$$

The PEP after averaging over the channel statistics becomes

$$P(\mathbf{S}_a \rightarrow \mathbf{S}_b) = E_{\mathbf{h}}\left[Q\left(\sqrt{\frac{\sum_{k=1}^{K} d_k^2|h[k]|^2}{2N_0}}\right)\right] \tag{2.12}$$

where the kth dimension-wise distance between $\mathbf{S}_a$ and $\mathbf{S}_b$ is defined as $d_k = |\mathbf{S}_a[k] - \mathbf{S}_b[k]|$. The dimension-wise minimum distance is expressed as $d_{k,\min} = \min_{a,b} d_k$.

At high SNR region, the upper bound on (2.12) according to [10] is given by

$$P(\mathbf{S}_a \rightarrow \mathbf{S}_b) \leq (1/4N_0)^{-|D(\mathbf{S}_a,\mathbf{S}_b)|} \prod_{k \in D(\mathbf{S}_a,\mathbf{S}_b)} d_k^{-2} \tag{2.13}$$

where $D(\mathbf{S}_a, \mathbf{S}_b) = \{k|d_k \neq 0, 1 \leq k \leq K\}$ and $N(\mathbf{S}_a, \mathbf{S}_b) = |D(\mathbf{S}_a, \mathbf{S}_b)|$ denotes the diversity order of SCMA. The MPD related to diversity order $N(\mathbf{S}_a, \mathbf{S}_b)$ is given by,

$$\text{MPD}^{(N(\mathbf{S}_a,\mathbf{S}_b))} = \min_{a,b}\left\{\prod_{k \in D(\mathbf{S}_a,\mathbf{S}_b)} d_k\right\}. \tag{2.14}$$

Observe from (2.13) that the MPD related to each diversity order from d_v to $d_v + \delta$ influences the upper bound on PEP; δ is selected judiciously based on performance-complexity trade-off. By complexity, we refer to the computational complexity involved in optimizing the codebook. The aim is to maximize the MPD of each such diversity order $N(\mathbf{S}_a, \mathbf{S}_b)$. The KPIs over the DL channel are considered as

$$d_{k,\min}, \quad \text{for } k = 1, 2, \ldots, K,$$
$$\text{MPD}^{(N(\mathbf{S}_a,\mathbf{S}_b))}, \quad \text{for } N(\mathbf{S}_a, \mathbf{S}_b) = d_v, d_v + 1, \ldots, d_v + \delta.$$

For the UL Rayleigh fading channel, suppose the transmitted codewords of J UEs constitute a matrix $\mathbf{U} = [\mathbf{x}_1, \ldots, \mathbf{x}_J]$ and the channel matrix is $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_J]$. Define a vector based on

the UL system model in (2.7)

$$\mathbf{u} = \Big[\sum_{j\in\phi_1} h_j[1]x_j[1], \ldots, \sum_{j\in\phi_K} h_j[k]x_j[k] \Big]^T \in \mathbb{C}^K,$$

where ϕ_k is the set of users overlap on the kth RE.

Let $\mathbf{u}_a$ and $\mathbf{u}_b$ be two different vectors corresponding to $\mathbf{U}_a$ and $\mathbf{U}_b$. Then the conditional PEP is given by

$$P(\mathbf{U}_a \to \mathbf{U}_b|\mathbf{H}) = Q\Bigg(\sqrt{\frac{||\mathbf{u}_a - \mathbf{u}_b||^2}{2N_0}} \Bigg).$$

At high SNR region, the upper bound on PEP is given by [10]

$$P(\mathbf{U}_a \to \mathbf{U}_b) \leq (1/4N_0)^{|D(\mathbf{S}_a,\mathbf{S}_b)|} \prod_{k\in D(\mathbf{U}_a,\mathbf{U}_b)} \lambda_k^{-2}, \tag{2.15}$$

where $\lambda_k^2 = \sum_{j\in\phi_k} |x_{j,a}[k] - x_{j,b}[k]|^2$ and $x_{j,a}[k]$ and $x_{j,b}[k]$ are the kth non-zero elements of codewords a and b of the jth user's codebook.

The MPD of the individual codebook is given by

$$\text{MPD}^{(d_v)} = \min_{a,b} \left\{ \prod_{k\in D(\mathbf{x}_a,\mathbf{x}_b)} |\mathbf{x}_a[k] - \mathbf{x}_b[k]| \right\}.$$

where $D(\mathbf{x}_a, \mathbf{x}_b) = \{k : |\mathbf{x}_a[k] - \mathbf{x}_b[k]| \neq 0, 1 \leq k \leq K\}$ and $\mathbf{x}_a$ and $\mathbf{x}_b$ are the codewords a and b of a particular user.

Observe from (2.15) that the MPD of the individual codebook determines the upper bound on the PEP. Therefore, $\text{MPD}^{(d_v)}$ is considered as the KPI for the UL channel.

Table 2.1 summarizes the KPIs used in the proposed method of SCMA codebook design for various channels.

Table 2.1: KPIs of the proposed codebook design method.

AWGN	DL Rayleigh fading	UL Rayleigh fading
MED_{sup}	$\text{MPD}^{(N(\mathbf{S}_a,\mathbf{S}_b))}$ $\forall N(\mathbf{S}_a, \mathbf{S}_b) = d_v, \ldots, d_v+\delta$	$\text{MPD}^{(d_v)}$

2.4 Proposed SCMA Codebook Design Procedure

This section begins by discussing the design of MC and then provides a detailed outline of the systematic steps involved in the design and optimization of codebooks. Multi-user codebooks are generated based on the underlined MCs, with each user employing a distinct MC derived from sub-constellations (SCs). Here, an SC refers to a one-dimensional complex constellation, similar to a rotated QAM constellation. In this work, various SCs are utilized to construct the MC, followed by codebook optimization. These diverse SCs form the foundation for the codebooks developed in this work. The choice of SCs in the codebook design significantly impacts the error rate performance of the SCMA system, as demonstrated by the KPIs in Table 2.2 and the numerical results discussed in Section 2.5.

Three different SCs are used to construct MCs in this work as follows:

(i) Algebraic lattice constellation (ALC).

(ii) Uniquely decomposable constellation groups (UDCG).

(iii) Rotated PAM (R-PAM).

Algebraic lattice constellation (ALC)

The concept of ALC, inspired by the idea of rotating $\mathbb{Z}^n$-lattice constellations as proposed in [50,51]. These rotations are applied to $\mathbb{Z}^n$-lattices with the objective of maximizing the signal space diversity order and MPD. Thus, ALC achieves maximum diversity [32,50]. The resulting codebooks, developed on the basis of ALC SCs, are referred to as ALC_{prop} in this work.

A rotation matrix is obtained by embedding the number field $\mathbb{K}$ (extension field of the set $\mathbb{Q}$ of all rational numbers) in $\mathbb{R}^n$, i.e., $\sigma : \mathbb{K} \rightarrow \mathbb{R}^n$ as

$$\sigma(x) = (\sigma_1(x), \sigma_2(x), \ldots, \sigma_n(x)),$$

where $x \in \mathbb{K}$ and $\sigma(x) \in \mathbb{R}^n$. The fundamental steps involved in the design of ALC are given

in [50]. The lattice generator matrix **M** is defined as

$$\mathbf{M} = \begin{bmatrix} \sigma_1(w_1) & \sigma_2(w_1)... & \sigma_n(w_1) \\ \sigma_1(w_2) & \sigma_2(w_2)... & \sigma_n(w_2) \\ \\ \sigma_1(w_n) & \sigma_2(w_n)... & \sigma_n(w_n) \end{bmatrix}$$

where $w_i \in \mathbb{K}$. A lattice can be written as an integer combination of the n independent vectors of **M**. The lattice $\Lambda(\mathbb{K})$ is constructed in the field $\mathbb{K}$ as

$$\Lambda(\mathbb{K}) = \{\mathbf{x} = \mathbf{M}\lambda \ \in \mathbb{R}^n | \lambda \in \mathbb{Z}^n\}.$$

The resulting ALC built on $\mathbb{Z}^n$, denoted as $\Lambda(\mathbb{K})$, achieves the highest signal space diversity order $L = n$ [50]. For the case of $n = 2$, i.e., a 2-dimensional integer lattice $\mathbb{Z}^2$, it generates two SCs, which are then used as the SCs for designing the MC.

Uniquely decomposable constellation group (UDCG)

SCMA codebooks based on UDCG, proposed in [26], are designed to form uniquely decomposable superimposed constellation points. UDCG ensures the maximum minimum superimposed Euclidean distance. In this work, UDCG is constructed following the approach in [26]. The resulting codebooks, developed on the basis of UDCG SCs, are referred to as $\text{UDCG}_{\text{prop}}$. The construction of UDCG SCs is given as follows.

$$\mathbf{s}_1 = Ae^{(j\frac{2\pi}{2^r}l)} \text{ and } \mathbf{s}_i = \mathbf{s}_1 e^{j\theta_i}, \ i = 2, \ldots, d_f,$$

where $A = [a_0, a_1, \ldots, a_V]^T$ with $a_0 < a_1 < \ldots < a_V$ and $V = \frac{M}{2^r} - 1$, and $\theta_i = \pm\frac{\pi}{r+\varepsilon}$. The number a_v can be generated as $a_v = a + vt$ with $a = 0.5, t = 1.15$, and $l \in \{0, 1, \ldots, 2^r - 1\}$ with $r \leq \log_2(M)$ and $\varepsilon > 0$.

R-PAM Constellation

R-PAM-based SCMA codebooks, presented in [23], form the foundation for designing $\text{R-PAM}_{\text{prop}}$ codebooks in this work. R-PAM ensures the larger MED between the codewords. The R-PAM SC generation process is as follows.

$$\mathbf{s}_1 = A_m(1+j) \text{ and } \mathbf{s}_i = e^{(j\theta_{i-1})}\mathbf{s}_1,$$

where $A_m = 2m - 1 - M, m = 1, \ldots, M$ and $\theta_{i-1} = (i-1)\frac{\pi}{Md_f}$, $i = 2, \ldots, d_f$.

2.4.1 MC design

In this section, the concept of MC involves the construction of a d_v-dimensional constellation by the composition of d_v SCs. There are three distinct types of SC candidate sets: ALC, UDCG, and R-PAM. These three sets of SC candidates correspond to three different sets of MC candidates, which, in turn, lead to three sets of distinct SCMA codebook candidate sets. The construction of MC for each UE is based on the generator matrix **G** as given in (2.16), which is derived from the FG matrix of the SCMA system as given in (2.3).

$$\boldsymbol{G} = \begin{bmatrix} \mathbf{s}_1 & 0 & \mathbf{s}_2 & 0 & \mathbf{s}_3 & 0 \\ 0 & \mathbf{s}_3 & \mathbf{s}_1 & 0 & 0 & \mathbf{s}_2 \\ \mathbf{s}_2 & 0 & 0 & \mathbf{s}_1 & 0 & \mathbf{s}_3 \\ 0 & \mathbf{s}_1 & 0 & \mathbf{s}_3 & \mathbf{s}_2 & 0 \end{bmatrix} \tag{2.16}$$

For example, the generator matrix **G** for an SCMA system overloaded with 150% is derived from the matrix **F** presented in Section 2.2.1. The matrix **G** is constructed by placing the SCs at the positions of the unit elements, following Latin-rectangular mappings. These mappings ensure that the nonzero elements in each row and column are distinct [13], which helps improve the performance [25]. As mentioned above, in this example, d_v=2, resulting in a two-dimensional MC. MC of the jth UE is denoted by MC_j. The first column of **G** corresponds to UE-1, where MC_1 is formed by combining SCs $\mathbf{s}_1$ and $\mathbf{s}_2$. Similarly, for UE-2, MC_2 is constructed by combining SCs $\mathbf{s}_3$ and $\mathbf{s}_1$.

$$\text{MC}_1 = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix}, \qquad \text{MC}_2 = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2. \end{bmatrix} \tag{2.17}$$

Similarly, for $j = 3, 4, 5, 6$, MC_j is derived from the corresponding SCs, as presented in $\boldsymbol{G}$. Subsequently, the codebook for the jth UE is generated by multiplying MC_j with its mapping

matrix $\mathbf{V}_j$, as defined in (2.2.1). The codebooks for UE-1 and UE-2 are constructed as follows.

$$\mathbf{C}_1 = \mathbf{V}_1 \times \mathrm{MC}_1, \qquad \mathbf{C}_2 = \mathbf{V}_2 \times \mathrm{MC}_2. \tag{2.18}$$

Similarly, for $j = 3, 4, 5, 6$, $\mathbf{C}_j$ can be obtained from MC_j and $\mathbf{V}_j$. In this way, three different SCMA codebook candidates $\mathrm{ALC}_{\mathrm{prop}}$, $\mathrm{UDCG}_{\mathrm{prop}}$, and $\text{R-PAM}_{\mathrm{prop}}$ are designed. The following section outlines the steps involved in the codebook optimization process.

2.4.2 Codebook optimization procedure

In this section, we discuss the steps involved in optimizing the codebooks. The codebook optimization for AWGN or Rayleigh fading channels consists of the following two steps.

- **Step 1**: Optimize user-specific rotation angles $\theta_1, \theta_2, \ldots, \theta_{d_f}$, which are used in the multiplexing operation.
- **Step 2**: Apply the dimensional permutation on each user's MC.

These two steps are carried out for the codebook optimization, and the procedure is outlined in **Algorithm** 1. The detailed descriptions of **Step 1** and **Step 2** are as follows.

Step 1: Determining the optimum rotation angles

This step determines the optimum rotation angles for both AWGN and DL Rayleigh fading channels. Since the same SCs ($\mathbf{s}_1$,$\mathbf{s}_2$, and $\mathbf{s}_3$) are repeated across all REs as mentioned in (2.16), optimizing the rotation angles in a single dimension is sufficient. $\mathbf{s}_i^{(k_i)}$ denotes the k_ith point in the SC $\mathbf{s}_i$. Let $\mathbf{S}_a[k] = \sum_{i=1}^{d_f} \mathbf{s}_i^{(k_i)} e^{(j\theta_i)}$ be the k-dimensional rotated superimposed signal point corresponding to $\mathbf{S}_a$, where a is a function of k_i, and $i = 1, \ldots, d_f$.

For the AWGN channel, the optimum rotation angles are determined by minimizing the following metric derived from the upper bound on (2.10):

$$\theta_1, \ldots, \theta_{d_f} = \underset{\theta_1,\ldots,\theta_{d_f}}{\arg\min} \sum_{a \neq b} \exp(-|\mathbf{S}_a[k] - \mathbf{S}_b[k]|^2). \tag{2.19}$$

For the DL Rayleigh fading channel, the optimum rotation angles are determined by maximizing the dimension-wise minimum euclidean distance between $\mathbf{S}_a[k]$ and $\mathbf{S}_b[k]$:

$$\theta_1, \ldots, \theta_{d_f} = \underset{\theta_1,\ldots,\theta_{d_f}}{\arg\max}\, d_{k,\min} = \underset{\theta_1,\ldots,\theta_{d_f}}{\arg\max} \min_{a,b} |\mathbf{S}_a[k] - \mathbf{S}_b[k]|. \tag{2.20}$$

We set $\theta_1 = 0$. The remaining rotation angles are obtained through an exhaustive search method.

Step 2: Dimensional permutation

In this step, the dimensional permutation is applied systematically to each user's MC. Each MC is selected according to the structure given by $\mathbf{G}$. Consider two SCs $\mathbf{s}_{i_1}$ and $\mathbf{s}_{i_2}$ for a given MC. Select symbols randomly from the M symbols of $\mathbf{s}_{i_1}$ for the first dimension of MC. Further, to select the symbols for the second dimension corresponding to $\mathbf{s}_{i_2}$, there are $M!$ possible ways of arranging these points. Therefore, each user makes $P = M!$ MCs available after dimensional permutation.

For the AWGN channel, we select and preserve the MCs that give the maximum $\text{MED}_{\text{MC}} = \min_{i,i'} \|\mathbf{c}_i - \mathbf{c}_{i'}\|$ for each user according to (2.21) in **Algorithm 1**, where $\mathbf{c}_i$ and $\mathbf{c}_{i'}$ denote two different codewords in the MC. Suppose $\mathcal{I}_j$ is the set of MC indices preserved for user j. Then we calculate MED_{sup} using the preserved MCs after multiplexing as given in (2.22) of **Algorithm 1**. This process determines the indices $\mathcal{E}_j$ of high quality codebooks for each user in the AWGN channel, based on the maximization of the MED.

For the DL Rayleigh fading channel, the dimensional permutation is applied to minimize the PEP related to each diversity order $N(\mathbf{S}_a, \mathbf{S}_b) = d_v, \ldots, d_v + \delta$. The objective is to minimize the overall PEP in (2.13). The PEP corresponding to each $N(\mathbf{S}_a, \mathbf{S}_b)$ is minimized by maximizing $\text{MPD}^{(N(\mathbf{S}_a,\mathbf{S}_b))}$ of the respective diversity order according to (2.23) and (2.25)given in **Algorithm 1**. From (2.24), we find the set $\mathcal{U}_1^{(d_v+1)}$ of users paired with user-1 sending different codewords in $\mathbf{S}_a$ and $\mathbf{S}_b$ which result in $N(\mathbf{S}_a, \mathbf{S}_b) = d_v + 1$ after superposition. From (2.25), we obtain the codebooks ofuser-1 which is associated with user-j'. From (2.26), one can obtain the user-1 codebooks which give the maximum $\text{MPD}^{(d_v+1)}$ with all its associated users in $\mathcal{U}_1^{(d_v+1)}$.

In the case of UL Rayleigh fading, $\text{MPD}^{(d_v)}$ is the KPI as discussed in Section 2.3. **Algorithm 1** gives the first priority to $\text{MPD}^{(d_v)}$ and accordingly the dimensional permutation is carried out as shown in (2.23). Therefore, the codebooks designed for the DL channel perform equally well for the UL channel.

Algorithm 1 SCMA codebook design for AWGN and Rayleigh fading channel

Input: $\mathbf{G}$, $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{d_f}\}$;
Output: $C = \{C_1, \ldots, C_J\}$.
Step 1: **Constellation rotation**
for $\theta_1, \theta_2, \ldots, \theta_{d_f} \leftarrow 1 : 180$ **do**
    Optimize $\theta_1, \theta_2, \ldots, \theta_{d_f}$ according to the criteria given in (2.19) (AWGN channel). or
    Optimize $\theta_1, \theta_2, \ldots, \theta_{d_f}$ according to the criteria given in (2.20) (DL Rayleigh fading channel).
end for
Step 2: **Dimensional permutation**
For AWGN channel
Perform dimensional permutation and out of $P = M!$ possibilities, preserve those MCs having the maximum MED_{MC} as given below

$$\mathcal{I}_j = \arg\max_{p=1,\ldots,P} \text{MED}_{\text{MC}} \tag{2.21}$$

where $\mathcal{I}_j$ denotes indices of the preserved user-j MCs.
Use the preserved MCs of each user to maximize MED_{sup}:

$$\{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_J\} = \arg\max_{u_j \in \mathcal{I}_j, j=1,\ldots,J} \text{MED}_{\text{sup}} \tag{2.22}$$

where $\mathcal{E}_j$ denotes the indices u_js selected from $\mathcal{I}_j$.
For Rayleigh fading channel
Step 2a: Perform dimensional permutation and preserve the MCs with the maximum $\text{MPD}^{(d_v)}$:

$$\mathcal{I}_j = \arg\max_{p=1,\ldots,P} \text{MPD}^{(d_v)} \quad , \forall j = 1, \ldots, J \tag{2.23}$$

Initialize: $j = 1$ and $N(\mathbf{S}_a, \mathbf{S}_b) = d_v + 1$;
Step 2b: For user-1, find those user-j' such that the two-user superimposed codewords $\mathbf{S}_a$ and $\mathbf{S}_b$ produces $N(\mathbf{S}_a, \mathbf{S}_b) = d_v + 1$:

$$\mathcal{U}_1^{(d_v+1)} = \{j' | \, \mathbf{S}_a \text{ and } \mathbf{S}_b \text{ differ in user-1and user-}j' \text{ codewords which results } N(\mathbf{S}_a, \mathbf{S}_b) = d_v + 1\}. \tag{2.24}$$

Step 2c: Use the preserved MCs in (2.23) of user-1 (from $\mathcal{I}_1$) and user-j' (from $\mathcal{I}_{j'}$) to maximize the $\text{MPD}^{(d_v+1)}$, $\forall j' \in \mathcal{U}_1^{(d_v+1)}$. Obtain the indices of user-1 codebooks according to the criteria given below:

$$\{\mathcal{E}_{1,j'}\} = \arg\max_{u_1 \in \mathcal{I}_1} \text{MPD}^{(d_v+1)} \qquad \forall j' \in \mathcal{U}_1^{(d_v+1)} \tag{2.25}$$

Step 2d: Extract the codebooks for user-1 which is intersection of $\mathcal{E}_{1,j'}$, $\forall j' \in \mathcal{U}_1^{(d_v+1)}$:

$$\{\mathcal{E}_1\} = \bigcap_{j' \in \mathcal{U}_1^{(d_v+1)}} \mathcal{E}_{1,j'}, \tag{2.26}$$

where $\mathcal{E}_1$ is the set of indices of good user-1 codebooks.
Step 2e: Repeat the above three steps for $N(\mathbf{S}_a, \mathbf{S}_b) = d_v + 2, \ldots, d_v + \delta$.
Step 2f: Repeat the above four steps $\forall j = 2, \ldots, J$.
Final Step (for both AWGN and Rayleigh): For every $j = 1, \ldots, J$, select one index randomly from $\mathcal{E}_j$ and store the corresponding codebook in C_j.

2.5 Simulations and Discussions

Simulations are conducted to evaluate the SER of the SCMA system for overloading factors of 150% ($J = 6$, $K = 4$, $\delta = 1$) and 200% ($J = 10$, $K = 5$, $\delta = 2$) across different channel

Table 2.2: MED_{sup} with optimum rotation angles for AWGN channel in 150% overloading.

SC	MED_{sup}	θ_{opt}
$\text{UDCG}_{\text{prop}}$	1.0046	(2.8798,0.2618)
$\text{R-PAM}_{\text{prop}}$	1.0954	(1.8326,0.526)
Chen [25]	1.0708	none
GAM [24]	0.8901	none

Table 2.3: KPIs with optimum rotation angles for Rayleigh fading channel in 150% overloading.

SC	$\text{MPD}^{(d_v)}/\text{MPD}^{(d_v+1)}$	$d_{k,\min}$	θ_{opt}
ALC_{prop}	1.000/0.4632	0.4181	(0.0524,2.3213)
$\text{UDCG}_{\text{Prop}}$	0.8318/0.6832	0.4820	(2.6005,1.8675)
$\text{R-PAM}_{\text{prop}}$	0.801/0.6901	0.4407	(1.0647,0.2967)
Zhang [26]	0.3364/0.1846	0.2128	(-0.7854,0.7854)
GAM [24]	0.4162/0.2931	0.3246	none
Chen [25]	1.000/0.4276	0.4517	none
Li [52]	0.702/0.3041	0.4128	none

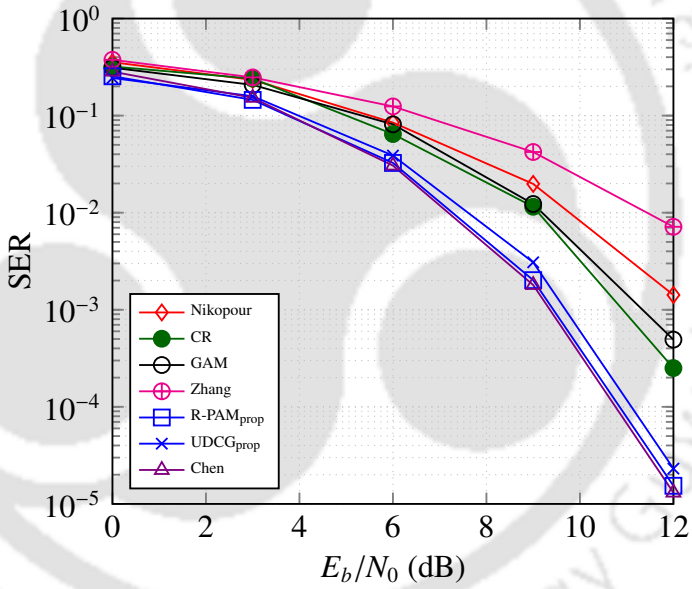


Fig. 2.5: SER performance comparison over AWGN channel for $M = 4$ and λ=150% overloading.

conditions. The SCMA detection process employs the MPA algorithm with a maximum of 10 iterations. The SER performance of the proposed codebooks (ALCprop, R-PAMprop, and UCG_{prop}) is compared to the existing reference codebooks from [22, 24–26, 40, 52].

Table 2.2 and Table 2.3 present a comparative analysis of the KPIs of different codebooks for the AWGN and Rayleigh fading channels, respectively. The proposed codebooks achieve the highest KPI values among all benchmark codebooks and exhibit nearly identical KPI values to the near-optimal codebooks proposed in Chen [25]. Furthermore, for a 150% overloading factor, the SER performance of the proposed codebooks is nearly equivalent to that of the Chen [25] codebook.

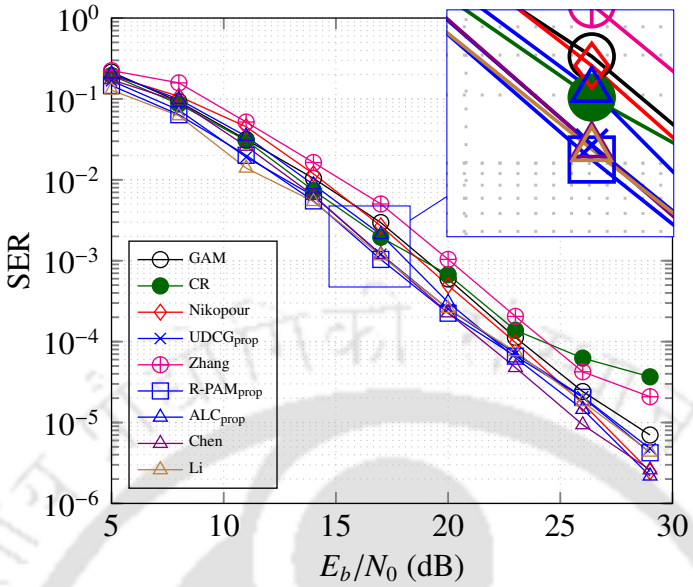


Fig. 2.6: SER performance comparison over DL channel for $M = 4$ and $\lambda = 150\%$.

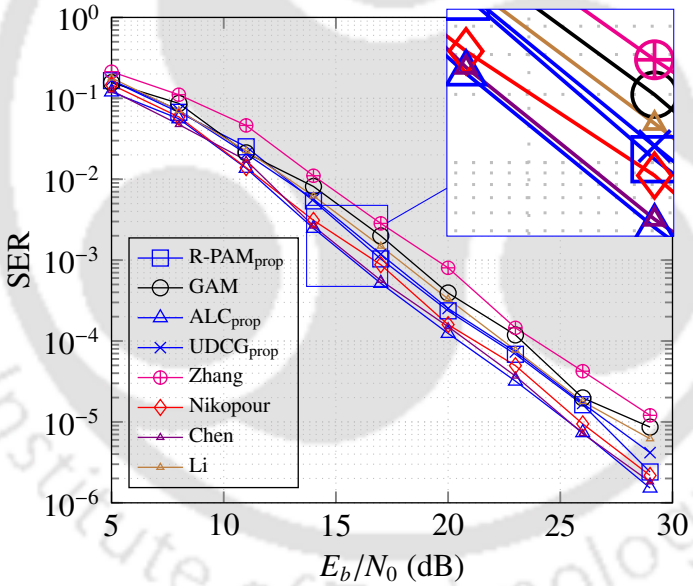


Fig. 2.7: SER performance comparison over UL channel for $M = 4$ and $\lambda = 150\%$.

In the AWGN channel, Fig. 2.5 illustrates that the proposed codebooks offer performance gains of approximately 4.5dB and 3dB over the UDCG [26] and GAM [24] codebooks, respectively, at SER = 10^{-3}.

In Rayleigh fading channels, the proposed codebooks show significantly improved performance in 200% overloaded scenarios. Fig. 2.9 explicitly shows that all three proposed codebooks over UL channel outperform Li [52] codebook with a gain of around 3 dB at SER = 10^{-5}. Observe from Fig. 2.8 that over DL channel, two of the proposed codebooks outperform Li [52]

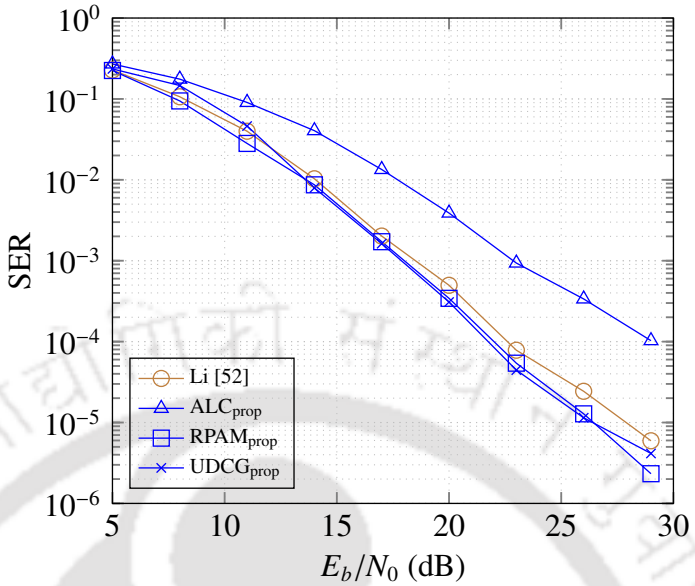


Fig. 2.8: SER performance comparison over DL channel for $M = 4$ and $\lambda = 200\%$ overloading.

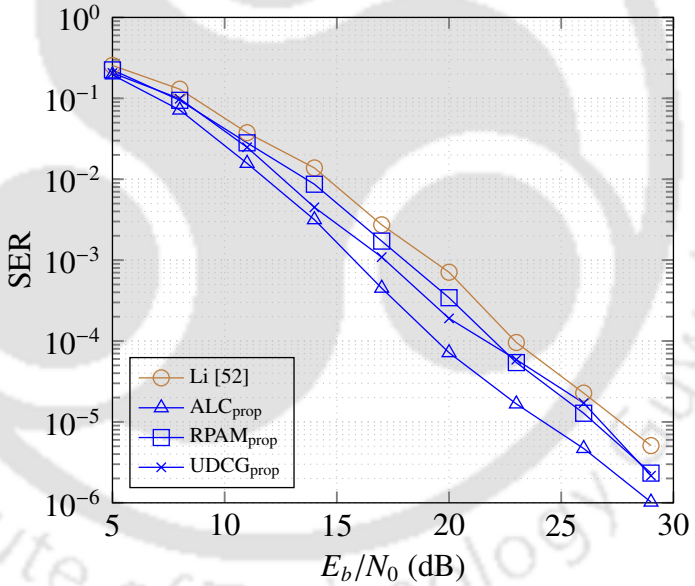


Fig. 2.9: SER performance comparison over UL channel for $M = 4$ and $\lambda = 200\%$ overloading.

with a gain of around 2 dB at SER $= 10^{-5}$. Fig. 2.6 shows that the proposed codebooks provides 0.9 dB gain over UDCG [26] and 0.7 dB gain over both Li [52] and GAM [24] codebooks at SER $= 10^{-5}$. From Fig. 2.7, observe that ALC_{prop} outperforms the other codebooks. It achieves a maximum of 2 dB gain over UDCG [26] and GAM [24], 1.2 dB gain over Li [52] codebooks. Higher SER gains are achieved in large scale SCMA parameters ($K = 5$, $J = 10$ and $M = 8$) as shown in Fig. 2.8, Fig. 2.9 and Fig. 2.10. This is one of the important features of the proposed codebook design. It can be seen in Fig. 2.7 and Fig. 2.9, that the codebooks

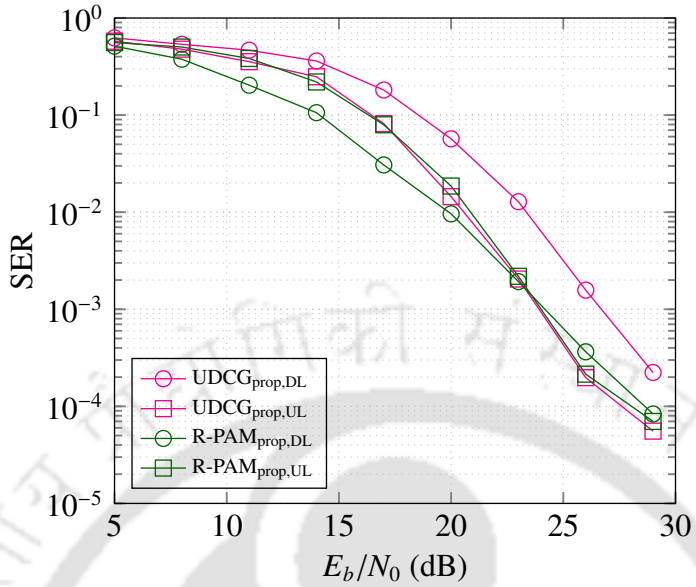


Fig. 2.10: SER performance comparison over UL and DL channel for $M = 8$ and $\lambda = 200\%$ overloading.

which are designed for DL Rayleigh fading channels can be used for the UL Rayleigh fading channels without any error performance degradation. The designed codebooks can be found at https://github.com/commletters/SCMA-codebooks.

2.6 Summary

In this chapter, we present a novel algorithm for designing SCMA codebooks tailored to various channel conditions. The algorithm introduces new KPIs specifically for DL Rayleigh fading channels and incorporates an enhanced sequence of optimization steps for codebook design. These advancements led to notable performance improvements, as demonstrated by the numerical experimental results discussed in Section 2.5. Moreover, the proposed algorithm, combined with the defined KPIs, simplifies the codebook design process for Rayleigh fading channels. Consequently, the codebook design for DL Rayleigh fading channels gives competitive performance also for UL Rayleigh fading channels.

This chapter explored sparse CD-NOMA systems, with a primary emphasis on SCMA codebook design. On the receiver side, MPA was utilized for SCMA detection. However, as previously discussed, MPA suffers from exponential computational complexity. The next chapter of this thesis shifts focus to dense CD-NOMA (DCMA) to harness the full diversity provided by the codebooks. Due to the dense FG structure of DCMA, MPA becomes impractical for

detection in terms of both complexity and convergence. Therefore, it is essential to develop a novel and efficient detection scheme for DCMA systems that also maintains low complexity and performs effectively for SCMA systems.

3

ADMM-based Detector for Large-scale MIMO Dense Code-domain NOMA Systems

Contents

3.1 Introduction

The existing literature on NOMA extensively studies two types of CD-NOMA systems. The first type is sparse CD-NOMA systems, which utilize low-density spreading sequence design (e.g., LDS) or sparse codebook design (e.g.,SCMA) [22]. The second type is dense CD-NOMA systems, which utilize dense spreading sequence design (e.g., overloaded CDMA or MUSA [15]) or dense codebook design (e.g., DCMA) [21]. This chapter focuses on the design of efficient detector for sparse and dense codebook-based NOMA systems, such are SCMA and DCMA, respectively.

Multi-input multi-output (MIMO) technology is another crucial enabler in improving spectral efficiency and system reliability [53]. MIMO aided CD-NOMA systems (MIMO-CD-NOMA) enhances the throughput and spatial diversity of CD-NOMA systems. This work considers three types of uplink (UL) MIMO-CD-NOMA systems: (1) Single-input multi-output (SIMO) aided CD-NOMA, where each user equipment (UE) has a single antenna [54]; (2) Spatial multiplexing-CD-NOMA (SMX-CD-NOMA), where each UE is equipped with multiple transmit antennas [55]; and (3) Spatial modulated CD-NOMA (SM-CD-NOMA), where each UE activates a single transmit antenna out of multiple transmit antennas at the UE [56].

The primary challenge in implementing MIMO-CD-NOMA systems lies in designing a multi-user signal detector that offers excellent performance while maintaining affordable complexity. The MPA is commonly used and effective for SCMA detection [30, 57, 58]. However, MPA exhibits exponential complexity with the codebook size (M), the number of overlapped UEs (d_f) on each resource element (RE). The computational complexity of MPA dramatically increases in SMX-CD-NOMA systems with the number of N_t transmit antennas [59] and MPA becomes impractical for SCMA detection. The linear complexity expectation propagation algorithm (EPA) is an efficient detector for coded MIMO-SCMA systems [60]. For uncoded SCMA systems, the EPA often suffers from high error floors [61]. An iterative Gaussian approximated MPA (GAMPA) is proposed for detecting MIMO-SCMA systems [62]. The Gaussian approximation of interference further degrades the performance [62]. All the above mentioned

nonlinear detectors operate based on sparse FG representation of SCMA systems. Therefore, these algorithms often face convergence issues when the FG contains short cycles.

To overcome the limited diversity issues (due to sparsity) in SCMA, researchers have recently started designing the DCMA systems [21]. DCMA leads to dense factor graphs with numerous short cycles, making MPA and EPA unsuitable solutions for such systems. Furthermore, to the best of authors' knowledge, no universal algorithm exists to solve the detection problem of various MIMO CD-NOMA systems. Thus, an alternate detection algorithm is needed to achieve the full diversity offered by DCMA with minimal detection complexity across different MIMO variants.

The generalized sphere decoder (GSD) is commonly employed for detecting single-input single-output (SISO)-based DCMA signals [21]. The GSD operates using a tree search algorithm, with its complexity determined by the number of tree nodes and the floating-point operations (FLOPs) required at each node. However, the computational cost of GSD escalates significantly for large-scale MIMO DCMA systems. Consequently, there is a need for an efficient detection algorithm to effectively integrate DCMA with various MIMO systems.

The existing research mainly focuses on nonlinear detectors (MPA, EPA, and sphere decoder) for CD-NOMA systems [57, 63]. On the other hand, linear minimum mean square error (LMMSE) equalization is a commonly used linear detector in standalone MIMO systems [64]. LMMSE is simple to implement and exhibits low computational complexity over nonlinear detectors. However, the performance of LMMSE degrades as the number of unknowns approaches the number of observations [65]. Several advanced signal recovery algorithms based on the approximate message passing (AMP) principle [66] have been developed, including orthogonal AMP (OAMP) [67] and memory AMP (MAMP) [68]. These algorithms are primarily utilized for sparse recovery [67] and multi-user MIMO (MU-MIMO) [69] detection problems.

Recently, the ADMM has been widely used to solve convex and non-convex problems in a distributed manner [70]. Glowinski and Marrocco first proposed the ADMM in the mid-1970s. Later on, Boyd *et al.* rigorously discussed various complex optimization problems that can be formulated and solved using ADMM in their tutorial [70]. The ADMM is formed by the com-

position of dual ascent and the method of multipliers. The dual decomposition properties of dual ascent make the ADMM solve in a distributed manner. ADMM has superior convergence properties and often is able to converge without the requirement of strict convexity. A comprehensive treatment of ADMM and its advancements are detailed in [71]. The preliminaries of ADMM is discussed in Section 3.2.2.

The ADMM is extensively applied to solve the linear programming (LP) problem in low-density parity-check code's (LDPC) decoding [72, 73]. The authors here have also shown that the ADMM significantly reduces the complexity of the decoder compared to the belief propagation (BP) methods. The works in [74] proposed a novel approach for multiuser detection in uplink grant-free NOMA systems, leveraging compressive sensing techniques and the ADMM. However this work is not related to the CD-NOMA systems. The ADMM-based infinity norm (ADMIN) iterative linear detector is proposed for massive MIMO systems in [75]. The authors have also proposed VLSI architecture for the ADMIN detector. The results show that ADMIN outperforms all the linear detectors in terms of SER performance and low hardware cost compared to the nonlinear BP-based detectors. Similarly, an ADMM-based QAM signal detector for massive MIMO systems is proposed in [65]. A distributed penalty sharing (DPS) ADMM method is designed to convert the maximum-likelihood (ML) detection problem into sharing optimization problem. This method shows good performance and complexity trade-offs for massive MIMO systems. The authors in [65, 75] show that the ADMM-based detector performs LMMSE equalization in the first iteration. The performance of the ADMM improves over LMMSE as iterations progress.

3.1.1 Focus and Contributions

In this chapter, we propose an ADMM-based iterative linear detector that solves the MUD problem of large-scale MIMO-aided CD-NOMA systems. It exhibits the best trade-off between the SER performance and computational complexity for SCMA and DCMA systems. The main contributions of this work are listed below:

• ***DCMA detection using an ADMM-based iterative linear equalizer*:** This work represents the first attempt to design a detector for dense CD-NOMA systems, specifically DCMA, using

the ADMM framework. The ADMM-based detector leverages the full diversity gain inherent to DCMA systems. It is an iterative linear equalizer that demonstrates substantial signal-to-noise ratio (SNR) improvements compared to its counterpart, the LMMSE detector. Notably, the ADMM detector achieves significant SNR enhancements for both SMX-DCMA and SM-DCMA configurations, as evidenced by the results.

• ***Conversion of ML detection problem into sharing optimization problem*** **:** The complexity of the exhaustive search involved in the ML MUD problem increases exponentially as the number of UEs and codebook size increase. To overcome this challenge, the ML problem is transformed into a non-convex sharing optimization problem. The novel box constraint relaxation (BCR) method on each user's complex vector codeword is proposed to solve the formulated sharing problem via the ADMM algorithm.

• ***Resource wise processing via ADMM algorithm*****:** The CD-NOMA system models, i.e., the relationship between the input signal and received signal, is reformulated so that the sharing ADMM algorithm can be readily applied for detection. The reformulation facilitates the resource-wise processing of the observation vector via ADMM detection. The resource-wise processing makes linear equalization possible at the receiver.

• ***Low computational complexity for various MIMO CD-NOMA systems*****:** The complexity of the proposed ADMM-based detector is independent of the CD-NOMA system's codebook size/modulation order (M), unlike the conventional MPA detector. The ADMM-based detector exhibits polynomial complexity with all other parameters, such as the numbers of antennas, UEs, and REs. Thus, it can be applied to large-size codebooks.

• ***Promising performance*****:** The comprehensive comparison reveals that the proposed detector delivers significant performance enhancements compared to various existing detectors. Specifically, the proposed soft-output-based ADMM detector, when applied to LDPC coded CD-NOMA systems, demonstrates notable improvements. One key aspect of the ADMM detector is its robustness in the presence of channel estimation errors (CEEs), which underscores its resilience and reliability in practical scenarios.

Notations: Lower case, bold lower case, and bold upper case letters denote scalars, vectors,

and matrices, respectively. $(\cdot)^T$ and $(\cdot)^H$ denote transpose and Hermitian transpose, respectively. $\|(\cdot)\|$ denotes the Euclidean norm of a vector. $\prod_{[-\alpha,\alpha]}(\cdot)$ denotes the Euclidean projection onto the interval $[-\alpha, \alpha]$. $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\mathrm{Re}(\cdot)$ denotes the real part of a complex variable. $\mathbb{R}^n$ and $\mathbb{C}^n$ denote n-dimensional real and complex vector spaces, respectively. $\mathcal{CN}(0, \sigma^2)$ denotes complex Gaussian distribution with zero mean and variance σ^2.

3.2 Preliminaries

This section primarily covers the fundamentals of the CD-NOMA system models and outlines the steps for addressing the sharing optimization problem using the ADMM approach. The SCMA system model preliminaries are elaborated in Section 2.2 of Chapter 2. The following discussion focuses on the preliminaries of the DCMA system model and the ADMM algorithm.

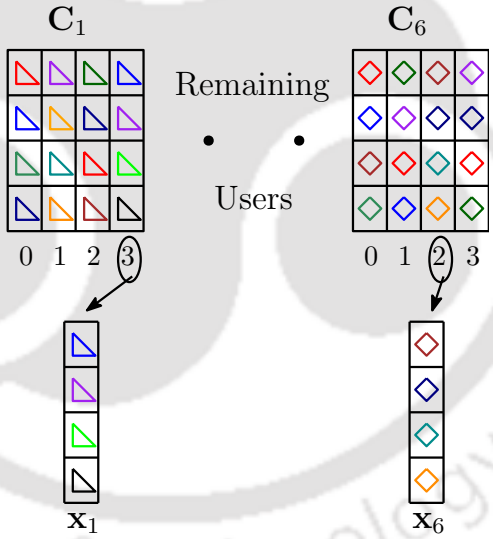


Fig. 3.1: DCMA codebook structure for $J = 6, K = 4$.

3.2.1 DCMA System Model

The error performance of SCMA is constrained by its limited diversity order, which corresponds to the number of non-zero entries in the codebook, as illustrated in Fig. 2.3 in **Chapter** 2. This limitation of SCMA can be overcome by using dense codebooks in the DCMA system [21]. Each UE in the DCMA system has access to all K REs. No zero entries are present in the K-dimensional dense codewords. Therefore, the DCMA system exploits the full diversity of multi-dimensional codewords. For example, the codebook structure of 150 % overloaded

DCMA system is shown in Fig. 3.1. The dense nature of DCMA allows all J UEs overlap on each RE, i.e., $N_u = J$. The observation vector of a DCMA system across different channels is identical to that of the SCMA system, as described in Section 2.2.2.2.

3.2.2 ADMM formulation for sharing problem

This subsection introduces the fundamental concepts of solving sharing problems using ADMM within a distributed optimization framework [70]. The ADMM is formed by combining the superior properties of dual ascent and the method of multipliers. This combination ensures the robustness of ADMM [70]. The generic sharing problem is

$$\min \quad \sum_{i=1}^{N} f_i(\mathbf{x}_i) + g\left(\sum_{i=1}^{N} \mathbf{x}_i\right), \tag{3.1}$$

where $\mathbf{x}_i \in \mathbb{R}^n, i = 1, \ldots, N$, the associated local cost function $f_i(\mathbf{x}_i)$ ($f_i : \mathbb{R}^n \to \mathbb{R}$) of subsystem i is handled by processor i, and g ($g : \mathbb{R}^n \to \mathbb{R}$) is the shared objective, whose argument is the sum of N variables. Each variable $\mathbf{x}_i$ is involved in minimizing the individual cost $f_i(\mathbf{x}_i)$, as well as the shared objective $g\left(\sum_{i=1}^{N} \mathbf{x}_i\right)$. The sharing problem can be converted into an ADMM problem by introducing an alternative variable, $\mathbf{z}_i \in \mathbb{R}^n$. Thus, the cost function is minimized over $\mathbf{x}_i$ and $\mathbf{z}_i$, alternatively. The problem (3.1) is converted to the following problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} f_i(\mathbf{x}_i) + g\left(\sum_{i=1}^{N} \mathbf{z}_i\right) \\ \text{s.t.} \quad & \mathbf{x}_i - \mathbf{z}_i = 0 \quad i = 1, \ldots, N. \end{aligned} \tag{3.2}$$

The augmented Lagrangian function for (3.2) can be written as

$$\mathcal{L}\left(\{\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^{N}\right) = \sum_{i=1}^{N} f_i(\mathbf{x}_i) + g\left(\sum_{i=1}^{N} \mathbf{z}_i\right) + \sum_{i=1}^{N} \langle \mathbf{x}_i - \mathbf{z}_i, \mathbf{y}_i \rangle + \frac{\rho}{2} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{z}_i\|_2^2, \tag{3.3}$$

where $\mathbf{x}_i, \mathbf{z}_i$ are called primal variables, $\mathbf{y}_i \in \mathbb{R}^n$ is the Lagrangian variable and $\rho > 0$ is called the penalty parameter. The function in (3.3) can be minimized by the ADMM steps [70] as follows

$$\mathbf{x}_i^{t+1} := \arg\min_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z}_i^t + \mathbf{u}_i^t\|_2^2 \right) \tag{3.4}$$

$$\mathbf{z}_i^{t+1} := \arg\min_{\mathbf{z}_i} g\left(\sum_{i=1}^{N} \mathbf{z}_i\right) + \frac{\rho}{2} \sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{u}_i^t - \mathbf{x}_i^{t+1}\|_2^2 \tag{3.5}$$

$$\mathbf{u}_i^{t+1} := \mathbf{u}_i^t + \mathbf{x}_i^{t+1} - \mathbf{z}_i^{t+1}, \tag{3.6}$$

where $\mathbf{u}_i$ is called a dual variable (scaled version of Lagrangian variable, $\mathbf{u}_i = \frac{\mathbf{y}_i}{\rho}$).
The problems (3.4) and (3.6) can be solved in parallel for $i = 1, \ldots, N$. Let $\mathbf{q}_i = \mathbf{u}_i^t + \mathbf{x}_i^{t+1}$. Then (3.5) can be rewritten as

$$\begin{aligned} \min \quad & g\left(\sum_{i=1}^{N} \mathbf{z}_i\right) + \frac{\rho}{2} \sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{q}_i\|_2^2 \\ \text{s.t} \quad & \bar{\mathbf{z}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \end{aligned} \tag{3.7}$$

with fixed variable $\bar{\mathbf{z}} \in \mathbb{R}^n$. The problem (3.7) has the following solution

$$\mathbf{z}_i = \mathbf{q}_i + \bar{\mathbf{z}} - \bar{\mathbf{q}}, \quad \text{where} \quad \bar{\mathbf{q}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_i = \bar{\mathbf{u}}^t + \bar{\mathbf{x}}^{t+1}, \tag{3.8}$$

$$\bar{\mathbf{u}}^t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{u}_i^t, \text{ and } \bar{\mathbf{x}}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i^{t+1}.$$

From (3.7) and (3.8), the $\bar{\mathbf{z}}$-update step can be simplified to the following unconstrained problem:

$$\min \quad g(N\bar{\mathbf{z}}) + \frac{\rho}{2} \sum_{i=1}^{N} \|\bar{\mathbf{z}} - \bar{\mathbf{q}}\|. \tag{3.9}$$

Substituting (3.8) for $\mathbf{z}_i^{t+1}$ into (3.6), the $\bar{\mathbf{u}}$-update step is

$$\mathbf{u}_i^{t+1} = \bar{\mathbf{u}}^t + \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{z}}^{t+1}. \tag{3.10}$$

The dual variables $\{\mathbf{u}_i^t\}_{i=1}^N$ are equal and can be replaced by a single variable $\mathbf{u}^t$. The ADMM steps are simplified as follows

$$\mathbf{x}_i^{t+1} := \arg\min_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{x}_i^t + \bar{\mathbf{x}}^t - \bar{\mathbf{z}}^t + u^t\|_2^2 \right) \tag{3.11}$$

$$\bar{\mathbf{z}}^{t+1} := \arg\min_{\bar{\mathbf{z}}} \left(g(N\bar{\mathbf{z}}) + \frac{N\rho}{2} \|\bar{\mathbf{z}} - u^t - \bar{\mathbf{x}}^{t+1}\|_2^2 \right) \tag{3.12}$$

$$\mathbf{u}^{t+1} := \mathbf{u}^t + \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{z}}^{t+1} \tag{3.13}$$

The original sharing problem (3.1) is decomposed into a three-step iterative optimization problem. The problem (3.11) can be solved in parallel for $i = 1, \ldots, N$. The step (3.12) solves the shared objective, and (3.13) is the dual variable update step.

3.3 ADMM-based Detection for UL MIMO CD-NOMA Systems

This section focuses on reformulating the relation between the input and observation signals of two different MIMO CD-NOMA system models: spatial multiplexing CD-NOMA (SMX-CD-NOMA) and spatial modulated CD-NOMA (SM-CD-NOMA). The reformulated models are exploited in resource-wise processing via the sharing ADMM algorithm. These models, along with the derivation of the ADMM steps, are described in the following:

3.3.1 System model for SMX-CD-NOMA

Consider a scenario where each UE is equipped with N_{t} transmitting antennas, and the BS is equipped with N_{r} receiving antennas as shown in Fig. 3.2. The total input data at each UE, $N_{\mathrm{t}}\log_2(M)$ bits, are divided into N_{t} parallel data streams. Each data stream is fed to the CD-NOMA encoder. Note that in the SMX-CD-NOMA system, all the UE antennas transmit data simultaneously. Fig. 3.3 illustrates the resource-wise processing at the BS. The received signal

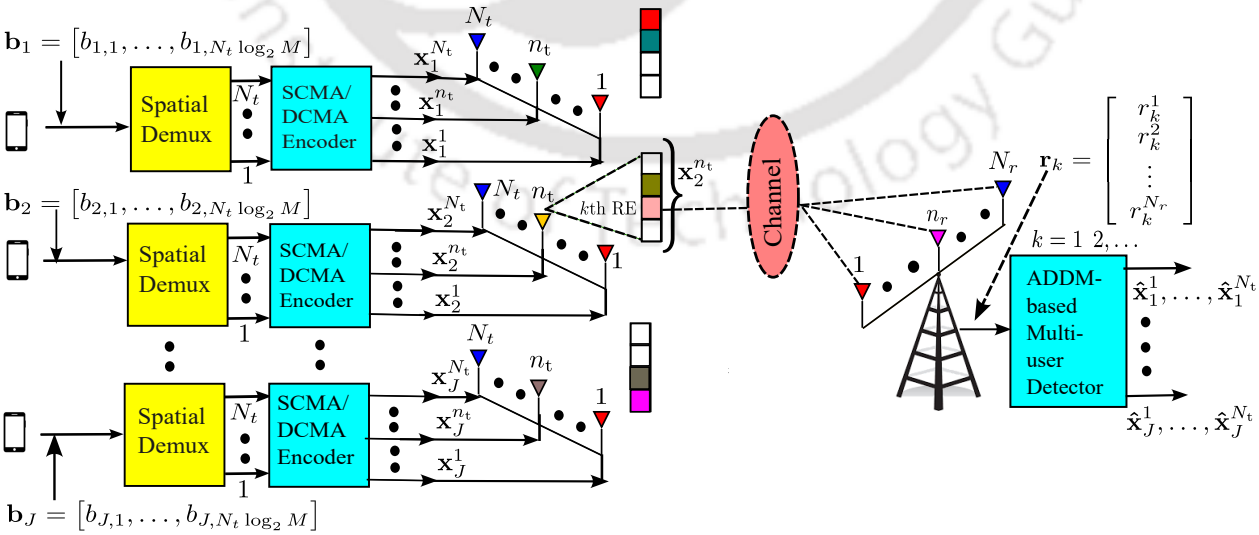


Fig. 3.2: SMX-CD-NOMA system model in the UL.

at the BS over the kth RE is given by

$$\mathbf{r}_k = \mathbf{H}_k \mathbf{x}_{\text{smx},k} + \mathbf{w}_k, \tag{3.14}$$

where

- $\mathbf{r}_k = [r_k^1 \dots r_k^{n_\text{r}} \dots r_k^{N_\text{r}}]^T$, $N_\text{r} \times 1$ observation vector.
- $\mathbf{w}_k \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_\text{r}})$, $N_\text{r} \times 1$ vector denotes the AWGN at the BS over kth RE.
- $\mathbf{x}_{\text{smx},k}$ is the $N_\text{u}N_\text{t} \times 1$ transmitted vector on kth RE. $\mathbf{x}_{\text{smx},k} = [\mathbf{x}_{1,k}^T \dots \mathbf{x}_{j,k}^T \dots \mathbf{x}_{N_\text{u},k}^T]^T$, and each $\mathbf{x}_{j,k} = [x_{j,k,1} \dots x_{j,k,n_\text{t}} \dots x_{j,k,N_\text{t}}]^T$ is $N_\text{t} \times 1$ vector corresponding to jth UE. Each $x_{j,k,n_\text{t}} \in \mathbf{x}_j^{n_\text{t}}$ is the symbol transmitted from n_tth antenna of jth UE over kth RE, where $\mathbf{x}_j^{n_\text{t}}$ is the codeword of jth UE transmitted from n_tth antenna. The set ϕ_k represents the set of UEs overlapping on kth RE given by

$$\phi_k = \{j : \mathbf{x}_j[k] \neq 0; 1 \leq j \leq J\}, \text{ and } |\phi_k| = N_\text{u},$$

where $N_\text{u} = d_\text{f}$ and $N_\text{u} = J$ for SCMA and DCMA, respectively.

The transmitted signal $\mathbf{x}_{\text{smx},k}$ can be rewritten to formulate the sharing-based detection problem in Section 3.3.3 i.e., $\mathbf{x}_{\text{smx},k} = \sum_{j=1}^{N_\text{u}} \mathbf{x}_{0j,k}$. The variable $\mathbf{x}_{0j,k} = [0 \dots 0 \;\; x_{j,k,1} \;\; x_{j,k,2} \dots x_{j,k,N_\text{t}} \;\; 0 \dots 0]^T$ is the $N_\text{u}N_\text{t} \times 1$ vector, and represents the jth overlapped UE on kth RE. The nonzero elements in $\mathbf{x}_{0j,k}$ are symbols transmitted from N_t antennas.

- $\mathbf{H}_k$ is the $N_\text{r} \times N_\text{u}N_\text{t}$ matrix given by

$$\mathbf{H}_k = [\mathbf{H}_{1,k} \dots \mathbf{H}_{j,k} \dots \mathbf{H}_{N_\text{u},k}]$$

where $\mathbf{H}_{j,k}$ represents the jth UE MIMO channel matrix of size $N_\text{r} \times N_\text{t}$.

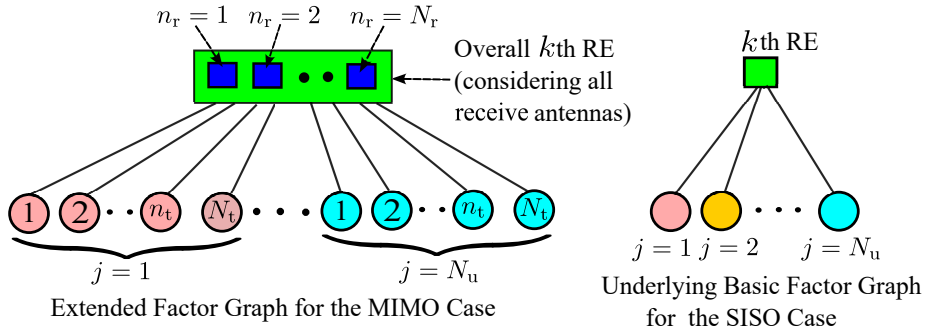


Fig. 3.3: Processing at kth RE of SMX-CD-NOMA system.

Example 1. *Consider $N_t = 2$, and according to the factor graph matrix for SCMA system*

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}. \tag{3.15}$$

The transmitted signal over the first RE in the SCMA system is

$$\mathbf{x}_{\text{smx},1} = \begin{bmatrix} \mathbf{x}_{1,1}^T & \mathbf{x}_{3,1}^T & \mathbf{x}_{5,1}^T \end{bmatrix}^T. \tag{3.16}$$

For DCMA, due to the dense structure of codebooks, all J UEs overlap on each RE [21]. The transmitted signal over the first RE is given by

$$\mathbf{x}_{\text{smx},1} = \begin{bmatrix} \mathbf{x}_{1,1}^T & \mathbf{x}_{2,1}^T & \dots & \mathbf{x}_{5,1}^T & \mathbf{x}_{6,1}^T \end{bmatrix}^T, \tag{3.17}$$

where each $\mathbf{x}_{j,1}$ is a 2×1 ($N_t \times 1$) vector given by

$$\mathbf{x}_{j,1} = \begin{bmatrix} x_{j,1,1} & x_{j,1,2} \end{bmatrix}^T.$$

The transmitted signal over each remaining RE ($\mathbf{x}_{\text{smx},k}, k = 2, \dots, K$) has a similar representation as in (3.16) and (3.17) for SCMA and DCMA, respectively.

For SCMA, $\mathbf{x}_{03,1}$ represents the 3rd UE's symbols as given by

$$\mathbf{x}_{03,1} = \begin{bmatrix} \underbrace{0 \;\; 0}_{\text{UE 1}} & \underbrace{x_{3,1,1} \;\; x_{3,1,2}}_{\text{UE 3}} & \underbrace{0 \;\; 0}_{\text{UE 5}} \end{bmatrix}^T. \tag{3.18}$$

For DCMA, $\mathbf{x}_{03,1}$ can be written as

$$\mathbf{x}_{03,1} = \begin{bmatrix} \underbrace{0 \;\; 0}_{\text{UE 1}} & \underbrace{0 \;\; 0}_{\text{UE 2}} & \underbrace{x_{3,1,1} \;\; x_{3,1,2}}_{\text{UE 3}} & \underbrace{0 \;\; 0}_{\text{UE 4}} & \underbrace{0 \;\; 0}_{\text{UE 5}} & \underbrace{0 \;\; 0}_{\text{UE 6}} \end{bmatrix}^T. \tag{3.19}$$

The variables $\{\mathbf{x}_{0j,k}\}_{k=1}^K$ are similarly represented as $\mathbf{x}_{03,1}$. The transmitted signal $\mathbf{x}_{\text{smx},k}$ can be rewritten for sharing problem formulation as, $\mathbf{x}_{\text{smx},k} = \sum_{j=1}^{3} \mathbf{x}_{0j,k}$ and $\mathbf{x}_{\text{smx},k} = \sum_{j=1}^{6} \mathbf{x}_{0j,k}$ for SCMA and DCMA, respectively. □

3.3.2 System Model for SM-CD-NOMA

Fig. 3.4 shows the system model for the SM-CD-NOMA system. Each UE is equipped with N_t transmitting antennas, in which only one antenna is active at any time, as shown in Fig. 3.4. The active antenna index of the jth UE is denoted by n_a^j. All other antennas remain silent in that particular time slot. Thus, the active antenna index is a spatial modulation symbol to transmit extra information bits. The information bit stream $\mathbf{b}_j$ of the jth UE, is split into

two parts [$\mathbf{b}_{ja}$,$\mathbf{b}_{jc}$] having $\log_2(N_t)$ and $\log_2(M)$ bits, respectively. Each UE transmits an overall $\log_2(N_t) + \log_2(M)$ bits from the active antenna. The observation vector over the kth RE at the

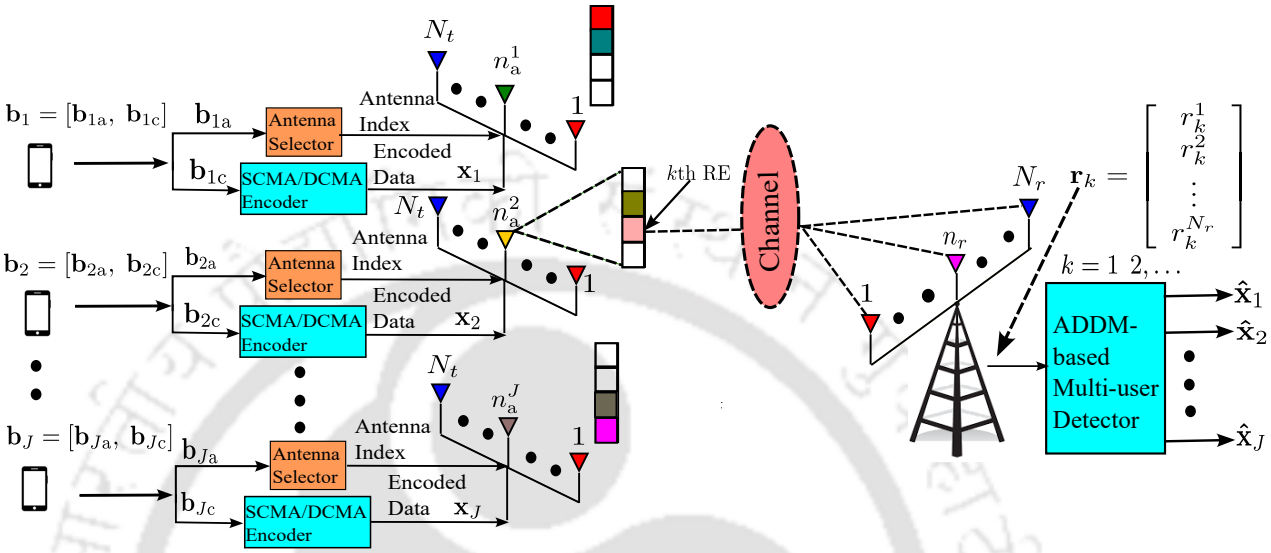


Fig. 3.4: SM-CD-NOMA system model in the UL.

BS is an $N_r \times 1$ vector given by

$$\mathbf{r}_k = \mathbf{H}_k \mathbf{x}_{sm,k} + \mathbf{w}_k, \tag{3.20}$$

where

- $\mathbf{r}_k$, $\mathbf{w}_k$, and $\mathbf{H}_k$ have the similar forms as in SMX-CD-NOMA for both DCMA and SCMA systems.
- $\mathbf{x}_{sm,k}$ is the $N_u N_t \times 1$ vector $\mathbf{x}_{sm,k} = [\mathbf{x}_{1,k}^T \ldots \mathbf{x}_{j,k}^T \ldots \mathbf{x}_{N_u,k}^T]^T$, and each $\mathbf{x}_{j,k} = [0 \ldots x_{j,k,n_a^j} \ldots 0]^T$ is $N_t \times 1$ vector corresponding to jth UE. The nonzero element x_{j,k,n_a^j} is the symbol transmitted from n_a^jth active antenna. The transmitted signal $\mathbf{x}_{sm,k}$ can be rewritten to formulate the sharing-based detection problem in Section 3.3.3 i.e., $\mathbf{x}_{sm,k} = \sum_{j=1}^{N_u} \mathbf{x}_{0j,k}^{n_a^j}$. The variable $\mathbf{x}_{0j,k}^{n_a^j} = [0 \ldots 0 \ldots \ x_{j,k,n_a^j} \ 0 \ldots \ 0 \ldots 0]^T$ represents the jth overlapped UE on kth RE. The nonzero element in $\mathbf{x}_{0j,k}^{n_a^j}$ is the symbol transmitted from n_a^jth active antenna.

Example 2. *Consider $N_t = 2$, and the factor graph matrix given in (3.15). Suppose the active antenna indices of the 6-users are $\{2,1,2,2,1,1\}$. The structure of transmit codewords $\mathbf{x}_{tx_{(N_t K \times J)}}$ from J UEs for SM-CD-NOMA is as follows [76]:*

$$\mathbf{x}_{tx_{(8\times6)}} = \begin{bmatrix} \mathbf{0}_{K\times1} & \mathbf{x}^1_{2_{K\times1}} & \mathbf{0}_{K\times1} & \mathbf{0}_{K\times1} & \mathbf{x}^1_{5_{K\times1}} & \mathbf{x}^1_{6_{K\times1}} \\ \mathbf{x}^2_{1_{K\times1}} & \mathbf{0}_{K\times1} & \mathbf{x}^2_{3_{K\times1}} & \mathbf{x}^2_{4_{K\times1}} & \mathbf{0}_{K\times1} & \mathbf{0}_{K\times1} \end{bmatrix}. \tag{3.21}$$

Each $\mathbf{x}_{j_{K\times 1}}^{n_{\mathrm{a}}^j}$ *in* $\mathbf{x}_{\mathrm{tx}}$ *represents the transmitted codeword from jth UE, and each column of* $\mathbf{x}_{tx}$ *carries information about active transmit antenna* (n_{a}^j) *as well as the codeword. In (3.21),* $\mathbf{0}_{K\times 1}$ *indicates zero power transmitted by a deactivated antenna. To facilitate the resource-wise processing at the receiver via ADMM, the transmitted signal over the first RE according to (3.15) in the SCMA system is modeled as*

$$\mathbf{x}_{\mathrm{sm},1} = \begin{bmatrix} \mathbf{x}_{1,1}^T & \mathbf{x}_{3,1}^T & \mathbf{x}_{5,1}^T \end{bmatrix}^T. \tag{3.22}$$

The transmitted signal over the first RE in the DCMA system is modeled as

$$\mathbf{x}_{\mathrm{sm},1} = \begin{bmatrix} \mathbf{x}_{1,1}^T & \mathbf{x}_{2,1}^T & \dots & \mathbf{x}_{5,1}^T & \mathbf{x}_{6,1}^T \end{bmatrix}^T, \tag{3.23}$$

where $\mathbf{x}_{1,1} = [0 \;\; x_{1,1,2}]^T$, $\mathbf{x}_{2,1} = [x_{2,1,1} \;\; 0]^T$, $\mathbf{x}_{3,1} = [0 \;\; x_{3,1,2}]^T$, $\mathbf{x}_{4,1} = [0 \;\; x_{4,1,2}]^T$, $\mathbf{x}_{5,1} = [x_{5,1,1} \;\; 0]^T$, *and* $\mathbf{x}_{6,1} = [x_{6,1,1} \;\; 0]^T$ *according to the considered active antenna indices of 6-users. The transmitted signals over the remaining REs* ($\mathbf{x}_{\mathrm{sm},k}, k = 2, \dots, K$) *have similar representations as (3.22) and (3.23) for SCMA and DCMA, respectively. For SCMA,* $\mathbf{x}_{03,1}^2$ *represents the symbol transmitted from 2nd active antenna of 3rd UE on 1st RE.* $\mathbf{x}_{03,1}^2$ *can be written as*

$$\mathbf{x}_{03,1}^2 = \Big[\underbrace{0 \;\; 0}_{\text{UE 1}} \;\; \underbrace{0 \;\; x_{3,1,2}}_{\text{UE 3}} \;\; \underbrace{0 \;\; 0}_{\text{UE 5}} \Big]^T. \tag{3.24}$$

For DCMA, $\mathbf{x}_{03,1}^2$ *can be written as*

$$\mathbf{x}_{03,1}^2 = \Big[\underbrace{0 \;\; 0}_{\text{UE 1}} \;\; \underbrace{0 \;\; 0}_{\text{UE 2}} \;\; \underbrace{0 \;\; x_{3,1,2}}_{\text{UE 3}} \;\; \underbrace{0 \;\; 0}_{\text{UE 4}} \;\; \underbrace{0 \;\; 0}_{\text{UE 5}} \;\; \underbrace{0 \;\; 0}_{\text{UE 6}} \Big]^T. \tag{3.25}$$

The variables $\{\mathbf{x}_{0j}^{n_{\mathrm{a}}^j}\}_{k=1}^K$ *are similarly represented as* $\mathbf{x}_{03,1}^2$. *The transmitted signal* $\mathbf{x}_{\mathrm{sm},k}$ *can be rewritten for sharing problem formulation as,* $\mathbf{x}_{\mathrm{sm},k} = \sum_{j=1}^{3} \mathbf{x}_{0j,k}^{n_{\mathrm{a}}^j}$ *and* $\mathbf{x}_{\mathrm{sm},k} = \sum_{j=1}^{6} \mathbf{x}_{0j,k}^{n_{\mathrm{a}}^j}$ *for SCMA and DCMA, respectively.* □

3.3.3 Large-scale MIMO UL CD-NOMA Detection via ADMM

This section discusses the problem formulation and algorithm design of ADMM-based MUD for the CD-NOMA system. The two different MIMO techniques, popularly known as spatial multiplexed (SMX) and spatial modulated (SM), are considered for the detection. The formulated MIMO-CD-NOMA system models, (3.14) and (3.20) are similar to the conventional MIMO system model. The ML detection problem for the MIMO CD-NOMA system in (3.14)

is given by

$$\min_{\mathbf{x}_{\text{smx,k}} \in \mathcal{X}_k^{N_\text{u}N_\text{t}}} \quad \|\mathbf{r}_k - \mathbf{H}_k \mathbf{x}_{\text{smx,k}}\|^2, \tag{3.26}$$

where $\mathcal{X}_k^{N_\text{u}N_\text{t}}$ denotes the multi-user signal constellation on kth RE, and it consists of symbols from N_u-UEs in concatenated form. However, solving ML detection problems, in general, is NP-hard. The exhaustive search method can be used to solve the ML detection problem. However, the exhaustive search is exponentially complex $O(M^{N_\text{u}N_\text{t}})$ [64], and thus, it is not feasible for large scale MIMO CD-NOMA systems. By relaxing the ML detection problem in (3.26), a sharing optimization problem is formed. Further, the sharing problem can be solved by using a distribution optimization framework. In this work, the ADMM algorithm is exploited to solve the formulated detection problem.

3.3.3.1 SMX-CD-NOMA

The ML detection problem in (3.26) can be converted into a sharing problem. Further, this problem can be solved in a distributed manner via the ADMM approach, as discussed in Section 3.2.2. ADMM allows parallel processing in MUD problems, guaranteeing a minimal computational time at the receiver, especially for large-scale CD-NOMA systems. The transmitted signal $\mathbf{x}_{\text{smx,k}}$ is given by $\mathbf{x}_{\text{smx,k}} = \sum_{j=1}^{N_\text{u}} \mathbf{x}_{0j,k}$. The problem (3.26) can be rewritten as

$$\begin{aligned} \min_{\mathbf{x}_{0j}} \quad & \|\mathbf{r}_k - \mathbf{H}_k \left(\sum_{j=1}^{N_\text{u}} \mathbf{x}_{0j,k} \right) \|^2 \\ \text{s.t} \quad & \mathbf{x}_{0j,k} \in \mathcal{X}_k^{N_\text{u}N_\text{t}} \quad j = 1, \ldots, J. \end{aligned} \tag{3.27}$$

Here, for the channel matrix $\mathbf{H}_k^{N_\text{r} \times N_\text{u}N_\text{t}}$, we consider $N_\text{r} > N_\text{u}N_\text{t}$. Each entry's real and imaginary parts in $\mathbf{x}_{0j,k}$ are restricted by set constraints. The set constraint must be converted into an interval constraint to convert (3.27) into a sharing problem. The BCR is used to relax the set constraints of each entry in the codeword [65]. Hence, each entry's real and imaginary parts in the jth UE codeword belong to $[-\alpha_j, \alpha_j]$ and $[-\beta_j, \beta_j]$. After relaxation, each element in the jth UE codebook is defined as $\tilde{\mathcal{X}}_j = \{x_j = x_{jR} + ix_{jI} | x_{jR} \in [-\alpha_j, \alpha_j], x_{jI} \in [-\beta_j, \beta_j]\}$, $\alpha_j =$

Algorithm 2 SMX CD-NOMA detection via sharing ADMM

Input: N_0, E_s and initialize $\{\mathbf{z}_{0j,k}\}_{j=1}^{N_u}, \bar{\mathbf{x}}_{0,k}, \mathbf{u}, \bar{\mathbf{z}}_{0,k}$ with zero vectors.
Output: $\bar{\mathbf{x}}_0^{(T)}$.
From (3.34) and (3.35), obtain optimal solutions (3.44) and (3.45) respectively.
Preprocessing
$\rho = \epsilon(N_0/E_s)$, where $\epsilon \in [0, 1]$.
Compute $(\mathbf{H}_k^H \mathbf{H}_k N_u + \rho \mathbf{I})^{-1}$.
Compute $\mathbf{H}_k^H \mathbf{r}_k$.
for $t = 1, 2, , \ldots T$ **do**
    Step:1 Update $\{\mathbf{z}_{0j,k}^{t+1}\}_{j=1}^{J}$ in parallel via (3.44) and compute $\bar{\mathbf{x}_1}$,
    Step:2
    Update $\bar{\mathbf{x}}_{0,k}^{t+1}$ via (3.45),
    Update $\mathbf{u}^{t+1}$ via (3.36).
end for
Compute $\hat{\mathbf{x}}_{\text{smx},k} = J\bar{\mathbf{x}}_{0,k}^{(T)}$ (for DCMA), and $\hat{\mathbf{x}}_{\text{smx},k} = d_f \bar{\mathbf{x}}_{0,k}^{(T)}$ (for SCMA).
After extracting $\hat{\mathbf{x}}_{j,N_t}$, apply the MED rule (3.46).

$\max |\mathbb{R}(\mathcal{X}_j)|, \beta_j = \max |\mathcal{I}(\mathcal{X}_j)|$. The highly complex MIMO-CD-NOMA ML detection problem in (3.27) is now ready to be converted into a non-convex distributed optimization problem. The penalty function introduced in (3.28) ensures the transmit codewords are from finite energy codebooks. Each variable $\mathbf{x}_{0j,k}$ is the choice of penalty function (local cost) corresponding to jth UE. The sharing problem ensures that each user adjusts its variable to minimize the individual penalty function as well as the shared objective function. The detector of SMX-CD-NOMA needs to detect the signals transmitted from N_t antennas of N_u UEs. Recall from Section 3.3.1, that the SMX codeword is given by, $\mathbf{x}_{\text{smx},k} = \sum_{j=1}^{N_u} \mathbf{x}_{0j,k}$. The resource-wise ADMM processing provides the soft spatial multiplexed transmitted vector $\hat{\mathbf{x}}_{\text{smx},k}$, for $k = 1, \ldots, K$. Then, the transmitted codeword indices are detected via the MED rule. The resource-wise ADMM processing is applied as follows

$$\begin{aligned} \min_{\mathbf{x}_{0j,k}} \quad & \frac{1}{2}\|\mathbf{r}_k - \mathbf{H}_k \left(\sum_{j=1}^{N_u} \mathbf{x}_{0j,k} \right)\|_2^2 + \sum_{j=1}^{N_u} \frac{\gamma_j}{2} \|\mathbf{x}_{0j,k}\|_2^2 \\ \text{s.t.} \quad & \mathbf{x}_{0j,k} \in \tilde{\mathcal{X}}_{j,\text{smx}}^{N_u N_t}, \end{aligned} \tag{3.28}$$

where $\gamma_j \geq 0$ is the penalty parameter. The problem in (3.28) is similar to the sharing ADMM problem in (3.1). Proceeding similarly as in Section 3.2.2, (3.28) can be written as

$$\begin{aligned} \min_{\mathbf{x}_{0j,k},\mathbf{z}_{0j,k}} \quad & \frac{1}{2}\|\mathbf{r}_k - \mathbf{H}_k\left(\sum_{j=1}^{N\mathrm{u}} \mathbf{x}_{0j,k}\right)\|^2 + \sum_{j=1}^{N\mathrm{u}} \frac{\gamma_j}{2}\|\mathbf{z}_{0j,k}\|_2^2 \\ \text{s.t} \quad & \mathbf{x}_{0j,k} = \mathbf{z}_{0j,k}, \quad \mathbf{x}_{0j,k} \in \tilde{\mathcal{X}}_k^{N_\mathrm{u}N_\mathrm{t}}, \quad \forall j,k \end{aligned} \tag{3.29}$$

The augmented Lagrangian function for (3.29) is:

$$\begin{aligned} \mathcal{L}\left(\{\mathbf{x}_{0j,k},\mathbf{z}_{0j,k},\mathbf{y}_{j,k}\}_{j=1}^{J}\right) = \frac{1}{2}\|\mathbf{r}_k - \mathbf{H}_k\left(\sum_{j=1}^{N_u} \mathbf{x}_{0j,k}\right)\|_2^2 + \sum_{j=1}^{N\mathrm{u}} \frac{\gamma_j}{2}\|\mathbf{z}_{0j,k}\|_2^2 + \sum_{j=1}^{N\mathrm{u}} \mathrm{Re}\langle \mathbf{x}_{0j,k} - \mathbf{z}_{0j,k}, \mathbf{y}_{j,k}\rangle \\ + \frac{\rho}{2}\sum_{j=1}^{N\mathrm{u}} \|\mathbf{x}_{0j,k} - \mathbf{z}_{0j,k}\|_2^2, \end{aligned} \tag{3.30}$$

where $\mathbf{x}_{0j,k}$ and $\mathbf{z}_{0j,k}$ are the primal variables, $\mathbf{y}_{j,k} \in \mathbb{C}^{N_\mathrm{u}N_\mathrm{t}}$ is Lagrangian multiplier of the jth UE and $\rho > 0$ is called the augmented Lagrangian penalty parameter.

$$\mathbf{z}_{0j,k}^{t+1} := \underset{\mathbf{z}_{0j,k}\in\tilde{\mathcal{X}}_{j,\mathrm{smx}}^{N_\mathrm{u}N_\mathrm{t}}}{\arg\min} \frac{\gamma_j}{2}\|\mathbf{z}_{0j,k}\|_2^2 - \langle \mathbf{y}_{j,k}^t, \mathbf{z}_{0j,k}\rangle + \frac{\rho}{2}\|\mathbf{x}_{0j,k}^t - \mathbf{z}_{0j,k}\|_2^2 \tag{3.31}$$

$$\mathbf{x}_{0j,k}^{t+1} := \underset{\mathbf{x}_{0j,k}}{\arg\min} \frac{1}{2}\|\mathbf{r}_k - \mathbf{H}_k\left(\sum_{j=1}^{N_\mathrm{u}} \mathbf{x}_{0j,k}\right)\|_2^2 + \sum_{j=1}^{N_\mathrm{u}} \langle \mathbf{y}_{j,k}^t, \mathbf{x}_{0j,k}\rangle + \frac{\rho}{2}\sum_{j=1}^{N_\mathrm{u}} \|\mathbf{z}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}\|_2^2 \tag{3.32}$$

$$\mathbf{y}_{j,k}^{t+1} := \mathbf{y}_{j,k}^t + \rho\left(\mathbf{z}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t+1}\right). \tag{3.33}$$

Letting $\mathbf{u}_{j,k} = \frac{\mathbf{y}_{j,k}}{\rho}$ (called a dual variable), the scaled form of ADMM steps are as follows [70] :

$$\mathbf{z}_{0j,k}^{t+1} := \underset{\mathbf{z}_{0j,k}\in\tilde{\mathcal{X}}_{j,\mathrm{smx}}^{N_\mathrm{u}N_\mathrm{t}}}{\arg\min} \frac{\gamma_j}{2}\|\mathbf{z}_{0j,k}\|_2^2 + \frac{\rho}{2}\|\mathbf{z}_{0j,k} - \mathbf{x}_{0j,k}^t + \mathbf{u}_{j,k}^t\|_2^2 \tag{3.34}$$

$$\mathbf{x}_{0j,k}^{t+1} := \underset{\mathbf{x}_{0j,k}}{\arg\min} \frac{1}{2}\|\mathbf{r}_k - \mathbf{H}_k\left(\sum_{j=1}^{N_\mathrm{u}} \mathbf{x}_{0j,k}\right)\|_2^2 + \frac{\rho}{2}\sum_{j=1}^{N\mathrm{u}} \|\mathbf{z}_{0j,k}^{t+1} - \mathbf{x}_{0j,k} + \mathbf{u}_{j,k}^t\|_2^2 \tag{3.35}$$

$$\mathbf{u}_{j,k}^{t+1} := \mathbf{u}_{j,k}^t + \left(\mathbf{z}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t+1}\right). \tag{3.36}$$

The equations (3.34) and (3.36) can be solved independently in parallel for $j = 1,\ldots,N_\mathrm{u}$ and $k = 1,\ldots,K$. The problem in (3.35) can be simplified according to (3.7), as follows

$$\min \quad \|\mathbf{r}_k - N_\mathrm{u}\mathbf{H}_k\bar{\mathbf{x}}_{0,k}\|_2^2 + \frac{\rho}{2}\sum_{j=1}^{N_\mathrm{u}} \|\mathbf{q}_j - \mathbf{x}_{0j,k}\|_2^2 \tag{3.37}$$

$$\text{s.t} \quad \bar{\mathbf{x}}_{0,k} = \frac{1}{N_\text{u}} \sum_{j=1}^{N_\text{u}} \mathbf{x}_{0j,k},$$

where $\mathbf{q}_j = \mathbf{u}_{j,k}^t + \mathbf{z}_{0j,k}^{t+1}$, $\bar{\mathbf{x}}_{0,k} \in \mathbb{C}^{N_\text{u}N_\text{t}}$ be a fixed variable. The problem can be solved by using the Lagrangian method of optimization, and the solution as follows

$$\mathbf{x}_{0j,k} = \mathbf{q}_j + \bar{\mathbf{x}}_{0,k} - \bar{\mathbf{q}}, \tag{3.38}$$

where $\bar{\mathbf{q}} = \frac{1}{N_\text{u}} \sum_{j=1}^{N_\text{u}} \mathbf{q}_j = \bar{\mathbf{u}} + \bar{\mathbf{z}}_{0,k}$ and $\bar{\mathbf{z}}_{0,k} = \frac{1}{N_\text{u}} \sum_{j=1}^{N_\text{u}} \mathbf{z}_{0j,k}$, $\bar{\mathbf{u}} = \frac{1}{N_\text{u}} \sum_{j=1}^{N_\text{u}} \mathbf{u}_j$. From (3.37) and (3.38), the $\bar{\mathbf{x}}_{0,k}$-update can be computed by using unconstrained optimization problem

$$\min \quad \|\mathbf{r}_k - N_\text{u}\mathbf{H}_k\bar{\mathbf{x}}_{0,k}\|_2^2 + \frac{\rho}{2} \sum_{j=1}^{N_\text{u}} \|\bar{\mathbf{x}}_{0,k} - \bar{\mathbf{q}}\|_2^2 \tag{3.39}$$

Substituting (3.38) in (3.36) gives

$$\mathbf{u}_{j,k}^{t+1} = \bar{\mathbf{u}} + \bar{\mathbf{z}}_{0,k} - \bar{\mathbf{x}}_{0,k} \tag{3.40}$$

The dual variables $\{\mathbf{u}_{j,k}\}_{j=1}^{N_\text{u}}$ are equal and can be replaced with a single variable $\mathbf{u}$. The ADMM steps are simplified as follows

$$\mathbf{z}_{0j,k}^{t+1} := \underset{\mathbf{z}_{0j} \in \tilde{\mathcal{X}}_\text{mu}^{N_\text{u}N_\text{t}}}{\arg\min} \frac{\gamma_j}{2}\|\mathbf{z}_{0j}\|_2^2 + \frac{\rho}{2}\|\mathbf{z}_{0j,k} - \mathbf{z}_{0j,k}^t - \bar{\mathbf{x}}_{0,k}^t + \mathbf{u}^t + \bar{\mathbf{z}}_{0,k}^t\|_2^2 \tag{3.41}$$

$$\bar{\mathbf{x}}_{0,k}^{t+1} := \underset{\bar{\mathbf{x}}_{0,k}}{\arg\min} \frac{1}{2}\|\mathbf{r}_k - N_\text{u}\mathbf{H}_k\bar{\mathbf{x}}_{0,k}\|_2^2 + \frac{\rho N_\text{u}}{2}\|\bar{\mathbf{x}}_{0,k} - \bar{\mathbf{z}}_{0,k}^{t+1} - \mathbf{u}^t\|_2^2 \tag{3.42}$$

$$\mathbf{u}^{t+1} := \mathbf{u}^t + \bar{\mathbf{z}}_{0,k}^{t+1} - \bar{\mathbf{x}}_{0,k}^{t+1} \tag{3.43}$$

(3.41) can be computed in parallel for $j = 1, \ldots, J$. The solutions for (3.41) and (3.42) are as follows:

$$\mathbf{z}_{0j,k}^{t+1} = \prod_{[-\alpha_j,\alpha_j]\&[-\beta_j,\beta_j]} \frac{\rho}{(\rho + \gamma_j)}(\mathbf{z}_{0j,k}^t + \bar{\mathbf{x}}_{0,k}^t - \mathbf{u}^t - \bar{\mathbf{z}}_{0,k}^t) \tag{3.44}$$

$$\bar{\mathbf{x}}_{0,k}^{t+1} = (\mathbf{H}_k^H\mathbf{H}_k N_\text{u} + \rho\mathbf{I})^{-1}(\mathbf{H}_k^H\mathbf{r}_k + \rho(\bar{\mathbf{z}}_{0,k}^{t+1} + \mathbf{u}^t)) \tag{3.45}$$

where $\mathbf{I}$ is an identity matrix of size $N_\text{u}N_\text{t} \times N_\text{u}N_\text{t}$. From the definitions of $\bar{\mathbf{x}}_{0,k}$ in (3.7) and $\mathbf{x}_{\text{smx},k}$ in (3.26), the soft decision SMX codeword corresponding to the kth RE is given by $\hat{\mathbf{x}}_{\text{smx},k} = N_\text{u}\bar{\mathbf{x}}_{0,k}$.

The sparse vector $\hat{\mathbf{x}}_{0j,k}$ corresponding to jth UE can be extracted from $\hat{\mathbf{x}}_{\text{smx},k}$. Let $\tilde{\mathbf{x}}_{j,k}$ be a $N_\text{t} \times 1$ vector formed after removing the zeros from sparse vector $\hat{\mathbf{x}}_{0j,k}$. Let $\hat{\mathbf{x}}_j^{n_\text{t}}$ be the ADMM output codeword of jth UE corresponding to n_tth transmit antenna. $\hat{\mathbf{x}}_j^{n_\text{t}}$ is formed after resource-wise processing is finished. The MED rule is applied to detect the transmitted codeword index corresponding to each UE and transmit antenna as follows:

$$\hat{p}_j^{n_\text{t}} = \arg\min_{p} ||\hat{\mathbf{x}}_j^{n_\text{t}} - \mathbf{x}_{j,p}||, \, j = 1, \ldots, J, \;\; n_\text{t} = 1, \ldots, N_\text{t}. \tag{3.46}$$

where $\mathbf{x}_{j,p}$ represents the pth codeword of the jth UE codebook $\mathcal{X}_j^{K\times M}$. The detailed ADMM-based detection procedure is given in **Algorithm** 2.

3.3.3.2 SM-CD-NOMA

The proposed ADMM algorithm can be further extended to SM-CD-NOMA systems. The reformulation system model (or) resource-wise processing allows us to consider the SM-CD-NOMA as the special case of the SMX-CD-NOMA. In SM, the modulation happens in both the signal and spatial domains, as discussed in Section 3.3.2. The receiver needs to estimate these two quantities. The variable $\mathbf{x}_{\text{sm},k}$ from Section 3.3.2, is given by $\mathbf{x}_{\text{sm},k} = \sum_{j=1}^{N_\text{u}} \mathbf{x}_{0j,k}^{n_\text{a}^j}$. The resource-wise ADMM processing provides the soft decision output spatial modulated transmitted vector $\hat{\mathbf{x}}_{\text{sm},k}$, for $k = 1, \ldots, K$. Then, the ℓ_1-norm rule is applied to detect the antenna index, and the MED rule is applied to detect the transmitted codeword index. The sharing problem of the SM-CD-NOMA system is similar to the problem in (3.28). Following the steps (3.34), (3.35), and (3.36), the sharing ADMM problem of SM-CD-NOMA can be solved. The solutions obtained for the sharing ADMM problem are similar to (3.44) and (3.45) and are repeated here for clarity:

$$\mathbf{z}_{0j,k}^{n_\text{a}^j,t+1} = \prod_{[-\alpha_j,\alpha_j]\&[-\beta_j,\beta_j]} \frac{\rho}{(\rho+\gamma_j)} (\mathbf{z}_{0j,k}^{n_\text{a}^j,t} + \bar{\mathbf{x}}_{0,\text{sm,k}}^t - \mathbf{u}^t - \bar{\mathbf{z}}_{0,\text{sm},k}^t) \tag{3.47}$$

$$\bar{\mathbf{x}}_{0,\text{sm},k}^{t+1} = (\mathbf{H}_k^H \mathbf{H}_k N_\text{u} + \rho \mathbf{I})^{-1} (\mathbf{H}_k^H \mathbf{r}_k + \rho(\bar{\mathbf{z}}_{0,\text{sm},k}^{t+1} + \mathbf{u}^t)). \tag{3.48}$$

The definitions of $\bar{\mathbf{x}}_{0,\text{sm},k}$ and $\bar{\mathbf{z}}_{0,\text{sm},k}$ are similar to SMX ADMM processing. The ADMM output SM codeword corresponding to kth RE is $\hat{\mathbf{x}}_{\text{sm},k} = N_\text{u} \bar{\mathbf{x}}_{0,\text{sm},k}$. After resource-wise processing, the

structure of the transmitted codewords for J UEs can be retrieved. In the presence of noise, the detected active transmit antenna index is given by

$$\hat{n}_{\mathrm{a}}^{j} = \arg\max_{n_{\mathrm{t}}}(\|\hat{\mathbf{x}}_{j}^{\mathrm{n_t}}\|_1), \quad \forall j \tag{3.49}$$

where $\|(\cdot)\|_1$ denotes the ℓ_1-norm of $(\cdot)$. Let $\hat{\mathbf{x}}_{j}^{\hat{n}_{\mathrm{a}}^{j}}$ be the ADMM output codeword corresponds to active transmit antenna $\hat{n}_{\mathrm{a}}^{j}$ of jth UE. The MED rule is applied to detect the codeword index corresponding to each UE

$$\hat{p}_j = \arg\min_{p} \|\hat{\mathbf{x}}_{j}^{\hat{n}_{a}^{j}} - \mathbf{x}_{j,p}\|, \quad \forall j. \tag{3.50}$$

Similar steps of **Algorithm** 2 are followed in the detection of the SM-CD-NOMA system.

3.3.3.3 Convergence of Algorithm 2

Theorem 3.1. *Suppose $\gamma_j > \rho, j = 1, \ldots, J$ and $\rho > \lambda_{\max}(\mathbf{H}_k^H \mathbf{H}_k)$ are satisfied. Then the sequence $\{\mathbf{x}_{0j,k}, \mathbf{z}_{0j,k}, \mathbf{y}_j\}_{j=1}^{J}$ generated by ADMM algorithm is convergent as given below*

$$\lim_{t\to\infty} \mathbf{x}_{0j,k}^{t} = \mathbf{x}_{0j,k}^{\star}, \quad \lim_{t\to\infty} \mathbf{z}_{0j,k}^{t} = \mathbf{z}_{0j,k}^{\star}, \quad \lim_{t\to\infty} \mathbf{y}_{j,k}^{t} = \mathbf{y}_{j,k}^{\star}, \quad \forall j = 1, \ldots, J, k = 1, \ldots, K. \tag{3.51}$$

Proof. Provided in Appendix A. □

Remark 1: *The above theorem indicates that the proposed sharing ADMM detection algorithm is guaranteed to converge to some stationary point under mild conditions on the selection of the ADMM penalty parameters ($\rho, \gamma_j > \rho, j = 1, \ldots, J$). The parameters are selected based on the given channel matrix $\mathbf{H}_k$ [65].*

3.4 Computational Complexity

The number of calculations in the proposed CD-NOMA detection via ADMM consists of two parts. Part-1 is iteration-independent (Pre-processing) steps, described in lines 6 and 7 in **Algorithm** 2, and Part-2 is iteration-dependent steps, from lines 9 to 12. The calculations in Part-1 are performed only once, i.e., before the ADMM iterations. In the SMX CD-NOMA system, Part-1 contains three steps of calculations such as $\mathbf{H}_k^H \mathbf{H}_k$, $(\mathbf{H}_k^H \mathbf{H}_k + \rho \mathbf{I})^{-1}$, and $\mathbf{H}_k^H \mathbf{r}$. The size of $\mathbf{H}_k$ is $N_{\mathrm{r}} \times N_{\mathrm{t}} d_{\mathrm{f}}$ and $N_{\mathrm{r}} \times N_{\mathrm{t}} J$ for SCMA and DCMA, respectively. Further, the numbers of FLOPs required to perform these three steps over the SCMA system are $(N_{\mathrm{r}})(N_{\mathrm{t}} d_{\mathrm{f}})^2, (N_{\mathrm{t}} d_{\mathrm{f}})^3$

Table 3.1: Computational complexity of different detectors for SMX-CD-NOMA systems

Detector	SCMA	DCMA
MPA	$(Kd_f^2 M^{(N_t d_f)} N_r + N d_f M d_v)T$	NA
EPA	$(KN_r d_f(2d_f + 2N_t M + 3) + 3KN_t d_f M)T$	NA
GAMPA	$(N_t d_f M + N_t d_f(2M + 1) + 5)T$	NA
GSD	NA	$(N_r K)J^2 + J^3/3 + (J^3 + (N_r K)J^2 + JN_r K) + J^3 + 2J^2 + E_{\text{FLOPS}}$
LMMSE	$(N_r)(N_t d_f)^2 + (N_t d_f)^3 + (N_r)(N_t d_f)$	$(N_r)(N_t J)^2 + (N_t J)^3 + (N_r)(N_t J)$
OAMP	$T((N_r)^2(N_t d_f) + (N_r)^3)$	$T((N_r)^2(N_t J) + (N_r)^3)$
Sharing ADMM	$(N_r)(N_t d_f)^2 + (N_t d_f)^3 + (N_r)(N_t d_f) + T(J(N_t d_f) + (N_t d_f)^2 + N_t d_f)$	$(N_r)(N_t J)^2 + (N_t J)^3 + (N_r)(N_t J) + T(J(N_t J) + (N_t J)^2 + N_t J)$

and $(N_r)(N_t d_f)$. The numbers of FLOPs required to perform the same steps over the DCMA system are $(N_r)(N_t J)^2$, $(N_t J)^3$, and $(N_r)(N_t J)$. The calculations in Part-2 need to be repeated in every iteration and contain mainly two steps. For SCMA system, these steps involve scalar multiplication of $N_t d_f \times 1$ vector for J UEs parallelly in (3.44), multiplication of $N_t d_f \times N_t d_f$ matrix with $N_t d_f \times 1$ and scalar multiplication with $N_t d_f \times 1$ vector in (3.45), in total $J(N_t d_f) + (N_t d_f)^2 + N_t d_f$ FLOPs approximately. For DCMA system, $(J(N_t J) + (N_t J)^2 + N_t J)$ FLOPs are required. The approximate total computational cost to implement the ADMM-based detector over the SMX-SCMA system and SMX-DCMA system are $(N_r)(N_t d_f)^2 + (N_t d_f)^3 + (N_r)(N_t d_f) + T(J(N_t d_f) + (N_t d_f)^2 + N_t d_f)$ and $(N_r)(N_t J)^2 + (N_t J)^3 + (N_r)(N_t J) + T(J(N_t J) + (N_t J)^2 + N_t J)$, respectively. Table 3.1 compares the total complexity of the sharing ADMM-based detection problem with other known detection schemes. In Table 3.1, NA stands for 'Not applicable'.

MPA is widely used for SCMA systems, and its complexity is exponential with M, N_t and d_f, as mentioned in Table 3.1 [63]. The computational burden will grow further for large-scale SCMA systems. Further, due to exponential complexity, MPA is not feasible in large-scale MIMO systems. The alternate low-complexity detectors available for MIMO-SCMA systems are EPA [60] and GAMPA [62]. The complexity of EPA and GAMPA is linear with M, N_t, and d_f. These detectors mainly solve the detection problem of sparse FG-based CD-NOMA systems.

A single tree search (STS) based soft-in soft-out (SISO) GSD [77] is applied to detect the signals of a spreading sequence-based DCMA system, also known as an overloaded CDMA system [21]. GSD performs key pre-processing steps to convert the rank deficient system into full rank one [78]. These steps include $\mathbf{H}^H\mathbf{H}$ in ($\mathbf{Q} = \mathbf{H}^H\mathbf{H} + \lambda\mathbf{I}_J$), Cholesky decomposition $\mathbf{Q} = \mathbf{D}^H\mathbf{D}$, and $(\mathbf{H}\mathbf{D}^{-1})^H\mathbf{r}$. These steps require $(N_r K)J^2$, $J^3/3$, and $(J^3 + (N_r K)J^2 + JN_r K)$ FLOPs, respectively. The rank-deficient linear system is thus converted into a full-rank one,

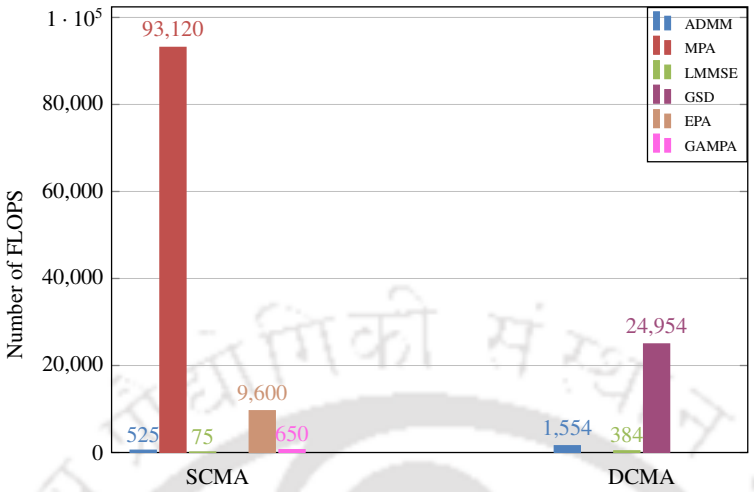


Fig. 3.5: Computational complexity comparison of various detectors.

and standard SD can be readily applied. The SD performs pre-processing steps, including **QR** decomposition and $\mathbf{Q}^H\mathbf{r}$. Further, the numbers of FLOPs to perform these steps are J^3 and $2J^2$, respectively. The major complexity of the SD lies in the tree search algorithm [79]. The expected complexity of SD is $E_{\text{FLOPS}} = \sum_{j=1}^{Nu} f_p(j)N_j$, where N_j is the average number of nodes visited in level-j of the tree and $f_p(j)$ is the number of FLOPs needed in level-j. It is given in [79], $f_p(j) = 2j + 11$, and N_j is roughly cubic in the number J of unknowns to be solved. The computational complexity of OAMP is primarily due to the inverse operation in the LMMSE step [67]. Fig. 3.5 illustrates the computational complexity comparison of various detection schemes for different CD-NOMA systems with $N_r = 4$, $K = 4$, $J = 6$, $d_v = 2$, and $M = 4$. It can be observed that the complexity of the ADMM-based detector is nearly equal to that of the LMMSE and OAMP detectors.

3.5 Simulations and Discussions

The simulations are performed with different values of λ, M, N_r, and N_t for the various MIMO-CD-NOMA systems mentioned in Section 3.3. $J = 6, K = 4$ and $J = 8, K = 4$ are considered for 150% and 200% overloaded CD-NOMA systems, respectively. The performance of the proposed ADMM algorithm is compared with the baseline algorithms, namely MPA, GAMPA, and EPA detectors for the SCMA system. Further, the ADMM detector's performance is compared with its counterpart, i.e., the LMMSE detector, for both SCMA and DCMA systems. The existing GSD detector for DCMA systems is also compared with the ADMM

detector. The codebooks designed in [80] and [21] are considered for SCMA and DCMA systems, respectively. Table 4.1 further details the simulation parameters. The sample simulation codes for this work are available at https://github.com/vikas2020-del/ADMM-based-detector-for-NOMA.

Table 3.2: Simulation parameters.

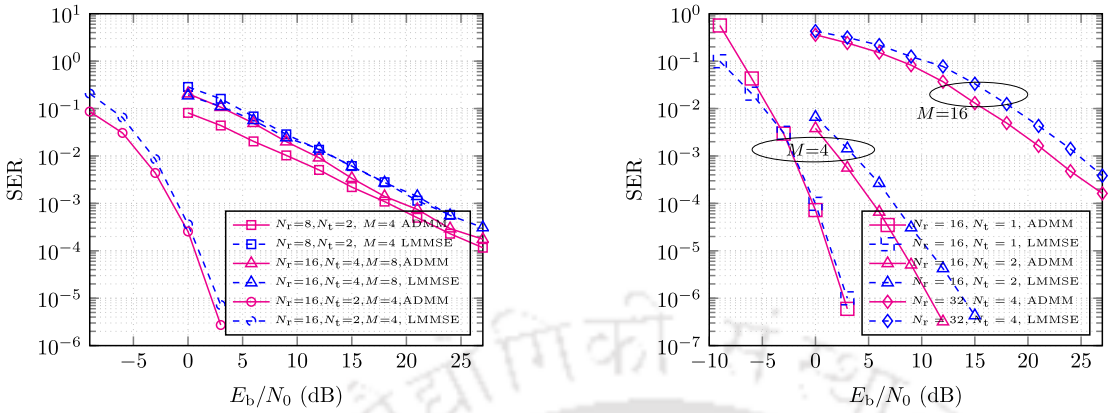
Parameter	Value
Modulation order (M)	$4, 8, 16$
Number of UEs (J)	$6, 8$
Number of overlapped UEs (N_u)	$d_f = 3, 4$ (SCMA), $J = 6, 8$
Number of resources (K)	$4, 4$
Number of transmit antenna (N_t)	$2, 4, 8$
Number of receive antenna (N_r)	$4, 8, 32$
Overloading factor (λ)	150 %, 200 %
Number of iterations (T)	15 (ADMM), 10 (MPA), 10 (EPA),10 (GAMPA)

3.5.1 SER Performance of SMX CD-NOMA Systems

Fig. 3.6 (a) and Fig. 3.6 (b) show the SER performance of the SMX-SCMA and the SMX-DCMA systems, respectively. The simulations are performed for 200% overloaded CD-NOMA systems with $J = 8, K = 4$. We consider $N_u = 4$ and $N_u = 8$ for SCMA and DCMA, respectively. In each case, the value of N_r is increased as the value of N_t increases to maintain $N_r = N_u N_t$ in (3.14). The performance of ADMM and LMMSE further improves when $N_r > N_u N_t$. The LMMSE performance degrades when the ratio of N_r to $N_u N_t$ is close to one, but in contrast, ADMM still maintains the decent performance. For both SCMA and DCMA systems, the ADMM outperforms the LMMSE, as shown in Fig. 3.6 (a) and Fig. 3.6 (b), respectively. However, significant SNR gains were recorded for SMX-DCMA systems. To highlight, the ADMM offers a maximum of 3 dB gain over the LMMSE in the DCMA system. This shows the SER performance advantage of ADMM over LMMSE with almost similar complexity, as shown in Fig. 3.5.

3.5.2 SER Performance of SM CD-NOMA Systems

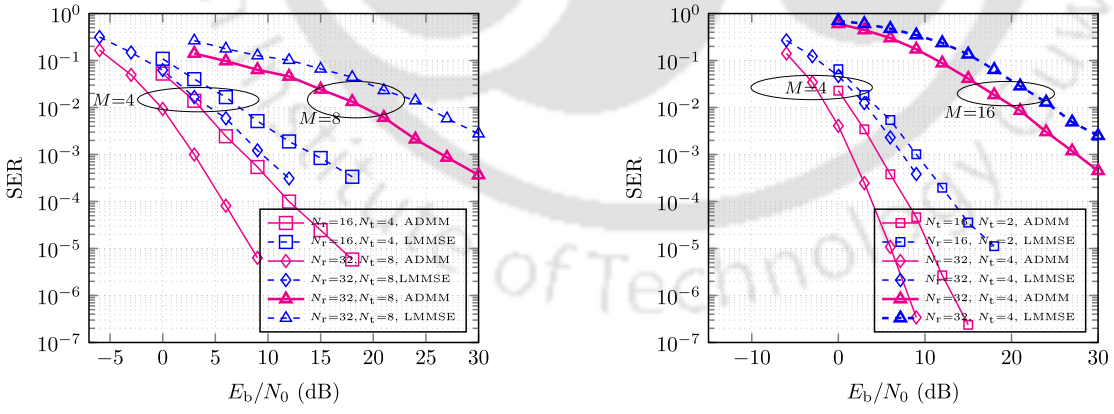
Fig. 3.7 (a) and Fig. 3.7 (b) show the SER performance of the SM-SCMA and SM-DCMA systems, respectively. The performance of the ADMM-based detector mainly depends on N_t, N_r, and M, as we have seen in the previous subsection. The observations mentioned above for SMX-



(a) (b)

Fig. 3.6: ADMM and LMMSE detectors' SER performance comparison for SMX-CD-NOMA systems (a) SMX-SCMA with λ = 200% , (b) SMX-DCMA with λ = 200%.

CD-NOMA systems apply to SM-CD-NOMA systems. In SM-CD-NOMA systems, each UE's codeword contains the active antenna and codeword index information of that particular UE. Fig. 3.7 (a) and Fig. 3.7 (b) depict the efficiency of ADMM in detecting the information both in the signal domain and spatial domain. In contrast, the LMMSE performs poorly in recovering the signal and spatial domain data. As a result, the ADMM offers a maximum of 6 dB gains over LMMSE in both SCMA and DCMA systems, as shown in Fig. 3.7 (a)and Fig. 3.7 (b), respectively.



(a) (b)

Fig. 3.7: ADMM and LMMSE detectors' SER performance comparison for SM-CD-NOMA systems (a) SM-SCMA with λ = 200%, (b) SM-DCMA with λ = 200%.

3.5.3 SER Performance: ADMM vs. Nonlinear Detectors

This subsection presents the comparative analysis of ADMM with benchmark nonlinear iterative detectors. For SCMA systems, the most prominent detectors, namely MPA, EPA, and

GAMPA, are considered. For DCMA systems, a GSD detector is considered for comparison.

3.5.3.1 ADMM vs. MPA

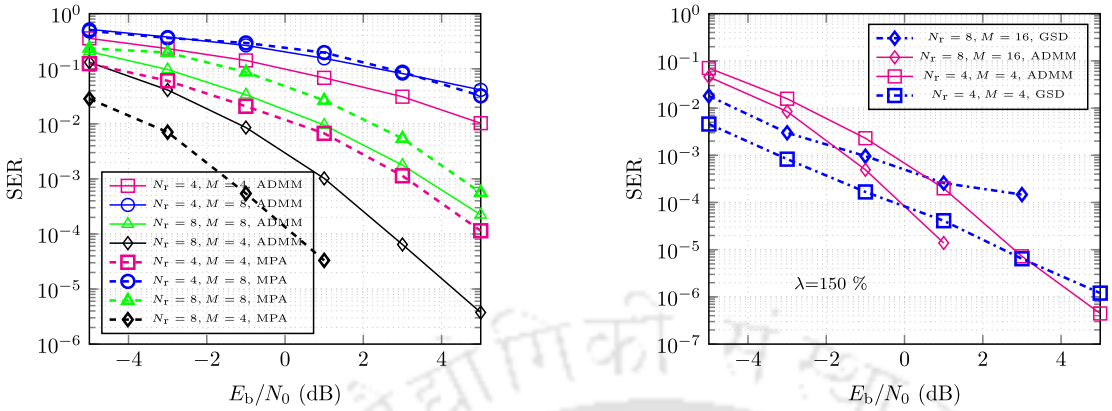
Fig. 3.8 (a) shows the SER performance comparison between the ADMM and MPA detectors for the SIMO-SCMA system. MPA is a near-optimal and exponentially complex detector available for SCMA systems [57]. MPA gives a maximum SNR gain of around 2.5 dB over ADMM at SER $= 10^{-2}$ for $N_r = 4$ and $M = 4$, as shown in Fig. 3.8 (a). As the modulation order M increases (M=8), the performance of ADMM converges to MPA, as observed for $N_r = 4, M = 8$ in Fig. 3.8 (a). Therefore, for $M = 8$, codebook distance properties influence MPA and ADMM-based detectors similarly. Further, for $N_r = 8, M = 8$, the ADMM detector shows an approximately 1.8 dB SNR gain at SER $= 10^{-3}$ over the MPA detector. The increase in the number of observations significantly improves the ADMM performance compared to MPA for large-size codebooks (i.e., $M = 8$).

3.5.3.2 ADMM vs. GSD

Fig. 3.8 (b) shows the SER performance comparison between the ADMM and GSD detectors for the DCMA system. Observe from Fig. 3.8 (b) that, at low SNR regions, GSD outperforms ADMM with a maximum of around 2 dB SNR gain at SER=10^{-4}. As SNR increases, GSD performance decreases compared to ADMM. Unlike GSD, as M increases, ADMM performance is enhanced by double the number of antennas received, as shown in Fig. 3.8 (b). In contrast to GSD, ADMM performance improves as the number of observations increases.

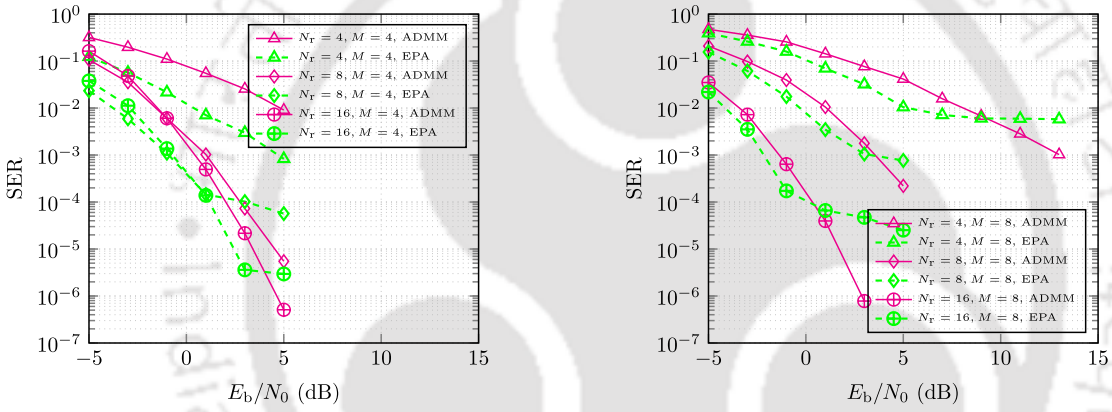
3.5.3.3 ADMM vs. EPA

The SER performance comparison between ADMM and EPA is shown in Fig. 3.9, and the observations are as follows. As shown in Fig. 3.9 (a), EPA outperforms the ADMM significantly for $M = 4$ size SCMA codebooks. For $M = 8$, the ADMM offers competitive performance to the EPA as shown in Fig. 3.9 (b). Further, as N_r (number of observations) increases, the performance gap between the ADMM and EPA reduces for both $M = 4$ and $M = 8$ cases. Moreover, EPA detection for uncoded SCMA systems suffers from high error floors [61]. The error floor level depends on N_r and M, which can be observed in Fig. 3.9 (a) and Fig. 3.9 (b).



(a) (b)

Fig. 3.8: SER performance comparison: ADMM vs. MPA and GSD detectors (a) SIMO-SCMA: ADMM vs. MPA with $\lambda = 150\%$, (b) SIMO-SCMA: ADMM vs. GSD with $\lambda = 150\%$.

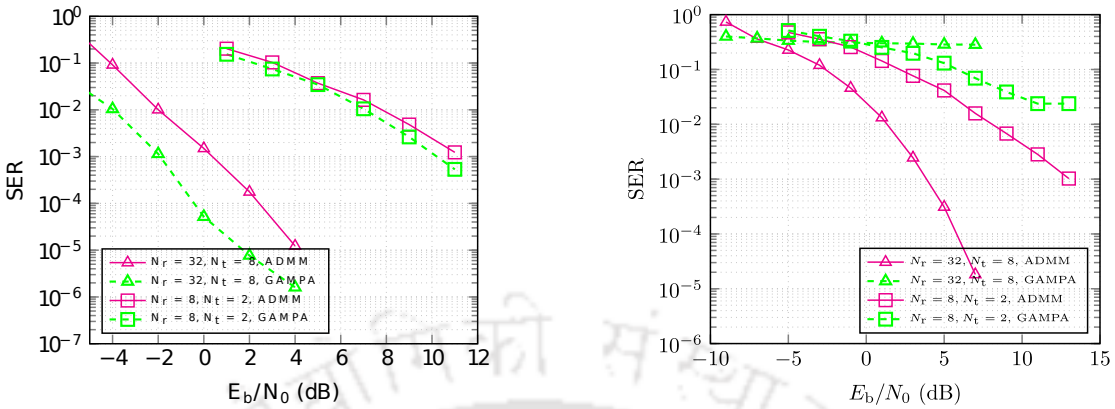


(a) (b)

Fig. 3.9: ADMM vs. EPA: SER performance for 200% overloaded SCMA system (a) $M = 4$, (b) $M = 8$.

3.5.3.4 ADMM vs. GAMPA

Fig. 3.10 depicts the SER performance comparison of ADMM and GAMPA detectors for SCMA systems. The Gaussian approximation of the interference (GAI) [62] and the number of short cycles lead to performance degradation. The GAI-based MPA performance degrades for lower values of N_t [81]. For instance, in the $M = 4$ and $\lambda = 150\%$ SCMA system, the GAMPA outperforms the ADMM as shown in Fig. 3.10 (a). As the number N_t of transmit antennas decreases, GAMPA approaches ADMM performance. For $M = 8$ as shown in Fig. 3.10 (b), the performance of GAMPA degrades and it is significantly inferior to the ADMM detector. Thus, ADMM outperforms GAMPA in many folds. In specific, the ADMM is a more appropriate detector than GAMPA for the large codebook size SCMA systems.



(a) (b)

Fig. 3.10: ADMM vs. GAMPA: SER performance for 200% overloaded SCMA system (a) $M = 4$ (b) $M = 8$.

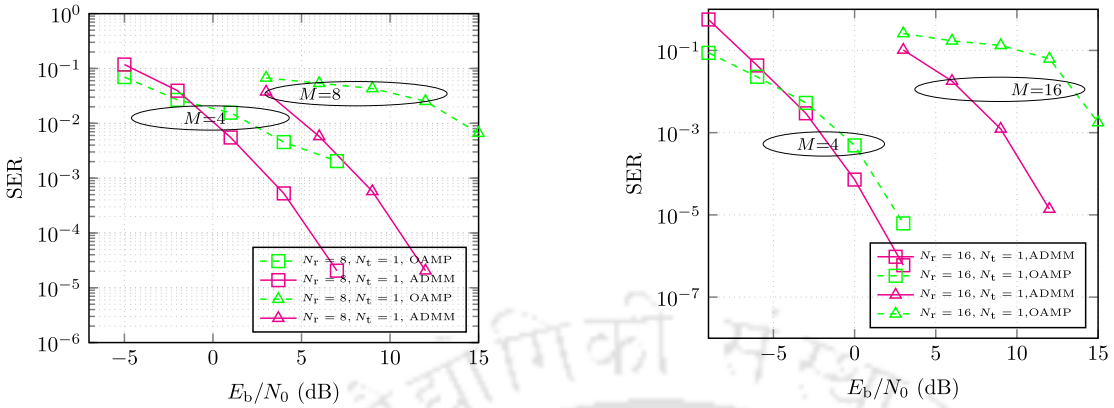
3.5.3.5 ADMM vs. OAMP

Fig. 3.11 illustrates the symbol error rate (SER) performance comparison between ADMM and OAMP for CD-NOMA systems. The OAMP algorithm requires a specific input signaling structure and assumes that entries are independently and identically distributed (i.i.d.). Resource-wise processing operates under the assumption of independent entries in the signal vector to be estimated. By incorporating resource-wise processing in (3.20), OAMP detection becomes feasible. Fig. 3.11 (a) and Fig. 3.11 (b) present the SER performance of SCMA and DCMA systems, respectively. The proposed ADMM detector yields signal-to-noise ratio (SNR) gains over OAMP of approximately 4 dB at an SER of 10^{-3} and 1.1 dB at an SER of 10^{-5} with $M = 4$ for SCMA and DCMA systems, respectively. These SNR gains become even more substantial for $M = 8$ and $M = 16$ for SCMA and DCMA, respectively. Therefore, ADMM emerges as a more efficient detector than OAMP for larger values of M, with computational complexity similar to that shown in Table 3.1.

3.5.4 SER Performance: Imperfect CSI

All the above simulations are analyzed by considering perfect CSI at the receiver. In practical scenarios, obtaining perfect CSI at the receiver is not feasible due to CEEs. Thus, it is important to show the performance of the proposed detector in the presence of CEEs. An imperfect channel can be modeled as [82]

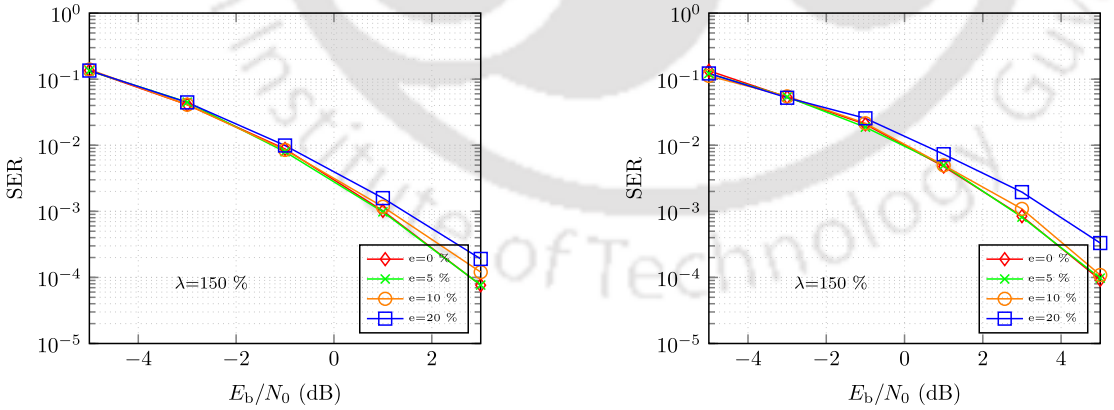
$$\hat{\mathbf{H}} = \mathbf{H} + e\,\mathbf{\Omega}, \tag{3.52}$$



(a) (b)

Fig. 3.11: ADMM vs. OAMP: SER performance for 200% overloaded SCMA system (a) SCMA, (b) DCMA.

where $e\mathbf{\Omega}$ is the estimation error that is uncorrelated with $\mathbf{H}$ and the entries of $\mathbf{\Omega}$ are i.i.d. complex Gaussian random variables with zero mean and unit variance, i.e., $\mathcal{CN}(0,1)$. The quantity e in (3.52) determines the variance of the CEE. Fig. 3.12 depicts the performance of the proposed ADMM-based detector for the imperfect CSI with $e = 0\ \%, e = 5\ \%, e = 10\ \%$, and $e = 20\ \%$. The simulations for SCMA and DCMA systems are shown in Fig. 3.12 (a) and Fig. 3.12 (b), respectively. Observe that the impact of the CEE on the ADMM-based detector's performance is minimal. Therefore, the ADMM-based detector can be applied in imperfect CSI scenarios.



(a) (b)

Fig. 3.12: SER performance of CD-NOMA system in imperfect CSI for $e = 0\ \%, e = 5\ \%, e = 10\ \%, e = 20\ \%$ (a) SIMO SCMA system with $N_r = 8$, $M = 4$, (b) SIMO DCMA system with $N_r = 8$, $M = 4$.

3.5.5 LDPC Coded CD-NOMA Systems

The bit error rate (BER) performance of LDPC-coded CD-NOMA systems is depicted in Fig. 3.13. The ratio $\frac{N_r}{N_u N_t} = 1$, and $\lambda = 200\%$ (with $J = 8$ and $K = 4$) overloading for

both SCMA and DCMA systems ensure a fair comparison. For channel coding, we utilized the rate-0.32 5G-New Radio (NR) LDPC code [83], featuring 352 message bits (k) and 1088 codeword bits (n). The minsum algorithm [84] serves as the decoder for the LDPC codes. The soft information generated by the ADMM detector is fed into the LDPC decoder as an input log-likelihood ratio (LLR). The soft output or LLR from the ADMM detector can be derived as follows.

Example 3. *Let us consider $\hat{\mathbf{x}}_j$ as the estimated codeword for the jth user. $\mathbf{x}_{j,m}$ is the mth codeword of the jth user codebook $\mathcal{X}_j^{K\times M}$ with M codewords. The estimation errors of M codewords are given by*

$$e_{j,m} = \|\hat{\mathbf{x}}_j - \mathbf{x}_{j,m}\| \ \forall m.$$

The approximate a posteriori probability (APP) obtained for the mth codeword after normalization, $e(\mathbf{x}_{j,m}|\mathbf{y})$ can be written as $\frac{1}{S_e}\frac{1}{e_{j,m}}$, where $S_e = \sum_{m=1}^{M}\frac{1}{e_{j,m}}$. The lth bit LLR value corresponding to mth codeword in the jth codebook is given by

$$\text{LLR}(b_{j,l}) = ln\frac{\sum_{\mathbf{x}_j\in\mathcal{X}_{j,m}^{1}} e(\mathbf{x}_{j,m}|\mathbf{y})}{\sum_{\mathbf{x}_j\in\mathcal{X}_{j,m}^{0}} e(\mathbf{x}_{j,m}|\mathbf{y})}. \tag{3.53}$$

□

Consider the estimated codeword $\hat{\mathbf{x}}_j$ for the jth user. Suppose $\mathbf{x}_{j,m}$ is the mth codeword of the jth user codebook $\mathcal{X}_j^{K\times M}$ with M codewords. The distance between $\hat{\mathbf{x}}_j$ and each of the M codewords is given by

$$d_{j,m} = \|\hat{\mathbf{x}}_j - \mathbf{x}_{j,m}\| \quad \forall m.$$

After normalization, the distance metric denoted by $d(\mathbf{x}_{j,m}|\mathbf{y})$ can be written as $\frac{1}{S_d}\frac{1}{d_{j,m}}$, where $S_d = \sum_{m=1}^{M}\frac{1}{d_{j,m}}$. The ith bit LLR value corresponding to mth codeword in the jth codebook can be considered as

$$\text{LLR}(b_{j,i}) = \log\frac{\sum_{\mathbf{x}_j\in\mathcal{X}_j^{i,0}} d(\mathbf{x}_{j,m}|\mathbf{y})}{\sum_{\mathbf{x}_j\in\mathcal{X}_j^{i,1}} d(\mathbf{x}_{j,m}|\mathbf{y})} \tag{3.54}$$

where $\mathcal{X}_j^{i,0}$ and $\mathcal{X}_j^{i,1}$ denote the set of codewords of which the ith bit corresponds to "0" and "1" for user-j, respectively. Fig. 3.13 shows the BER comparison of uncoded and LDPC coded for both SCMA and DCMA systems. Due to full diversity order, DCMA offers multi-fold SNR

gains over SCMA for uncoded and coded systems. The soft output ADMM detector provides a significant performance gain for the coded CD-NOMA system over uncoded CD-NOMA system.

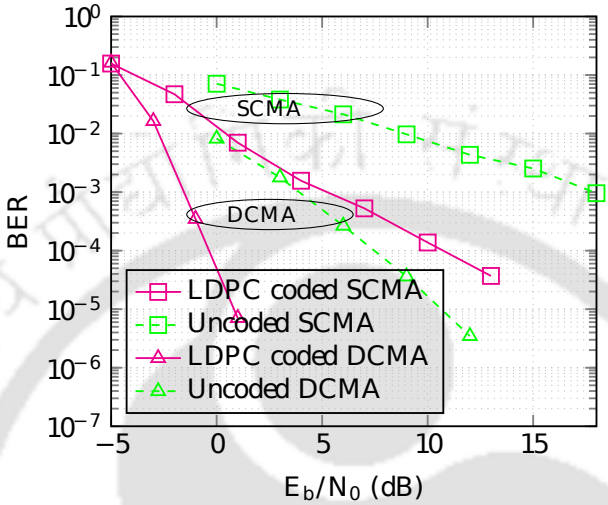
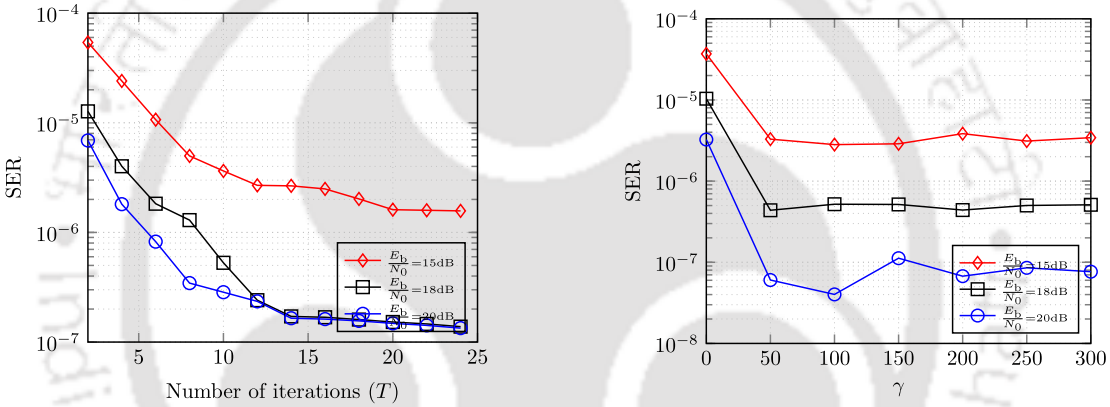


Fig. 3.13: SIMO SCMA system with $N_r = 8$, $M = 4$.

3.5.6 Convergence of ADMM and Selection of Parameters

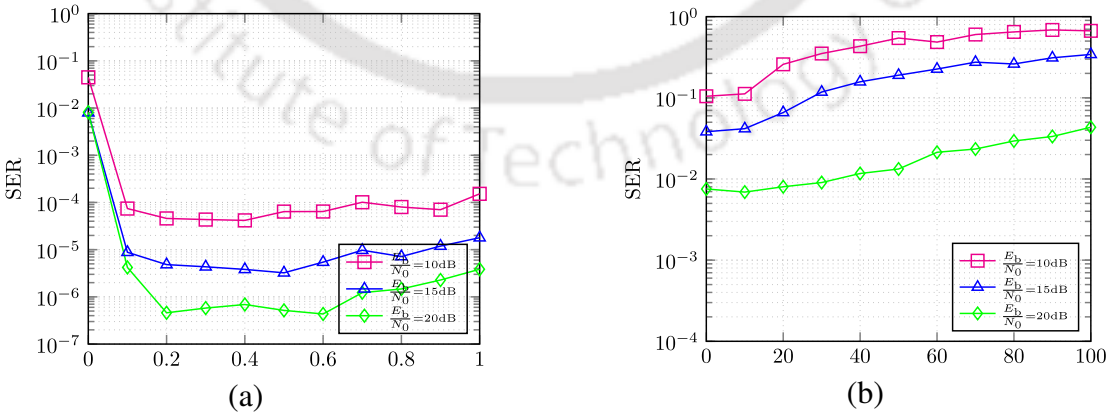
The convergence analysis of the ADMM-based detector in the DCMA system is performed based on the Monte Carlo simulations. Fig. 3.14 (a) shows the impact of ADMM iterations on the SER performance. The proposed detector exploits the iterative nature of ADMM, and it converges after some finite number of iterations. The SER performance improves as the number of iterations increases. The improvement in SER becomes marginal after a certain number of iterations. For all considered $\frac{E_b}{N_0}$ values, ADMM converges after 15 iterations, as shown in Fig. 3.14 (a). The three lemmas in **Appendix** A provide the guidelines for parameter selection to improve the detector's performance. Further, the penalty parameters $\{\gamma_j\}_{j=1}^{J}$ and ρ are selected based on the empirical observations given in Fig. 3.14 (b) and Fig. 3.15 to maximize the SER performance. The augmented Lagrangian parameter ρ is selected as the scalar multiple, ϵ, of reciprocal of SNR [75] as given in **Algorithm 2**. The impact of γ on SER performance in the ADMM-based detector is analyzed through Monte Carlo simulations, and the results are given in Fig. 3.14 (b). For low values of γ, the ADMM performance degrades, and for large values, the variation in SER becomes inconsequential. When γ is close to zero, the impact of the penalty

terms is nullified. Thus, the alternative variables in the penalty function lose importance in minimizing the objective function. Observe from the Fig. 3.14 (b) that the ADMM performs better in the range $\{\gamma_j\}_{j=1}^{J} \in [50\ \ 100]$. As we mentioned in Algorithm 2, ρ is a function of ϵ. The importance of the selection of ϵ in ρ is shown in Fig. 3.15. For low values of ϵ, the SER performance improves initially and then becomes marginal after a certain value, as shown in Fig. 3.15 (a). This shows the algorithm's convergence for low values of ϵ. The performance degrades for large values of ϵ as shown in Fig. 3.15 (b). It can be observed that the algorithm does not converge for large values of ϵ.



(a) (b)

Fig. 3.14: Impact of T and γ on the ADMM-based detector's SER performance (a) SIMO SCMA system with $N_r = 8$, $M = 4$, (b) SIMO DCMA system with $N_r = 8$, $M = 4$.



(a) (b)

Fig. 3.15: ϵ vs. SER performance (a) SIMO SCMA system with $N_r = 8$, $M = 4$, (b) SIMO DCMA system with $N_r = 8$, $M = 4$.

3.6 Summary

This chapter presents a novel multi-user detection (MUD) algorithm based on the ADMM framework, specifically designed for UL large-scale MIMO CD-NOMA systems. The proposed algorithm is effectively applied to both SMX-CD-NOMA and SM-CD-NOMA configurations, and its SER performance is analyzed under various system parameter variations.

In this work, the ML detection problem is reformulated as a shared optimization problem, efficiently addressed using the ADMM algorithm. The limitations of existing CD-NOMA detection methods, such as MPA, EPA, GAMPA, OAMP, MMSE, and GSD, are thoroughly evaluated and compared with the proposed ADMM-based detector. Additionally, a soft-output ADMM detector is implemented for LDPC-coded CD-NOMA systems.

This work is further extended to address the JADD problem in grant-free NOMA systems. The exploration of ADMM-based detectors for practical mMTC networks is presented in Chapter 4.

4

Joint Activity and Data Detection in CB based Grant-free NOMA Systems

Contents

4.1 Introduction

As mentioned above, NOMA systems have been well investigated to address several challenges of 6G wireless networks [2, 3]. The overloading nature of NOMA supports massive connectivity with a limited number of REs. In this chapter, the term NOMA primarily refers to CD-NOMA. OMA allocates each RE exclusively to a single user for transmission. In OMA, each UE schedules network access after receiving a BS grant through a four-step hand shake procedure [4, 85]. This procedure is popularly known as the grant-based random access procedure. This approach is prevalent in human-type communication (HTC) e.g. long-term evolution (LTE), which involves fewer devices with large amounts of data. It introduces additional overhead and latency, which is often negligible compared to the large amount of data transmission. This assumption cannot hold for mMTC networks.

6G wireless networks hold great promise for mMTC, where a vast number of devices are connected to the network. These devices typically transmit short-packet data sporadically. However, the conventional OMA scheme may become a bottleneck for mMTC, as it struggles to support massive connectivity. One of the key challenges in 6G networks is achieving massive connectivity with limited resources. Additionally, URLLC is a critical requirement for 6G wireless networks.

NOMA, with its potential for massive connectivity, is regarded as a key enabler for grant-free communication. In a grant-free approach, users can transmit data whenever needed, without waiting for permission from the base station. The overloading capability of NOMA facilitates massive connectivity, while grant-free access supports URLLC. Consequently, grant-free NOMA simultaneously addresses the requirements of massive connectivity and URLLC [4].

In mMTC networks, only 10% of the devices are active even during peak times due to sporadic traffic [31]. NOMA enables grant-free access, allowing devices to connect without scheduling and transmit data as needed. The BS lacks real-time knowledge of device activity, which requires mechanisms to identify active users. The main challenge in grant-free NOMA is JADD, with an acceptable performance and complexity trade-off [31, 74, 86–89]. The transmit-

ted signal structure affects JADD, resulting in two variants of CD-NOMA, as discussed below.

As mentioned in Chapter 1, two main CD-NOMA variants have been studied in the literature: (1) sequence-based NOMA and (2) CB-based NOMA. This chapter specifically focuses on grant-free CB-based NOMA systems with an emphasis on MUD for JADD.

4.1.1 Related Literature

Extensive research on MUD for grant-free NOMA in JADD has focused on spread-sequence-based CD-NOMA in mMTC networks traffic [31,74,86–89]. Due to sporadic traffic assumption in grant-free NOMA systems, JADD problem is formulated as a sparse recovery problem [90]. Standard compressive sensing (CS) algorithms are employed to solve the formulated JADD problem. These algorithms can be classified into greedy-based, Bayesian inference (e.g., approximate MPA), and convex optimization approaches. Each of these approaches offers distinct advantages and limitations, as detailed below in the context of JADD.

CS-based orthogonal matching pursuit (OMP)-aided MUD is proposed for downlink (DL) code division multiple access (CDMA) systems [91]. The structured sparsity-based MUD algorithm known as structured iterative support detection (SISD) is proposed for the JADD of grant-free NOMA systems [31]. Structured sparsity is a restricted assumption of the activity of the devices. Later, research has been conducted on dynamic activity scenarios. A dynamic system model is considered, and a dynamic CS (DCS)-based MUD is proposed based on the subspace pursuit (SP) algorithm [86]. In addition, the temporal correlation between adjacent time slots is leveraged to enhance detection. However, DCS-based MUD, assuming that the sparsity level is known, uses the support set of the prior slot without accounting for the partial overlap with the current set. A DCS algorithm that leverages prior information for JADD in dynamic grant-free NOMA was proposed, exploiting temporal correlation for notable performance gains [87]. BCS-based MUD models, designed for block sparsity in UL grant-free NOMA, further improve performance over structured sparsity but lack adaptability to dynamic scenarios [88]. Existing detectors, mainly greedy-based (e.g. OMP, SP), compute local support sets iteratively with expensive inverse operations and are not guaranteed for the global optimum, especially for less sparse signals or under significant noise [92]. These methods also perform

poorly with coherent measurement matrices.

The approximate message passing (AMP) is another approach based on MPA to solve the JADD problem in grant-free NOMA [89]. AMP relies on the assumption that the entries of the channel matrix are IID to carry out state evolution (SE) analysis. It performs poorly when the channel matrix is not IID [93]. An orthogonal AMP-based MUD is proposed for JADD of grant-free NOMA systems [94]. Both of these works exploit the structured sparsity model for performance enhancement. However, this sparsity model is often unrealistic for mMTC networks and does not account for dynamic user activity.

Convex optimization approaches for sparse recovery, particularly LASSO, have been widely studied in CS [95]. As stated in Chapter 3, ADMM has emerged as an efficient solution for convex and non-convex problems [70]. In addition, its decomposition and parallel computing capabilities make it suitable for large-scale data. Strong theoretical convergence properties make ADMM more robust in a wide range of applications. Later, ADMM is adapted to solve the LASSO problem [70]. LASSO-based MUD for JADD in grant-free NOMA uses ADMM, which leverages structured sparsity for improved detection and adapts to dynamic scenarios [74]. However, blindly relying on temporal correlation between slots can degrade performance, suggesting room for further enhancement in ADMM-based MUD for dynamic scenarios.

The comprehensive survey outlined above introduces various models for grant-free NOMA systems and proposes new algorithms to tackle the multi-user detection challenge for JADD. However, all of these studies focus on JADD for grant-free spread-sequence-based NOMA systems. To the best of the authors' knowledge, no prior work has explored the design of a CB-based NOMA system for grant-free scenarios or developed a detection algorithm to address the JADD problem in such systems. As discussed earlier, CB-based grant-free NOMA offers several advantages over sequence-based grant-free NOMA. Additionally, the inherent block-sparsity structure in CB-based grant-free NOMA enhances the robustness of sparse signal reconstruction.

4.2 Focus and Contributions

The comprehensive survey highlights various grant-free NOMA models and algorithms that address MUD for JADD, all focusing on spread-sequence-based systems. However, no prior work has investigated the CB-based NOMA design for grant-free scenarios or developed JADD-specific detection algorithms. CB-based grant-free NOMA outperforms sequence-based systems by leveraging block-sparsity for more robust sparse signal reconstruction. This chapter formulates two block-sparsity-aided optimization problems: sparse group LASSO for SCMA and group LASSO for DCMA to address the JADD challenge within block-sparsity frameworks. To solve these, an ADMM-based detection algorithm is developed. For grant-free NOMA systems, we introduce a dynamic block sparsity model in a frame-wise manner to tackle critical challenges in mMTC networks. A robust prior information-aided ADMM algorithm is specifically designed for JADD in dynamic conditions, distinguishing it from previous SP algorithms [87]. A quality parameter is introduced to support reliable active-user detection that quantifies the accuracy of prior active-user information from preceding slots for use in the current slot. Accurately estimating this parameter without precise active user support poses a challenge, and setting an inaccurate quality parameter, as in [87], adversely affects the performance of JADD. Thus, we propose a simple two-step process: Step 1 derives the quality of the prior information, and Step 2 leverages it to optimize JADD performance. This proposed detector not only enhances detection accuracy but also maintains lower complexity relative to existing models, particularly in dynamic scenarios.

The detailed contributions of this work are listed below.

• ***Block CS (BCS) models for CB-based CD-NOMA systems***

The proposed grant-free CB-based CD-NOMA (SCMA and DCMA) systems leverage the inherent block sparsity in their input signals, modeled via BCS of CD-NOMA codewords. This framewise dynamic block sparsity model enhances MUD performance for JADD, with results supported by extensive numerical experiments.

• ***Low complexity JADD problem formulation***

The computationally intensive detection problem in SCMA and DCMA is addressed through a convex optimization framework, reducing the MUD for JADD to block sparse recovery problems. The JADD problem for sparse and dense CB-based CD-NOMA systems is formulated using two block sparse recovery approaches: sparse-group LASSO for SCMA and group LASSO for DCMA, enabling efficient detection in grant-free systems.

• ***Use of ADMM algorithm to solve the JADD problem***

A parallel computing-aided ADMM-based detection algorithm is proposed to solve the formulated JADD problem. The distributive nature of ADMM makes it suitable for large-scale CD-NOMA systems, which are expected to exist in mMTC networks. Hence, the proposed detector is scalable, unlike the MPA and GSD detectors available for both SCMA and DCMA systems, respectively, in the literature.

• ***Block-sparsity based dynamic compressive sensing (DCS) approach***

The random arrivals and departures of users in the mMTC network further challenge the JADD. To address this, a dynamic system model with practical use cases is proposed based on realistic assumptions. A DCS approach via the ADMM algorithm leverages temporal correlation to compute robust prior information, which is integrated into ADMM iterations, enhancing MUD performance in the dynamic model.

• ***Efficiency of the algorithm***

The proposed algorithm achieves modest accuracy within a small number of iterations, which has proven sufficient for acceptable performance, as confirmed by extensive numerical experiments. Additionally, the computational complexity of the proposed detector is lower compared to that of the benchmark detector, making it particularly advantageous for JADD in dynamic scenarios.

4.3 UL CB-based grant-free NOMA system models

This section presents block-sparsity models for UL single-input multi-output (SIMO) grant-free SCMA and DCMA systems with multi-antenna BS. CB-based grant-free NOMA systems utilize a block-sparsity structure to improve JADD's performance. This section introduces two

block-sparse grant-free NOMA systems. Section 4.3.1 introduces a one-shot model for single-time slot transmission, while Section 4.3.2 describes a dynamic frame-wise model for random user access over time.

4.3.1 One-shot block sparsity model (Model-1)

This section models a CB-based NOMA system, assuming a single transmission time slot, where the BS is equipped with N_{r} antennas, serving J number of users with K number of resources. Active users transmit codewords, while inactive users remain silent, represented by zero blocks. The joint augmented codebook $\mathcal{X} \triangleq \{\mathcal{X}_a \cup \mathbf{0}\}$ includes codewords for active and inactive users. The observation on n_{r}th antenna at the BS is given by

$$\mathbf{r}^{(n_{\mathrm{r}})} = \sum_{j=1}^{J} \mathrm{diag}\left(\mathbf{h}_j^{(n_{\mathrm{r}})}\right)\mathbf{x}_j + \mathbf{w}^{(n_{\mathrm{r}})}.\Psi\Psi$$

The entire observation at BS is given by

$$\mathbf{r} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{4.1}$$

where

- $\mathbf{r} = \left[\mathbf{r}^{(1)^T}, \mathbf{r}^{(2)^T}, \ldots, \mathbf{r}^{(N_{\mathrm{r}})^T}\right]^T \in \mathbb{C}^{N_{\mathrm{r}}K}$, $\mathbf{r}^{(n_{\mathrm{r}})} \in \mathbb{C}^{K\times 1}$.
- $\mathbf{w}^{(n_{\mathrm{r}})} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{KN_{\mathrm{r}}})$ denotes the additive white Gaussian noise (AWGN) at the receiver and
$$\mathbf{w} = [\mathbf{w}^{(1)^T}, \mathbf{w}^{(2)^T}, \ldots, \mathbf{w}^{(N_{\mathrm{r}})^T}]^T$$
- $\mathbf{h}_j^{n_{\mathrm{r}}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ denotes the Rayleigh fading channel vector between the jth UE and n_{r}th receive antenna.

- $$\mathbf{H} = \begin{bmatrix} \text{diag}(\mathbf{h}_1^{(1)}), \ldots, \text{diag}(\mathbf{h}_j^{(1)}), \ldots, \text{diag}(\mathbf{h}_J^{(1)}) \\ \text{diag}(\mathbf{h}_1^{(2)}), \ldots, \text{diag}(\mathbf{h}_j^{(2)}), \ldots, \text{diag}(\mathbf{h}_J^{(2)}) \\ \\ \\ \text{diag}(\mathbf{h}_1^{(N_r)}), \ldots, \text{diag}(\mathbf{h}_j^{(N_r)}), \ldots, \text{diag}(\mathbf{h}_J^{(N_r)}) \end{bmatrix}$$

- $\mathbf{x} = \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_J^T\right]^T \in \mathbb{C}^{JK\times 1}$, where $\mathbf{x}_j = \left[x_{j,1}, x_{j,2}, \ldots, x_{j,K}\right]^T \in \mathbb{C}^{K\times 1}$ is the codeword of the jth user.

In a CB-based grant-free NOMA system, the transmitted signal $\mathbf{x}$ exhibits block sparsity, where each user's codeword forms a unique block. For JADD, user activity detection utilizes this blockwise sparsity. The active user set is defined as

$$\Lambda = \{j \in \{1, 2, \ldots, J\} \mid \|\mathbf{x}_j\|_2 \neq 0\}.$$

The sparsity level denoted as S_l, is defined as $S_l = |\Lambda|$.

4.3.2 Frame-wise dynamic block-sparsity model (Model-2)

This subsection presents a practical system model for mMTC, addressing the limitation of the static one-shot model, which ignores temporal correlation. A frame-wise structured sparsity model [31,88] is unsuitable for dynamic user activity. Therefore, a new system model is needed to accommodate varying user activity over time. This section integrates three concepts: block sparsity, dynamism, and frame-wise design to develop a system model for practical grant-free CB-based NOMA systems. The block-sparsity model introduced here is inspired by the one-shot static model discussed in Section 4.3.1. This model incorporates a frame-wise dynamic transmission approach within the uplink CB-based NOMA system, capturing the dynamism present in sporadic traffic networks, such as mMTCs. In real-world scenarios, the set of active users in the network changes continuously and gradually. A frame consists of multiple time slots that allow users to enter or leave the network at any time. This temporal correlation can be used to detect active users using the dynamic block compression sensing (BCS) algorithm

discussed in the following section. The entire observation of size $N_r K \times 1$ at BS in the time slot l is given by

$$\mathbf{r}^{[l]} = \mathbf{H}^{[l]}\mathbf{x}^{[l]} + \mathbf{w}^{[l]}, \quad l = 1, \ldots, L \tag{4.2}$$

where $\mathbf{x}^{[l]}$ may vary across different time slots, and the active user support set in (4.2) exhibits the following property:

$$\Lambda^{[1]} \neq \Lambda^{[2]} \neq \ldots = \Lambda^{[L]},$$

where $\Lambda^{[l]} = \{j \in \{1, 2, \ldots, J\} \mid \|\mathbf{x}_j^{[l]}\| \neq 0\}$ indicates the set of active users in the lth time interval. In this model, the set of active users varies over different time intervals.

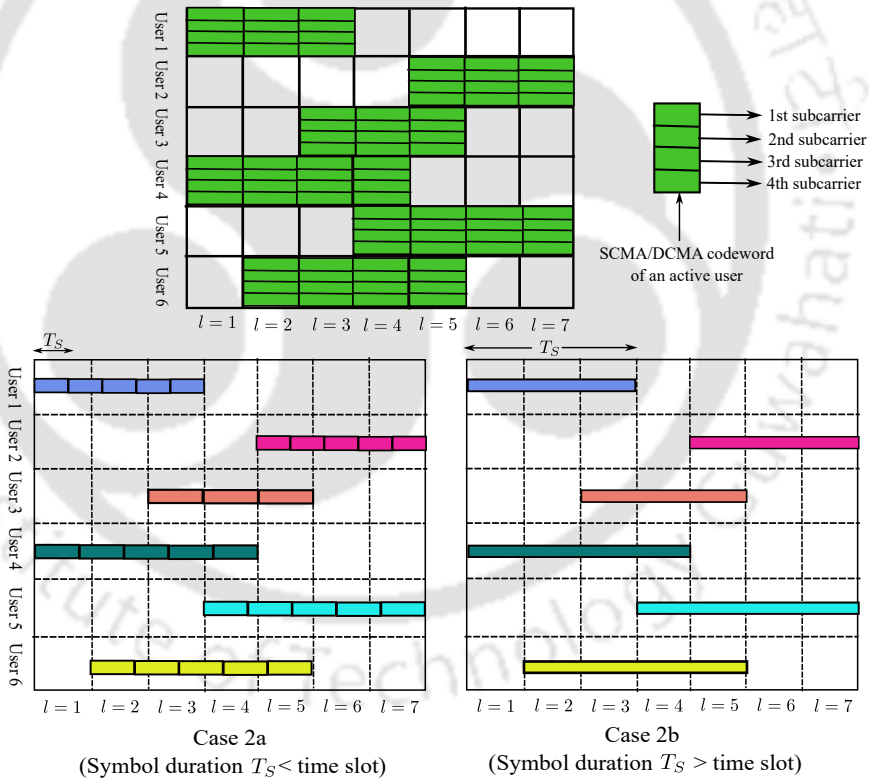


Fig. 4.1: Two scenarios involved in Case-2 of frame-wise dynamic block sparsity model.

The system model described in (4.2) implies that the observations vary across time intervals, making joint processing of the entire frame of observations infeasible. The goal of the JADD problem for (4.2) is to detect both active users and their transmitted data in each individual time interval. Specifically, observation $\mathbf{r}^{[l]}$ in a given time interval aids in recovering $\mathbf{x}^{[l]}$ for that interval. Additionally, the temporal correlation between active user sets across consecutive

time slots can be leveraged for user detection, as many users transmitting data in previous time slots are likely to remain active in the current time slot. Therefore, the estimated active user set and the data from the previous time interval can serve as prior information for the current one. By integrating this prior information, the proposed JADD algorithm in Section 4.4 enhances the performance of the frame-wise dynamic block-sparsity system model. The dynamic model considered in this paper addresses two distinct use cases within mMTC networks:

(i) **Case-1 (Fast-changing user activity):** Active users' behavior changes rapidly.

(ii) **Case-2 (Slow-changing user activity):** Active users' behavior changes more gradually.

Case-1: In the first scenario, which typically occurs during short packet transmissions, users remain active for only a brief period. This leads to the assumption that the activity time of a user is strictly less than the duration of a time slot within a frame, making the activity of the user independent of different time slots. Therefore, JADD must be performed independently for each time slot.

Case-2: In the second scenario, which generally occurs with moderate packet transmissions or when time slots within a frame are of short duration, the user activity period exceeds the length of a time slot. This means that some active users are likely to transmit data across consecutive time intervals, creating a temporal correlation between the active user sets. This scenario has two possible subcases, illustrated in Fig. 4.1. This temporal correlation can enhance the performance of MUD in the following ways.

- **Case 2a:** When the duration of the symbol is shorter than or equal to the time interval (that is, the duration of the symbol $\leq$ time interval), a temporal correlation exists only through the sets of active users in adjacent time intervals as shown in Case 2a of Fig. 4.1. In this case, the prior information is presented as a quality support set between time slots.
- **Case 2b:** When the duration of the symbol is longer than the time slot (i.e., symbol duration¿time slot), temporal correlation extends not only across the active user sets but also includes the signal information between the time slots as shown in Case 2b of Fig. 4.1. Here, the prior information consists of both the quality support set and the signal information shared across time slots.

The detailed formulation and algorithm for JADD in relation to (4.1) and (4.2) are discussed in Section 4.4.

4.4 ADMM based JADD for system models in (4.1) and (4.2)

The system models for UL CB-based grant-free NOMA are provided in (4.1) for the one-shot scenario and in (4.2) for the dynamic scenario, respectively. The transmit signal vector $\mathbf{x}$ in (4.1) is a concatenation of codewords from J users. In a typical mMTC scenario, only a small number of users are active, while the rest remain inactive and do not participate in transmission. This leads to the assumption on $\mathbf{x}$ in (4.1) be a block-sparse vector. Additionally, the BS has no prior knowledge of which users are active. The key challenge is to accurately detect the set of active users, followed by data detection, a process known as JADD. In CB-based NOMA systems, the transmit vector exhibits a block sparsity structure. A detailed discussion on block-sparsity signal structures presented in [96]. This block sparsity arises because the non-zero entries in the codewords are clustered, with each codeword (or block) corresponding to a specific user. The two main variants of CB-based NOMA, namely sparse CB-based NOMA (SCMA) with sparse codewords and dense CB-based NOMA (DCMA) with dense codewords, necessitate two distinct JADD formulations. This section presents an ADMM-based JADD method for both SCMA and DCMA systems.

4.4.1 JADD for one-shot block sparsity model

This subsection aims to design an MUD for JADD in the one-shot block sparsity system discussed in Section 4.3.1 for both sparse and dense CB-based NOMA systems.

4.4.1.1 JADD for sparse CD-NOMA System

SCMA is an effective example of a CB-based sparse NOMA system. This section presents the formulation of the MUD problem for SCMA systems in a grant-free scenario. As mentioned earlier, grant-free NOMA systems allow users to access the network without requiring any grant from the BS, allowing multiple users to share the same resources. Additionally, the BS has no prior knowledge of which users are active at any given time on the network. Consider a

CD-NOMA system with J users in total, where only J_a are active users ($J_a << J$) and the rest remain inactive. The SCMA system exhibits two types of sparsity: user-wise sparsity, since many users are inactive and do not transmit data, and within-user sparsity, as active users transmit sparse codewords with only a few nonzero entries. The detection problem for grant-free SCMA systems can be formulated as a sparse group LASSO problem. An ADMM algorithm is proposed to solve this sparse-group LASSO problem in a distributed manner. The sparse group LASSO for the SCMA system is defined as

$$\min_{\mathbf{x}} \quad \|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 + \underbrace{\alpha_1 \sum_{j=1}^{J} \|\mathbf{x}_j\|_2}_{\text{user wise sparsity}} + \underbrace{\alpha_2 \|\mathbf{x}\|_1}_{\text{within user sparsity}} \tag{4.3}$$

The problem (4.3) includes three terms. The first term represents the fidelity of the data, while the remaining two represent penalties that promote the sparse structure of $\mathbf{x}$. The first penalty is the mixed norm $\ell_{2,1}$, which encourages user- or block-wise sparsity, indicating active users in the system. The second penalty is the ℓ_1 norm, which promotes the sparsity within the user or element, referring to the sparse codewords in the SCMA system. These two penalty terms help find a solution with fewer non-zero entries while preserving the block sparsity structure. The parameters $\alpha_1 > 0$ and $\alpha_2 > 0$ represent the regularization (penalty) parameters for user activity and data detection, respectively, allowing control over the sparsity of the solution. The objective function in problem (4.3) is convex with two non-smooth penalty terms. It can be solved using a subdifferential calculus [97]. Optimal selection of penalty parameters improves the performance of the detector. However, the problem in (4.3) ignores the fact that the penalty imposed on active users should be lower than that for inactive users. The uniform penalization used in (4.3) results in performance degradation. Therefore, in this work, a nonuniform or reweighted penalization strategy is adopted to address this problem, as shown below:

$$\min_{\mathbf{x}} \quad \|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{x}_j\|_2 + \alpha_2 \|\mathbf{x}\|_{1,\mathbf{v}} \tag{4.4}$$

where $\|\mathbf{x}\|_{1,\mathbf{v}} = \sum_{i=1}^{JK} v_i |x_i|$, x_i is the ith entry of the vector $\mathbf{x}$. The weights w_j and v_i are selected as inversely proportional to the norm and absolute values of the $\mathbf{x}_j$ and x_i of the previous iteration,

respectively as follows

$$w_j^t = \frac{1}{\|\mathbf{x}_j^{t-1}\| + \epsilon} \quad v_i^t = \frac{1}{|x_i^{t-1}| + \epsilon}. \tag{4.5}$$

Here, ϵ is a small constant chosen to ensure numerical stability, and t indicates the ADMM iteration. According to (4.5), users and codeword entries with lower magnitude values in the previous iteration are penalized more in the current iteration. As the iterations progress, the penalties imposed on active users (first penalty) and nonzero entries (second penalty) gradually approach zero.

ADMM algorithm

The ADMM steps for the problem in (4.4) are easy to implement and have a guaranteed convergence property [70]. In the ADMM formulation, the variable $\mathbf{x}$ is split by introducing two auxiliary variables, $\mathbf{q}$ and $\mathbf{z}$. The resulting ADMM problem formulation is given by

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z},\mathbf{q}} \quad & \|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j\|_2 + \alpha_2 \|\mathbf{q}\|_{1,\mathbf{v}} \\ \text{s.t} \quad & \mathbf{x} = \mathbf{z}, \mathbf{x} = \mathbf{q}. \end{aligned} \tag{4.6}$$

The augmented Lagrangian function for (4.6) is given by

$$\begin{aligned} \mathcal{L}\left(\mathbf{x}, \{\mathbf{z}_j\}_{j=1}^J, \mathbf{q}, \mathbf{y}_1, \mathbf{y}_2\right) = \frac{1}{2}\|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j\|_2 + \alpha_2 \|\mathbf{q}\|_{1,\mathbf{v}} + \mathrm{Re}\langle \mathbf{y}_1, \mathbf{x} - \mathbf{z}\rangle + \mathrm{Re}\langle \mathbf{y}_2, \mathbf{x} - \mathbf{q}\rangle \\ + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|^2 + \frac{\rho}{2}\|\mathbf{x} - \mathbf{q}\|^2 \end{aligned} \tag{4.7}$$

where $\mathbf{x}, \mathbf{q}$, and $\mathbf{z}$ are the primal variables and $\mathbf{z} = [\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_J^T]^T$, $\mathbf{y}_1$ and $\mathbf{y}_2$ are the Lagrangian multipliers and $\rho > 0$ is the augmented Lagrangian penalty parameter. The ADMM update equations are given by

$$\begin{aligned} \mathbf{x}^{t+1} &:= \arg\min_{\mathbf{x}} \mathcal{L}\left(\mathbf{x}, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{q}^t, \mathbf{y}_1^t, \mathbf{y}_2^t\right) \\ &:= \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 + \mathrm{Re}\langle \mathbf{y}_1, \mathbf{x} - \mathbf{z}\rangle + \mathrm{Re}\langle \mathbf{y}_2, \mathbf{x} - \mathbf{q}\rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|^2 + \frac{\rho}{2}\|\mathbf{x} - \mathbf{q}\|^2 \end{aligned} \tag{4.8a}$$

$$\begin{aligned}\mathbf{q}^{t+1} &:= \arg\min_{\mathbf{q}} \mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{q}, \mathbf{y}_1^t, \mathbf{y}_2^t\right) \\ &:= \arg\min_{\mathbf{q}} \alpha_2 \|\mathbf{q}\|_{1,\mathbf{v}} + \mathrm{Re}\langle \mathbf{y}_2, \mathbf{x} - \mathbf{q}\rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{q}\|^2 \\ \mathbf{z}^{t+1} &:= \arg\min_{\mathbf{z}} \mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j\}_{j=1}^J, \mathbf{q}^{t+1}, \mathbf{y}_1^t, \mathbf{y}_2^t\right) \\ &:= \arg\min_{\mathbf{z}} \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j\|_2 + \mathrm{Re}\langle \mathbf{y}_1, \mathbf{x} - \mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|^2\end{aligned} \tag{4.8b}$$

$$\mathbf{y}_1^{t+1} := \mathbf{y}_1^t + \rho(\mathbf{x}^{t+1} - \mathbf{z}^{t+1}) \tag{4.8c}$$

$$\mathbf{y}_2^{t+1} := \mathbf{y}_2^t + \rho(\mathbf{x}^{t+1} - \mathbf{q}^{t+1}) \tag{4.8d}$$

Let $\mathbf{u}_1 = \frac{\mathbf{y}_1}{\rho}$ and $\mathbf{u}_2 = \frac{\mathbf{y}_2}{\rho}$, called the scaled dual variables [70]. The scaled version of ADMM steps are as follows:

$$\mathbf{x}^{t+1} := \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z} + \mathbf{u}_1\|^2 + \frac{\rho}{2}\|\mathbf{x} - \mathbf{q} + \mathbf{u}_2\|^2 \tag{4.9a}$$

$$\mathbf{q}^{t+1} := \arg\min_{\mathbf{q}} \alpha_2 \|\mathbf{q}\|_{1,\mathbf{v}} + \frac{\rho}{2}\|\mathbf{x} - \mathbf{q} + \mathbf{u}_2\|^2 \tag{4.9b}$$

$$\mathbf{z}^{t+1} := \arg\min_{\mathbf{z}} \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j\|_2 + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z} + \mathbf{u}_1\|^2 \tag{4.9c}$$

$$\mathbf{u}_1^{t+1} := \mathbf{u}_1^t + (\mathbf{x}^{t+1} - \mathbf{z}^{t+1}) \tag{4.9d}$$

$$\mathbf{u}_2^{t+1} := \mathbf{u}_2^t + (\mathbf{x}^{t+1} - \mathbf{q}^{t+1}). \tag{4.9e}$$

The augmented Lagrangian function in (4.7) is now decomposed into simpler optimization sub-problems, as shown in (4.9). The first three steps correspond to the optimization of the primal variables, while the remaining two steps involve updating the dual variables. The problem in (4.9c) can be simplified and solved in a distributed manner for $j = 1, \ldots, J$, and can equivalently be rewritten as:

$$\sum_{j=1}^{J} \min_{\mathbf{z}_j} \alpha_1 w_j \|\mathbf{z}_j\|_2 + \frac{\rho}{2}\|\mathbf{x}_j - \mathbf{z}_j + \mathbf{u}_{1j}\|^2. \tag{4.10}$$

The problem in (4.10) can be solved separately for each j with respect to the variable $\mathbf{z}_j$. The main challenge lies in computing the optimal solutions for the optimization problems in (4.9). For the problem in (4.9a), simplified from (4.8a), it is observed that $\mathcal{L}\left(\mathbf{x}, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{q}^t, \mathbf{y}_1^t, \mathbf{y}_2^t\right)$ is convex with respect to $\mathbf{x}$ when $\rho > 0$ and $\mathbf{H}^H\mathbf{H}$ is positive semidefinite (PSD). Furthermore, it

becomes strongly convex with respect to $\mathbf{x}$ when the condition $\mathbf{H}^H\mathbf{H}+\rho\mathbf{I} \geq (\lambda_{\min}(\mathbf{H}^H\mathbf{H})+\rho)\mathbf{I}$ is fulfilled according to [98,99]. For the problem in (4.9b), derived from (4.8b), $\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{q}, \mathbf{y}_1^t, \mathbf{y}_2^t\right)$ is convex with respect to the variable $\mathbf{q}$ when $\rho > 0$ and becomes strongly convex as ρ increases. Similarly, for the problem in (4.9c), derived from (4.8c), $\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j\}_{j=1}^J, \mathbf{q}^{t+1}, \mathbf{y}_1^t, \mathbf{y}_2^t\right)$ is convex with respect to the variable $\mathbf{z}$ when $\rho > 0$, and eventually becomes strongly convex due to its second term when $\rho > 0$. Therefore, the optimal solution for each of these optimization problems can be obtained by setting the gradient with respect to the corresponding variables equal to zero, as follows:

$$\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{x}, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{q}^t, \mathbf{y}_1^t, \mathbf{y}_2^t\right) = 0, \tag{4.11}$$

$$\partial_{\mathbf{q}}\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{q}, \mathbf{y}_1^t, \mathbf{y}_2^t\right) \in 0, \tag{4.12}$$

$$\partial_{\mathbf{z}_j}\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j\}_{j=1}^J, \mathbf{q}^{t+1}, \mathbf{y}_1^t, \mathbf{y}_2^t\right) \in 0. \tag{4.13}$$

where ∂ indicates the sub-gradient operator. The first terms in both (4.9b) and (4.9c) are nondifferentiable, which means that they must be addressed using subdifferential methods. Solutions to (4.9b) and (4.9c) can be obtained through scalar soft-thresholding and block soft-thresholding operations, commonly known as shrinkage operations [70].

The ADMM solutions for the problems (4.9a), (4.9b), and (4.9c) are as follows:

$$\mathbf{x}^{t+1} := (\mathbf{H}^{\mathrm{H}}\mathbf{H} + 2\rho\mathbf{I})^{-1}(\mathbf{H}^H\mathbf{y} + \rho(\mathbf{z}^k + \mathbf{q}^k - \mathbf{u}_1^k - \mathbf{u}_2^k)) \tag{4.14a}$$

$$q_i^{t+1} := \frac{(x_i^{t+1} + u_{2i}^t)}{|x_i^{t+1} + u_{2i}^t|}\left(|x_i^{t+1} + u_{2i}^t| - \frac{\alpha_2 v_i^t}{\rho}\right)_+, \quad \forall i \tag{4.14b}$$

$$\mathbf{z}_j^{t+1} := \frac{(\mathbf{x}_j^{t+1} + \mathbf{u}_{1j}^t)}{\|\mathbf{x}_j^{t+1} + \mathbf{u}_{1j}^t\|}\left(\|\mathbf{x}_j^{t+1} + \mathbf{u}_{1j}^t\| - \frac{\alpha_1 w_j^t}{\rho}\right)_+, \quad \forall j \tag{4.14c}$$

where (4.14b) and (4.14c) involve the scalar soft-thresholding and block soft-thresholding operations, respectively.

4.4.1.2 JADD for dense CD-NOMA system

The DCMA codewords are dense, and clustered nonzero values occur densely within the codeword. The JADD problem for the DCMA system is modified by keeping $\alpha_2 = 0$, which means that there is no sparsity within the user in the DCMA. The group LASSO-based JADD

problem, in this case, is given by

$$\min_{\mathbf{x}} \quad \|\mathbf{r}-\mathbf{H}\mathbf{x}\|^2 + \underbrace{\alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{x}_j\|_2}_{\text{user wise sparsity}}. \tag{4.15}$$

The problem (4.15) is similar to the problem in (4.4) if $\alpha_2 = 0$. The sparse-group LASSO problem is now converted into the group LASSO problem. The group LASSO problem helps compute the solution with a sparse set of blocks; if it includes a user (active) in the transmission, all block entries will be nonzero. By following the steps (4.6) and (4.7), the ADMM steps are given by:

$$\mathbf{x}^{t+1} := \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{r}-\mathbf{H}\mathbf{x}\|^2 + \mathrm{Re}\langle \mathbf{y}, \mathbf{x}-\mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}\|^2, \tag{4.16a}$$

$$\mathbf{z}^{t+1} := \arg\min_{\mathbf{z}} \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j\|_2 + \mathrm{Re}\langle \mathbf{y}, \mathbf{x}-\mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}\|^2, \quad \text{and} \tag{4.16b}$$

$$\mathbf{y}^{t+1} := \mathbf{y}^t + \rho(\mathbf{z}^{t+1} - \mathbf{x}^{t+1}). \tag{4.16c}$$

The three ADMM solutions are derived below following the similar procedure of ADMM for sparse group LASSO in Section 4.4.1.1 are given by:

$$\mathbf{x}^{t+1} := (\mathbf{H}^H\mathbf{H} + \rho\mathbf{I})^{-1}(\mathbf{H}^H\mathbf{y} + \rho(\mathbf{z}^k - \mathbf{u}^k)), \tag{4.17a}$$

$$\mathbf{z}_j^{t+1} := \frac{(\mathbf{x}_j^{t+1} + \mathbf{u}_{1j}^t)}{\|\mathbf{x}_j^{t+1} + \mathbf{u}_{1j}^t\|}\left(\|\mathbf{x}_j^{t+1} + \mathbf{u}_{1j}^t\| - \frac{\alpha_1 w_j^t}{\rho}\right)_+, \quad \forall j \tag{4.17b}$$

$$\mathbf{u}^{t+1} := \mathbf{u}^t + (\mathbf{x}^{t+1} - \mathbf{z}^{t+1}). \tag{4.17c}$$

Algorithm 3 outlines the key steps of the proposed group-LASSO-based JADD approach via the ADMM algorithm for the one-shot model. Additionally, **Algorithm** 4 provides a detailed procedure for a robust prior information-aided ADMM applied to a frame-wise dynamic model. In [100], a threshold-aided support detection scheme is proposed for the sparse recovery problem, based on the principle of first significant jump (FSJ). This concept is applied in **Algorithm** 3 for support detection when the true sparsity level is unknown. Subsequently, it was adapted for **Algorithm** 4, which is designed for dynamic scenarios. The FSJ occurs within the

Algorithm 3 Group LASSO based JADD via FSJ aided ADMM algorithm for Model-1

Input: $\mathbf{r}$,$\mathbf{H}$, and maximum number of iterations (T).
Output: $\hat{\Lambda}$, and $\hat{\mathbf{x}}$.
Initialization: Initialize the vectors $\mathbf{z}$,$\mathbf{q}$,and $\mathbf{u}_1$ with zeros.
Pre-processing
Compute $(\mathbf{H}^H\mathbf{H} + 2\rho\mathbf{I})^{-1}$.
Compute $\mathbf{H}^H\mathbf{r}$.
for $t = 1, 2, , \ldots T$ **do**
    Step:1 Update $\mathbf{x}^{(t+1)}$ via (4.17a),
    Step:2 Update $\mathbf{z}_j^{(t+1)}$ via (4.17b), for $j = 1, 2, \ldots, J$
    Step:3 Update $\mathbf{u}_1^{(t+1)}$ via (4.17c),
(Computing primal residuals)

$$\mathbf{r}_p^{(t+1)} = \mathbf{x}^{(t+1)} - \mathbf{z}^{(t+1)}, \quad \|\mathbf{r}_p^{(t+1)}\|_2$$

(Computing dual residual)

$$\mathbf{r}_d^{(t+1)} = \rho(\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}) \quad \|\mathbf{r}_d^{t+1}\|_2$$

$$\epsilon^{\text{pri}} = \sqrt{J * K}\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\{\|\mathbf{x}^{(\mathbf{t+1})}\|_2, \|\mathbf{z}^{(\mathbf{t+1})}\|_2\}$$

$$\epsilon^{\text{dual}} = \sqrt{J * K}\epsilon^{\text{abs}} + \epsilon^{\text{rel}}\|\rho\mathbf{u}_1^{(t+1)}\|$$

    if $\|\mathbf{r}_p^{(t+1)}\|_2 \leq \epsilon^{\text{pri}} \wedge \mathbf{r}_d^{(t+1)} \leq \epsilon^{\text{dual}}$ **then**
    Stop the iteration
    end if
end for
Post-iteration processing
(Computing FSJ)
$\widetilde{\mathbf{x}} = [\|\mathbf{x}_1^{(T)}\|_2, \|\mathbf{x}_2^{(T)}\|_2, \ldots, \|\mathbf{x}_J^{(T)}\|_2]$
First sort the elements of $\widetilde{\mathbf{x}}$ in ascending order,

$$\widetilde{\mathbf{x}}_{\text{sort}} = \text{sort}(\widetilde{\mathbf{x}})$$

i.e. $\widetilde{\mathbf{x}}_{\text{sort}}[1] \leq \widetilde{\mathbf{x}}_{\text{sort}}[2] \leq, \ldots, \leq, \widetilde{\mathbf{x}}_{\text{sort}}[J]$, where $\widetilde{\mathbf{x}}_{\text{sort}}[p]$ indicates the p-th largest entry in $\widetilde{\mathbf{x}}_{\text{sort}}$.
The FSJ computed from the consecutive difference of sorted $\hat{\mathbf{x}}_{\text{sum}}$ is given by

$$\widetilde{\mathbf{x}}_{\text{sort}}[p+1] - \widetilde{\mathbf{x}}_{\text{sort}}[p] > \frac{\alpha\|\widetilde{\mathbf{x}}\|_\infty}{N_{\text{r}}} \tag{4.18}$$

The smallest p in (4.18) indicates the FSJ, then $\beta = \widetilde{\mathbf{x}}_{\text{sort}}[p]$. and the partial support set

$$\mathcal{T} = \{j : \widetilde{\mathbf{x}}_j > \beta\} \tag{4.19}$$

Sparsity level $\hat{\text{S}}_1 = |\mathcal{T}|$.
(Computation of final support set)
$\widetilde{\mathbf{x}} = [\|\mathbf{x}_1\|_2, \|\mathbf{x}_2\|_2, \ldots, \|\mathbf{x}_J\|_2]$
Find out the set of $\hat{\text{S}}_1$ user indices in $\widetilde{\mathbf{x}}$, whose $\|\mathbf{x}_j\|_2$ values are largest among $j = 1, 2, \ldots, J$ is indicated by $\Xi(\widetilde{\mathbf{x}}, \hat{\text{S}}_1)$
(Final detected support set)
$\hat{\Lambda} = \Xi(\widetilde{\mathbf{x}}, \hat{\text{S}}_1)$.
(Final estimated signal information)
$\hat{\mathbf{x}} = \mathbf{x}^{(T)}[\{1, 2, \ldots, J\} \setminus \hat{\Lambda}] = 0$.

estimated sequence, where the true nonzero elements are large in magnitude and few in number, whereas the false ones are small in magnitude and more numerous. This creates a distinct boundary between the true non-zero and false non-zero elements.

Based on primal and dual feasibility, the necessary and sufficient optimality conditions can be applied to the ADMM detection problem [70]. This leads to two key quantities for checking residual convergence: primal and dual residuals. For problem (4.15), the primal residual is given by $\mathbf{r}_p^{(t)} = \mathbf{x}^t - \mathbf{z}^t$ and the dual residual by $\mathbf{r}_d^{(t)} = \rho(\mathbf{z}^t - \mathbf{z}^{t+1})$. These residuals also serve as stopping criteria in the **Algorithm** 3. The ADMM terminates, signaling convergence when the residuals fall below the feasibility tolerances ϵ^{pri} and ϵ^{dual}. Here, $\epsilon^{\text{abs}} > 0$ and $\epsilon^{\text{rel}} > 0$ are the absolute and relative tolerance constants, respectively. Values of ϵ^{abs}=10^{-3} and ϵ^{rel}=10^{-3} or

10^{-4} are commonly effective as stopping criteria [70].

4.4.1.3 Therotical convergence of Algorithm 3

The convergence of the ADMM algorithm for convex problems is quite extensively investigated. It is worth mentioning that the ADMM is known to converge with a mild selection of the parameter ρ.

Theorem 4.1. *Suppose ρ is large enough and that conditions $\rho > 0$, $\rho(\rho + \lambda_{\min}(\mathbf{H}^H\mathbf{H})) \geq 2\lambda^2_{\max}(\mathbf{H}^H\mathbf{H})$, and $\rho > \lambda_{\max}(\mathbf{H}^H\mathbf{H})$ are satisfied. Then the sequence $\{\mathbf{x}^t, \{\mathbf{z}_j^t\}_{j=1}^J, \mathbf{y}^t\}$ generated by ADMM steps in (4.17) is convergent as given below*

$$\lim_{t\to\infty} \mathbf{x}^t = \mathbf{x}^\star, \quad \lim_{t\to\infty} \mathbf{z}_j^t = \mathbf{z}_j^\star, \forall j \quad \lim_{t\to\infty} \mathbf{y}^t = \mathbf{y}^\star. \tag{4.20}$$

Proof. Provided in Appendix B.1. □

Remark 1: *The above theorem proves the convergence of the proposed ADMM algorithm for the group LASSO problem formulated for DCMA systems. The conditions on ρ given for problem (4.15) holds for problem (4.3). The theorem proves that the proposed ADMM detection algorithm for the Group LASSO-based detection problem is guaranteed to converge to some stationary point under mild conditions on selecting the ADMM penalty parameter ρ. The parameters are selected based on the given channel matrix* $\mathbf{H}$.

4.4.2 JADD for frame-wise dynamic block sparsity model

The sparse group LASSO and group LASSO problems are formulated for MUD in one-shot grant-free SCMA and DCMA systems, respectively, in Section 4.4.1. Furthermore, an ADMM-based detection algorithm is developed to jointly recover user activity and data for the one-shot block sparsity model. This section addresses JADD in a more practical scenario by considering the frame-wise dynamic system model introduced in Section 4.3.2. The ADMM-based detection algorithm is adapted for this dynamic system model. Unlike (4.4) and (4.15), which do not require prior knowledge of the active user set and signal estimate due to one-shot transmission, the dynamic scenario benefits from incorporating prior information. An effective method to utilize prior knowledge of both the support and signal in sparse recovery is proposed in [101]. Later, this method is enhanced to enable joint processing of observations within a static frame-wise model, resulting in improved performance for MUD [74]. The prior information-aided group LASSO is formulated for DCMA systems to address the multi-user detection problem in Case-2, as discussed in Section 4.3.2, which pertains to dynamic scenarios.

Algorithm 4 Robust prior-information aided JADD for frame-wise dynamic block-sparsity model via FSJ-ADMM

Input: $\mathbf{r}^{[l]}$, $\mathbf{H}^{[l]}$ for $l = 1, 2, \ldots, L.$, and T.
Output: $\hat{\Lambda}^{[l]}$, and $\hat{\mathbf{x}}^{[l]}$ for $l = 1, 2, \ldots, L$.
Initialization: Initialize the vectors $\mathbf{z}$, and $\mathbf{u}$ with zeros.
Step-i: Repeat **Algorithm** 3 for $l = 1, 2, \ldots, L$ independently and compute $\hat{\Lambda}_{\text{p1}}^{[l]}$ and $\hat{\mathbf{x}}_{\text{p}}^{[l]}$, and $I_{\text{p}}^{[l]}, \forall l$.

$$\hat{\mathbf{X}} = [\hat{\mathbf{x}}^{[1]}, \hat{\mathbf{x}}^{[2]}, \ldots, \hat{\mathbf{x}}^{[L]}]$$

Firstly perform operation in (4.22)
if $\hat{\mathbf{X}} \in$ Case-1 in (4.22) **then**
Stop the algorithm
else
Continue to **Step-ii**
Step-ii:
if $\hat{\mathbf{X}} \in$ Case-2a in (4.23) **then**
Apply Algorithm 3 to the problem given in (4.24) and compute the $\hat{\Lambda}_{\text{p2}}^{[l]}$ for $l = 1, 2, \ldots, L$.
(Compute the final active support set)
$\hat{\Lambda}_{\text{fin}}^{[l]} = \Xi(\widetilde{\mathbf{x}}, \hat{\mathrm{S}}_{\text{l}}^{[l]})$.
(Final estimated signal information)
$\hat{\mathbf{x}}_{\text{fin}}^{[l]} = \mathbf{x}^{(T)}[\{1, 2, \ldots, J\} \setminus \hat{\Lambda}_{\text{fin}}^{[l]}] = 0$, for $l = 1, 2, \ldots, L$.
else if $\hat{\mathbf{X}} \in$ Case-2b in (4.23) **then**
Apply Algorithm 3 to the problem given in (4.26) by incorporating the corresponding solutions given in (4.27) for $l = 1, 2, \ldots, L$.
(Compute the final active support set)
$\hat{\Lambda}_{\text{fin}}^{[l]} = \Xi(\widetilde{\mathbf{x}}, \hat{\mathrm{S}}_{\text{l}}^{[l]})$.
(Final estimated signal information)
$\hat{\mathbf{x}}_{\text{fin}}^{[l]} = \mathbf{x}^{(T)}[\{1, 2, \ldots, J\} \setminus \hat{\Lambda}_{\text{fin}}^{[l]}] = 0$, for $l = 1, 2, \ldots, L$.
end if
end if

A similar procedure applies to the sparse group LASSO detection problem in (4.3) for SCMA systems. In the dynamic system model, the set of active users changes randomly across time slots and, with high probability, shows a temporal correlation between consecutive time slots. Performing MUD independently for each time slot, as in **Step-i** of **Algorithm** 4, can lead to performance degradation. Incorporating this temporal correlation into the ADMM iterations can significantly enhance the performance of the proposed MUD.

There are two cases in the dynamic system model discussed in Section 4.3.2. Note that the first case involves user activity changing at a very fast rate, while the second case occurs when user activity changes at a slow or moderate rate. These two scenarios lead to different approaches for JADD in the dynamic system model.

Case-1: This can be considered the worst-case scenario, which occurs when user activity

changes at a very fast rate. In this case, the user activity is not shared between adjacent time slots, meaning there is no temporal correlation between the active user sets across consecutive time slots within a frame. As a result, the observations in each time slot can be processed independently as follows:

$$\min_{\mathbf{x}^{[l]}} \quad \|\mathbf{r}^{[l]} - \mathbf{H}^{[l]}\mathbf{x}^{[l]}\|^2 + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{x}_j^{[l]}\|_2, \quad \forall l \tag{4.21}$$

Case-2: In this scenario, user activity changes slowly within a frame, leading to a high probability that common active users will be present across adjacent time slots. This creates a temporal correlation between the active user sets in consecutive time slots. Consequently, ADMM-aided JADD can be performed in two key steps, as discussed below.

Step-i: Apply Algorithm 3 independently to (4.21) for each time slot $l = 1, 2, \ldots, L$, to determine the active user support set $\Lambda_{\mathrm{p1}}^{[l]}$ and signal estimate $\hat{\mathbf{x}}_{\mathrm{p}}^{[l]}$. From $\hat{\mathbf{x}}_{\mathrm{p}}^{[l]}$, the codewords or signal index value set $I_{\mathrm{p}}^{[l]}$ for $\Lambda_{\mathrm{p1}}^{[l]}$ can be easily extracted. Here, $\hat{\mathbf{x}}_{\mathrm{p}}^{[l]}$ provides the active user signal information, while $I_{\mathrm{p}}^{[l]}$ represents the index values of the active user signals transmitted during the lth time slot. The subscript 'p' signifies the prior information extracted from **Step-i** for use in **Step-ii**.

After performing **Step-i**, compute the common support set between the adjacent time slots as follows

$$\begin{aligned} \Lambda_{\mathrm{p1}}^{[l]} \cap \Lambda_{\mathrm{p1}}^{[l-1]} &= \emptyset, \quad \hat{\mathbf{X}} \in \textbf{Case-1} \\ &\neq \emptyset, \quad \hat{\mathbf{X}} \in \textbf{Case-2}, \quad \forall l \end{aligned} \tag{4.22}$$

Note that $\hat{\mathbf{X}} = [\hat{\mathbf{x}}^{[1]}, \hat{\mathbf{x}}^{[2]}, \ldots, \hat{\mathbf{x}}^{[L]}]$.

If $\hat{\mathbf{X}} \in$ **Case-1**, which represents the worst-case scenario, there are no common users found between the adjacent time slots. Consequently, the active user support set $\hat{\Lambda}_{\mathrm{p1}}^{[l]}$ and the signal estimate $\hat{\mathbf{x}}_{\mathrm{p}}^{[l]}$ calculated independently from (4.21), for $l = 1, 2, \ldots, L$, are the final estimates for the **Case-1** .

If $\hat{\mathbf{X}} \in$ **Case-2**, one can perform **Step-ii** in two ways, as given below. Compute the common

signal index values between the adjacent time slots as follows

$$\begin{aligned} I_{\mathrm{p}}^{[l]} \cap I_{\mathrm{p}}^{[l-1]} &= \emptyset, \quad \hat{\mathbf{X}} \in \textbf{Case-2a} \\ &\neq \emptyset, \quad \hat{\mathbf{X}} \in \textbf{Case-2b}, \quad \forall l. \end{aligned} \tag{4.23}$$

Step-ii: The temporal correlation exists between the active user sets and the signal information between the adjacent time slots in a frame, which can be used in two ways to enhance performance. As discussed in Section 4.3.2, the JADD of two sub-cases within **Case-2**, are outlined below.

Case-2a: Symbol duration<time slot

In this scenario, we assume that the symbol duration is shorter than the time slot. As a result, the symbols transmitted by the active users in adjacent time slots are different. The temporal correlation exists only through the sets of active user sets across the adjacent time slots, while the symbols transmitted by these users within the frame are independent. Therefore, the active user support set computed in the previous time slot can be utilized to compute the active user support set in the current time slot.

Active user sets are temporally correlated and dynamic throughout the frame. The user can come and go at any time in the frame. Consequently, blindly using the active set in the previous time slot leads to a degradation in performance [86]. Instead, we can compute the partial information of the lth time interval from the $(l-1)$th time interval. It is important to note that neither the active user sets nor the signal values are known at the receiver to exploit the temporal correlation. Therefore, the proposed algorithm uses the computed support set and signal values from **Step i** as partial information for **Step ii**.

If $\hat{\mathbf{X}} \in$ Case-2a, it indicates that some common active users exist between the time slots. The algorithm then proceeds to execute **Step-ii** by incorporating the partial information obtained in **Step-i** as prior knowledge as follows

$$\min_{\mathbf{x}^{[l]}} \quad \|\mathbf{r}^{[l]} - \mathbf{H}^{[l]}\mathbf{x}^{[l]}\|^2 + \alpha_1 \sum_{j=1}^{J} w_{j,p}^{l} w_{j,d} \|\mathbf{x}_j^{[l]}\|_2$$

$$, l = 1, 2, \ldots, L \tag{4.24}$$

where $w^{l}_{j,p}$ is the weight associated with the prior information of the jth user in lth time slot and $w_{j,d}$ is the weight associated with democratic penalization, which is the same as given in (4.5). $w^{t}_{j,p}$ can be defined as

$$\begin{aligned} w^{l}_{j,p} &= 0, \quad j \in q^{[l]} \\ &= 1, \quad j \notin q^{[l]} \end{aligned} \tag{4.25}$$

where $q^{[l]} = \Lambda_{\mathrm{p2}}^{[l-1]} \cap \Lambda_{\mathrm{p1}}^{[l]}$ represents the quality set, the temporally correlated active user set between $(l-1)$th and lth time slots. Note that $\Lambda_{\mathrm{p1}}^{[l]}$ represents the active user support set for the lth (current) time slot as determined in Step-i, while $\Lambda_{\mathrm{p2}}^{[l-1]}$ corresponds to the active user support set for the $(l-1)$th time slot as determined in Step-ii (current step). The quality set $q^{[l]}$ provides reliable prior information for the lth time slot. Hence, it improves the reconstruction quality on the lth time slot. The term $w^{l}_{j,p}$ ensures the block-wise sparse vector $\mathbf{x}$ remains sparsest outside the set $q^{[l]}$ in lth time slot, while $w_{j,d}$ ensures the democratic penalization for the users, which are outside the quality support set $q^{[l]}$. Consequently, the second term in (4.24) does not penalize the active users whose locations are already known, i.e., those in $q^{[l]}$.

Case-2b: Symbol duration>time slot

Consider a scenario in which the duration of the symbol exceeds the time slot of a frame, as shown in Case 2b of Fig. 4.1. In this case, it is highly likely that the slowly varying active user transmits the same information across multiple adjacent time slots within the frame. As a result, the temporal correlation between both the active user sets and the signal estimates across time slots further enhances the reconstruction accuracy.

If **X** falls under Case-2b, the quality information for the current time slot includes both the partial support set and the signal estimate values of the previous time slot. Consequently, the group LASSO problem in (4.24) from **Step-ii** of Case-2a is modified accordingly. By incorporating the partial signal information corresponding to the quality support set $q^{[l]}$, the group

LASSO problem for Case-2b is formulated as follows:

$$\min_{\mathbf{x}^{[l]}} \quad \|\mathbf{r}^{[l]} - \mathbf{H}^{[l]}\mathbf{x}^{[l]}\|^2 + \alpha_1 \sum_{j=1}^{J} w_{j,p}^{l} w_{j,d} \|\mathbf{x}_j^{[l]}\|_2 + \frac{\mu^l}{2}\|\mathbf{x}^{[l]} - \hat{\beta}_{q^{[l]}}^{[l-1]}\|^2, l = 1, 2, \ldots, L \tag{4.26}$$

where $\hat{\beta}_{q^{[l]}}^{[l-1]}$ is the partial signal information of the lth time slot taken from the $(l-1)$th time slot. It contains non-zero values for users of the quality support set, $q^{[l]}$. The nonzero values indicate the signal estimates taken from the previous time slot. The factor μ^l is defined as $\mu^l := C/(\|\mathbf{r}^{[l]} - \mathbf{H}^{[l]}\hat{\beta}_{q^{[l]}}^{[l-1]}\|)$, ensures a larger μ^l when the signal estimate is accurate. Here C is a small constant chosen empirically. The second term in (4.26) ensures that the computed signal remains sparsest outside the set $q^{[l]}$. The third term in (4.26) enforces that the estimated signal stays sufficiently close to $\hat{\beta}_{q^{[l]}}^{[l-1]}$.

By following the similar procedure given in Section 4.4.1.1, the ADMM solutions derived for the problem (4.26) are given by

$$\mathbf{x}^{[l],t+1} := (\mathbf{H}^{[l]^{\mathrm{H}}}\mathbf{H}^{[l]} + (\mu^l + \rho)\mathbf{I})^{-1}(\mathbf{H}^{[l]^{\mathrm{H}}}\mathbf{r}^{[l]} + \mu^l \hat{\beta}_{q^{[l]}}^{[l-1]} + \rho(\mathbf{z}^{[l],t} - \mathbf{u}^{[l],t})) \tag{4.27a}$$

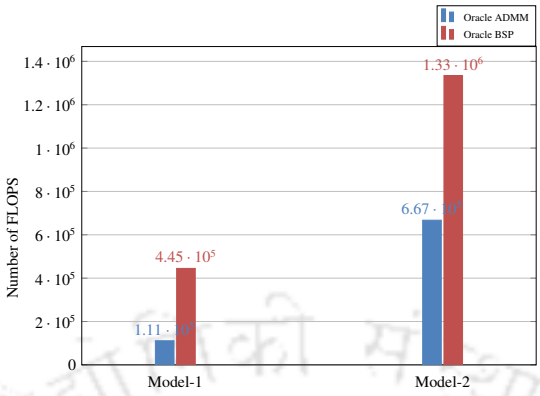
$$\mathbf{z}_j^{[l],t+1} := \frac{\mathbf{p}_j}{\|\mathbf{p}_j\|_2}\left(\|\mathbf{p}_j\|_2 - \frac{\alpha_1 w_{j,d} w_{j,p}^{l}}{\rho}\right)_+, \quad \forall j \tag{4.27b}$$

$$\mathbf{u}^{[l],t+1} := \mathbf{u}^{[l],t} + (\mathbf{x}^{[l],t+1} - \mathbf{z}^{[l],t+1}) \tag{4.27c}$$

where, $\mathbf{p}_j = \mathbf{x}_j^{[l],t+1} + \mathbf{u}_j^{[l],t}$.

4.5 Computational Complexity

This section examines the computational complexity of the proposed ADMM-based detection algorithms, as outlined in **Algorithm** 3 and **Algorithm** 4. **Algorithm** 4 builds upon **Algorithm** 3, requiring a greater number of computations to effectively address the practical challenges of MUD in dynamic environments. The computational complexity is evaluated in terms of the number of FLOPs or complex multiplications. The required number of FLOPs for executing these algorithms is discussed in the following: **Algorithm** 3 consists of three parts: pre-processing, iterative processing, and post-processing. The pre-processing includes matrix multiplications and matrix inverse operations such as $\mathbf{H}^H\mathbf{H}$, $\mathbf{H}^H\mathbf{r}$, and $(\mathbf{H}^H\mathbf{H} + 2\rho I)^{-1}$.



h!

Fig. 4.2: Computational complexity comparison of various detectors.

The size of the **H** matrix is $N_{\mathrm{r}}K \times JK$. The number of FLOPs required to perform these three operations is $(N_{\mathrm{r}}K)(KJ)^2$, $(N_{\mathrm{r}}K)(KJ)$, and $(KJ)^3$, respectively. The iterative processing steps are iteration-dependent and must be repeated in each iteration. The computations in **Step:1** in (4.17a) include scalar multiplication with a vector of size $KJ \times 1$ and the multiplication of a matrix of size $KJ \times KJ$ with a vector of size $KJ \times 1$. The number of FLOPs required to perform these computations is $(KJ) + (KJ)^2$. The **Step:2** in (4.17b) needs to be performed for $j = 1, 2, \ldots, J$, and which includes the multiplication of $K \times 1$ vector with a scalar. The number of FLOPs required to perform **Step:2** is KJ. **Step 3** and the post-iteration processing in Algorithm 3 do not involve any complex multiplications and contribute negligible complexity. Consequently, the total number of FLOPs required for the iterative processing of **Algorithm** 3 is: $(N_{\mathrm{r}}K)(KJ)^2 + (KJ)^3 + (N_{\mathrm{r}}K)(KJ) + T(2KJ + (KJ)^2)$.

Algorithm 4 proposed for a frame-wise dynamic block-sparsity model must perform the **Algorithm** 3 for $l = 1, 2, \ldots, L$ time slots. **Algorithm** 4 addresses the critical issues in JADD of dynamic scenarios and is performed in two steps. **Step-i** performs independent JADD and **Step-ii** performs the JADD by incorporating the prior information computed from **Step-i**. The number of FLOPs required to perform in **Step-i** of the **Algorithm** 4 are $L((N_{\mathrm{r}}K)(KJ)^2 + (KJ)^3 + (N_{\mathrm{r}}K)(KJ) + T((KJ) + (KJ)^2))$. The prior information aided ADMM iterations in **Step-ii** further improve the performance at the cost of increasing the complexity. In **Step-ii** , **Algorithm** 3 needs to be performed a second time for $l = 1, 2, \ldots, L$ by incorporating the prior information computed from **Step-i** . It can be observed in **Algorithm** 4 that if $\hat{\mathbf{X}} \in$ Case-1, then the algorithm stops after **Step:i**. Hence, the complexity is the same as that of **Step: i**. If $\hat{\mathbf{X}} \in$

Case-2, then the algorithm performs **Step:i** and **Step:ii**. Hence, the total number of FLOPs for **Algorithm** 4 with total T iterations with frame size of L is given by $2L((N_rK)(KJ)^2 + (KJ)^3 + (N_rK)(KJ) + T((KJ) + (KJ)^2))$.

The BSP algorithm tackles the JADD problem with a static block sparsity model in the frame, as outlined in [88]. It serves as a reference for comparison. The key implementation steps and computational complexity analysis are provided in [88], with the required FLOPs for the Oracle BSP algorithm summarized here. **Support estimate:** The support estimation step requires $(J(K^2N_r) + JK + 2J)$ FLOPS. **Least squares estimation (LSE):** The number of FLOPs required for LSE is $(|\tilde{T}|K)^3 + 2KN_r(|\tilde{T}|K)^2$, where $\tilde{T}$ represents the merged set of support obtained from the estimation of the support. **Support pruning:** Support pruning involves $(KJ+J)$ FLOPs. **Signal estimation:** The signal estimation requires $(J_aK)^3 + 2KN_r(J_aK)^2$ FLOPs, where J_a is the size of the true support set. **Residue computation:** Residue computation involves $(KN_r)(KJ)$ FLOPs. The computational complexity of the iterative BSP algorithm, considering a total of T_{BSP} iterations, is given by:$((J(K^2N_r) + JK + 2J) + ((|\tilde{T}|K)^3 + 2KN_r(|\tilde{T}|K)^2) + (KJ + J) + ((J_aK)^3 + 2KN_r(J_aK)^2) + (KN_r)(KJ))T_{BSP}$.

The prior information-aided adaptive subspace pursuit algorithm (PIA-ASP) [87] can be applied to the frame-wise dynamic block-sparsity model. PIA-ASP serves as a suitable baseline for comparison with the proposed algorithm. In PIA-ASP, JADD is performed in a single step. The algorithm repeats the BSP iterations in each time interval with a frame size of L. As a result, the computational complexity is expressed as:$((J(K^2N_r) + JK + 2J) + ((|\tilde{T}|K)^3 + 2KN_r(|\tilde{T}|K)^2) + (KJ + J) + ((J_aK)^3 + 2KN_r(J_aK)^2) + (KN_r)(KJ))T_{BSP}L$.

The main complexity in both the proposed ADMM and baseline BSP algorithms arises from matrix inversion. In ADMM, this occurs only once in preprocessing, whereas BSP requires it in every iteration, increasing computational cost. For the one-shot model, the number of FLOPs required to perform the ADMM and BSP algorithms is 1.11×10^5 and 4.45×10^5, respectively. For the frame-wise dynamic model, the number of FLOPs required to perform ADMM and PIA-ASP is 6.67×10^5 and 1.33×10^6, respectively. The values of the parameters used to compute the FLOPs are provided in Table 4.1. A comparison of the computational complexity

of these two algorithms is illustrated in Fig. 4.2.

4.6 Simulations and discussions

The numerical results of the proposed detection algorithm for various CB-based grant-free system models are discussed in this section. The simulation parameters are listed in Table 4.1. The fixed J_a given in Table 4.1 applies to the one-shot model, while in the frame-wise dynamic model, J_a varies across the time slots.

Table 4.1: Simulation parameters.

Parameter	Value
Number of potential UEs (J)	8
Number of active UEs (J_a)	3
Number of resources (K)	4
Modulation order (M)	4
Frame size (L)	7
Number of receive antenna (N_r)	16
Overloading factor (λ)	200 %
Number of iterations (T)	10

4.6.1 SER performance: One-shot block sparsity model

This subsection presents the simulation results of the proposed detection algorithm for the one-shot block sparsity model in the uplink scenario. Fig. 4.3 compares the performance of the symbol error rate (SER) of various detection schemes for DCMA, SCMA, and large-scale sequence-based NOMA systems.

Fig. 4.3 (a)illustrates the SER performance for DCMA systems. Oracle LSE is used as a reference for comparison. Oracle LSE assumes that the true support set is known at the receiver and uses LSE to estimate the signal information on this known support set. Oracle ADMM, on the other hand, assumes that the sparsity level S_1 is known to the receiver and performs ADMM iterations, as outlined in **Algorithm** 3, to jointly estimate both activity and signal information. From Fig. 4.3, we can observe that Oracle ADMM achieves performance comparable to that of Oracle LSE. The ADMM algorithm aided by FSJ performs JADD without prior knowledge of S_1, resulting in some performance degradation compared to Oracle methods. However, the adapted FSJ technique still provides competitive performance. However, the FSJ-aided BSP performs worse than the FSJ-aided ADMM. The detector based on FSJ-ADMM achieves an

SNR gain of approximately 1.2 dB at an SER of 10^{-3} compared to the FSJ-BSP algorithm, as shown in Fig. 4.3 (a). Fig. 4.3 (b) and Fig. 4.3 (c) present the SER performance comparison of different detection schemes for SCMA and sequence-based NOMA systems, respectively, where similar trends and observations hold as in the results of the DCMA system. Fig. 4.3 (c) highlights two key insights: (1) CB-based grant-free NOMA (DCMA and SCMA) significantly outperforms sequence-based NOMA, and (2) the proposed detector is efficient in large-scale settings with $J = 108$ and a spreading sequence of length $N = 72$.

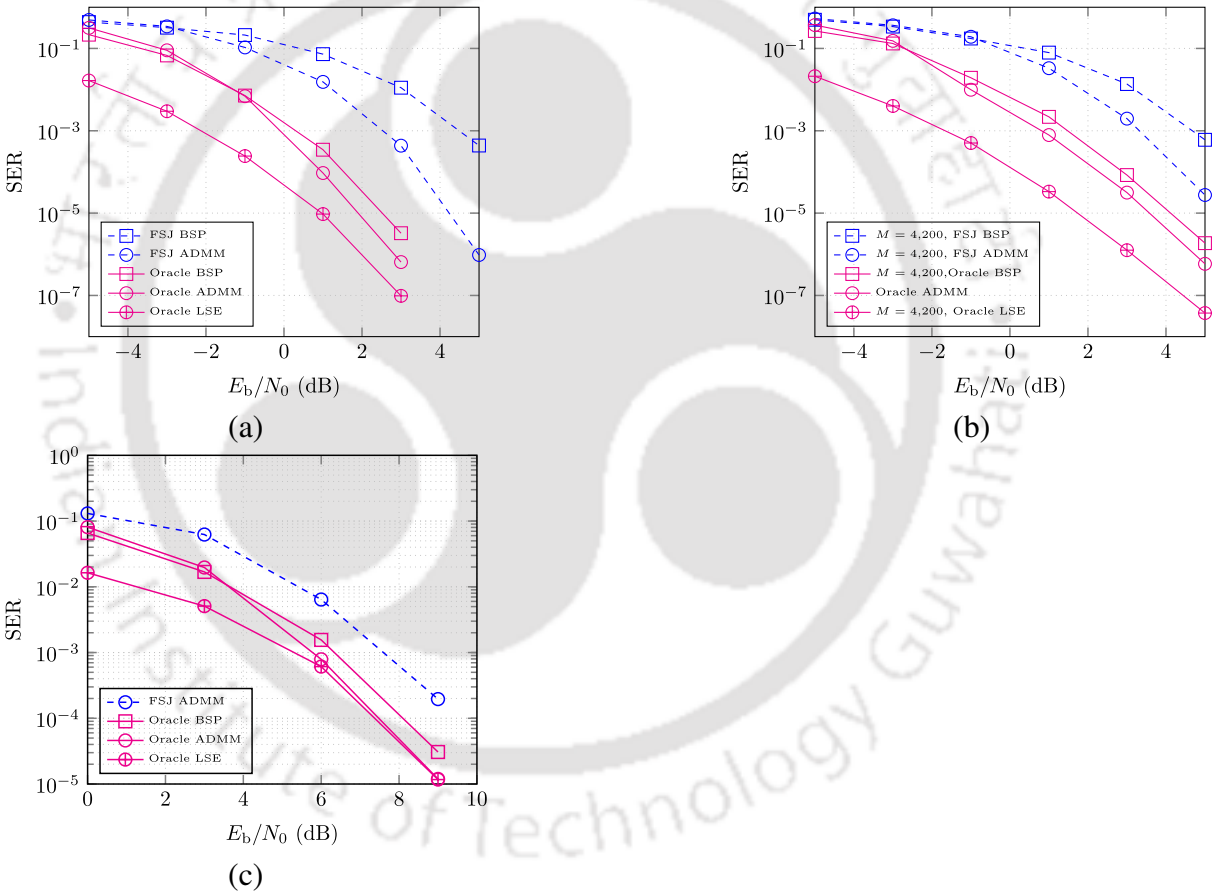


Fig. 4.3: SER performance of one shot static CD-NOMA systems (a) DCMA system with J=8 and K=4, (b) SCMA system with J=8 and K=4, (c) SS-based NOMA system with J=108 and N=72.

4.6.2 SER performance: Frame-wise dynamic block sparsity model

This subsection presents a series of numerical experiments on the proposed frame-wise dynamic block sparsity model. The average SER performance is analyzed on the basis of this model, highlighting the advantages of various practical use cases and their respective detection methodologies discussed in previous sections. Fig. 4.4 illustrates the comparison of SER

performance for different scenarios in an uplink DCMA system under dynamic conditions. In Fig. 4.4 (a), the SER performance curves for Algorithm 4 are shown in various cases. It is clear that the performance of Case-1 is inferior to that of the other cases. This is due to the lack of temporal correlation between the active user sets across adjacent time slots in Case-1, resulting in a lack of quality information for collaborative processing of time slots. In this work, Case-1 is considered the worst-case scenario. Conversely, Case-2, while possessing quality information, performs similarly to Case-1 if this information is not utilized, making its performance comparable to the worst-case scenario. Moreover, Fig. 4.4 (a) demonstrates the advantages of leveraging quality information for improved performance. In Case-2a, quality information is represented by common sets of active users in time slots. Using this quality information, Case-2a achieves an SNR gain of approximately 1.2 dB at an SER of 10^{-4} compared to Case-2, which lacks quality information. Case 2 without quality information represents a blind approach to exploiting temporal correlation, similar to the idea presented in [74]. In contrast, Case-2b incorporates quality information in the form of both common active user sets and their corresponding signal details. Consequently, the enhanced quality information in Case-2b results in superior performance compared to all other cases.

Fig. 4.4 (b) compares the SER performance of different detection schemes for Case 2. This includes Oracle ADMM's and FSJ-aided ADMM's performance, both with and without quality information. In both ADMM schemes, the inclusion of quality information significantly enhances performance compared to the versions without it. The SER performance of the two ADMM schemes with quality information is much closer to the benchmark performance than their counterparts without quality information. Furthermore, FSJ-aided ADMM with quality information achieves an SNR gain of nearly 2 dB at an SER of 10^{-5} over FSJ-aided ADMM without quality information. Furthermore, Fig. 4.4 (b) includes the performance of existing PIA-ASP algorithm in [87]. The performance of PIA-ASP is almost similar to that of the FSJ ADMM without quality-assisted ADMM. The quality-assisted FSJ ADMM outperforms the PIA-ASP scheme by nearly 1.9 dB in SNR gain at an SER of approximately 10^{-5}. Although the PIA-ASP scheme improves detector performance by incorporating the quality parameter (s_p),

it assumes that this parameter is inaccurately known to the receiver, limiting its effectiveness. Furthermore, s_p in PIA-ASP represents the size of the common active users between two consecutive time slots. In contrast, the proposed MUD leverages quality information in the form of both partial active user support sets $q^{[l]}$ and signal information $\hat{\beta}_{q^{[l]}}$. Rather than assuming that the quality information is known to the receiver, it is computed from **Step-i**. Fig. 4.4 underscores the importance of using quality information in ADMM iterations to enhance reliability in practical scenarios.

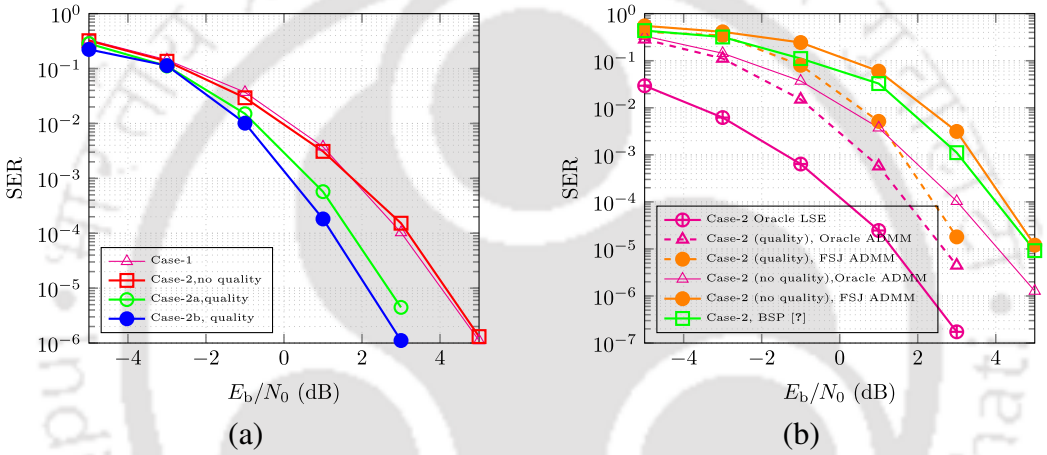


Fig. 4.4: SER performance: Frame-wise dynamic block-sparsity model (a) Various cases in dynamic system, (b) Oracle ADMM vs. adaptive ADMM.

4.6.3 SER performance : channel estimations errors (CEEs)

The proposed algorithm considers perfect CSI available at the receiver. The extensive simulations in this section show the effectiveness of the ADMM-based MUD for JADD in various scenarios. This section highlights the performance of the proposed detector against CEEs. In practice, the channel information is not available to the receiver and must be estimated before the data detection. Thus, the perfect channel estimation is not possible due to the channel estimation errors that occur in the estimation process. As a result, imperfect CSI is often available to the receiver. Thus, it is important to analyze the detector's performance in the presence of CEEs. A standard model is used to introduce the CEEs as follows

$$\widetilde{\mathbf{H}} = \mathbf{H} + \delta\mathbf{\Omega} \tag{4.28}$$

where $\mathbf{\Omega}$ indicates the CEE assumed to be uncorrelated with $\mathbf{H}$. The entries of $\mathbf{\Omega}$ are i.i.d complex Gaussian random variables with zero mean and unit variance, denoted as $\mathcal{CN}(0, 1)$. The parameter δ determines the variance of $\mathbf{\Omega}$. Fig. 4.5 show the performance of the proposed detector for the imperfect channel matrix given in (4.28) with $\delta = 0\%$, $\delta = 5\%$, $\delta = 10\%$, and $\delta = 20\%$. Observe that the ADMM-based MUD is robust against CEEs. These results show that the proposed detector performs acceptably in imperfect CSI scenarios.

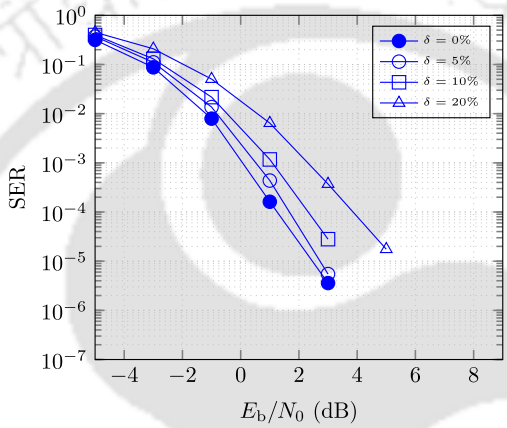


Fig. 4.5: SER performance: imperfect CSI scenario

4.6.4 Convergence

This subsection presents an empirical analysis of the residual convergence of the proposed ADMM-based detection algorithm, demonstrating the convergence behavior of the algorithm. A more rigorous theoretical analysis is provided in the Appendix B, establishing the guaranteed theoretical convergence properties of the ADMM-based detection algorithm. Furthermore, empirical observations on convergence help to select key ADMM parameters, specifically ρ and α_1 in (4.15), and T in **Algorithm** 3. These parameters significantly influence both the convergence speed and the recovery performance of the algorithm. This work chooses the ADMM parameters based on the empirical observations shown in Fig. 4.6. However, optimal parameter selection could further enhance the algorithm's speed and accuracy.

Figs. 4.6 (a) and 4.6 (b) illustrate the primal and dual residual convergence for **Algorithm** 3, showing that the proposed ADMM-based detection algorithm converges in a few tens of iterations. From Fig. 4.6, it is clear that the parameters ρ and α_1 significantly affect the convergence

speed. A larger ρ imposes a heavier penalty on primal feasibility violations, leading to lower primal residuals. In contrast, smaller ρ values result in smaller dual residuals. However, excessively large or small values of ρ are detrimental to both primal and dual residual convergence. On the other hand, α_1 in (4.15) controls the sparsity of the solution in (4.17b). Higher α_1 values intensify the penalty L_1, producing a highly sparse or even all-zero solution for $\mathbf{z}^{(t)}$, making the dual residual $\mathbf{r}_d^{(t)}$ zero. This is the reason why Fig. 4.6 (b) excludes the curves for $\alpha_1 > 1$. The selection of ρ and α_1 is interdependent, and achieving a balance between the two is crucial for the convergence speed and reconstruction accuracy. In summary, moderate values of ρ and smaller values of α_1 are preferable for the proposed ADMM detection algorithm. Fig. 4.6 shows that the algorithm converges quickly in primal and dual residuals with $\rho = 10$ and $\alpha_1 = 0.5$. The proposed algorithm converges to modest accuracy in almost ten iterations. This is sufficient to recover the signal with acceptable SER performance discussed in Sections 4.6.1 & 4.6.2.

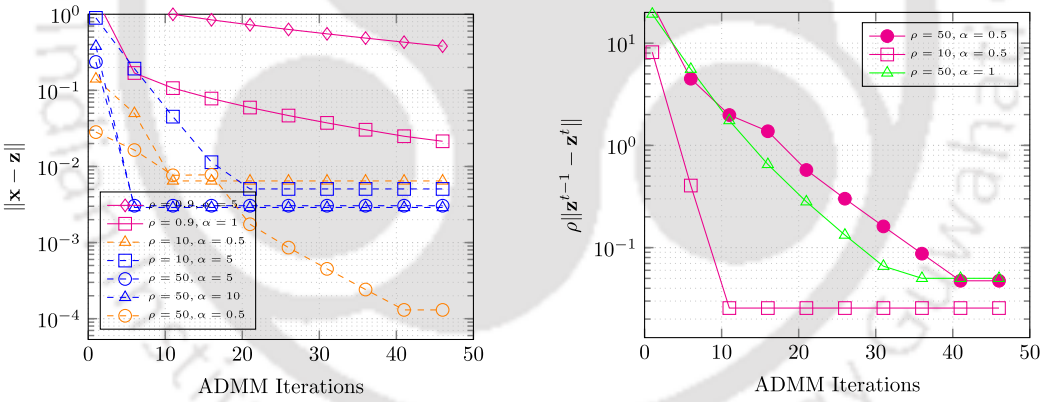


Fig. 4.6: Residual Convergence (a) Primal residual convergence, (b) Dual residual convergence.

4.7 Summary

This work proposed a robust prior information-assisted ADMM detection algorithm for CB-based grant-free NOMA systems. In this framework, grant-free NOMA systems based on CBs are introduced, utilizing the inherent block-sparsity structure to enhance the performance of JADD. For MUD in DCMA and SCMA systems, a convex optimization-based BCS framework is employed with group LASSO and sparse group LASSO, respectively. Additionally, practical challenges in mMTC networks are effectively modeled as a frame-wise dynamic block sparsity

problem and addressed using the ADMM framework. The benefits of the proposed detector are thoroughly analyzed in terms of computational complexity and performance.

5

Conclusions and Future Research Directions

Contents

This chapter presents the conclusions of the thesis and highlights its significance in advancing NGMA. Additionally, it outlines future directions, emphasizing the potential of CD-NOMA systems. This includes exploring the integration of CD-NOMA with key enablers of 6G.

This thesis focused on the modeling and optimization of CD-NOMA systems, addressing both transmitter and receiver aspects. This study addressed three primary challenges: optimal codebook design for SCMA systems, detection for both SCMA and DCMA systems, and JADD for CB-based grant-free NOMA systems. The conclusions drawn from these works are summarized below.

One of the key challenges in CD-NOMA systems is the optimal design of codebooks from MDCs. We began by exploring codebook designs for both uplink and downlink sparse CD-NOMA systems, proposing three distinct codebooks for the SCMA system under various channel conditions with 150% and 200% overloading factors. A thorough analysis of pair-wise error probability led to the development of new key performance indicators (KPIs), which contributed to performance improvements. Additionally, a new algorithm was proposed that systematically optimizes the codebooks.

Next, we focused on designing an efficient detector adaptable to both large-scale MIMO sparse and dense CD-NOMA systems. This work formulated a sharing optimization problem to address the ML detection challenge in CD-NOMA systems while maintaining low computational complexity. An efficient distributed optimization method, namely the ADMM algorithm, was then applied to solve the formulated problem. The proposed detector is particularly advantageous for DCMA systems, as it outperforms other viable alternatives in terms of both error-rate performance and computational complexity. A comprehensive comparison was conducted between the ADMM detector and state-of-the-art high-performance detectors. The soft-output ADMM detector was also applied, and its error performance was evaluated in LDPC-coded CD-NOMA systems. Additionally, the impact of ADMM parameters, such as the number of iterations and penalty factors, was thoroughly analyzed. A theoretical convergence analysis demonstrated that the ADMM detection algorithm is guaranteed to converge with appropriately chosen penalty parameters.

Finally, we addressed the MUD challenge in more practical grant-free NOMA systems. This work proposed a robust prior-information aided ADMM detection algorithm for CB based grant-free NOMA systems. In this context, CB-based grant-free NOMA systems were introduced, leveraging the inherent block sparsity structure for JADD. A BCS framework was employed to further enhance the performance of the multi-user detector in JADD. Two CB-based grant-free NOMA system models were presented to handle sporadic traffic scenarios in mMTC networks. A simple one-shot system model was introduced, highlighting the block sparsity structure and motivating the ADMM formulation for CB-based grant-free NOMA systems. Additionally, a frame-wise dynamic block sparsity model was proposed for continuous data transmission. The temporal correlation between active user support sets was leveraged as prior information, further improving detection performance.

5.1 Directions for Future Research

This section outlines potential future directions for this work. Several key suggestions are presented below.

- **Optimum and scalable codebook design:** In Chapter 1, near-optimal SCMA codebooks are designed to achieve improved performance in both UL and DL Rayleigh fading channels. However, designing an optimal codebook for given values of J, K, and M remains an open problem and a crucial direction for future research. Furthermore, scalable codebook design will play a significant role in addressing the challenges of mMTC networks.
- **Large-scale CD-NOMA systems:** Currently, small-scale CD-NOMA systems with overloading factors of 150% and 200%, featuring configurations like 4 REs with 6 User Equipments UEs, and 5 REs with 10 UEs, have been extensively studied. However, in grant-free NOMA scenarios, it is essential to address sporadic traffic patterns typically exists in mMTC networks. Grant-free NOMA necessitates large-scale CD-NOMA system to serve the users in mMTC networks. Moreover, the existing codebooks are relatively small in size, making them unsuitable for applications that require high data rates. This limitations affects the practical deployment of current CD-NOMA systems in 6G wireless

networks. Therefore, it is crucial to explore large-scale codebook design methods and low-complexity detection algorithms to meet these demands.

- **Joint Channel estimation and MUD in CB-based grant-free NOMA systems:** In Chapter 4, we proposed a robust prior-information-aided ADMM algorithm for JADD in CB-based grant-free NOMA systems. This work assumed perfect CSI at the receiver and aimed to reduce the computational overhead at the access point while maintaining competitive error rate performance in practical grant-free networks. A potential extension of this work involves relaxing the assumption of perfect CSI at the receiver. Consequently, formulating the MUD problem as joint channel estimation and activity detection, followed by data detection, particularly in CB-based grant-free systems. Furthermore, addressing this problem in the dynamic scenarios discussed in Chapter 4 presents an even more challenging and intriguing research direction.
- **Asymptotic Performance Analysis of the ADMM Detection Algorithm:** This work introduces ADMM-based detection algorithms for CD-NOMA systems in two distinct contexts. The first addresses the sharing problem formulated for multi-user detection, while the second tackles the group LASSO problems designed for JADD. For future research, an in-depth asymptotic performance analysis of the ADMM detector in both contexts could be investigated using the Convex Gaussian Min-Max Theorem (CGMT) [102]. Moreover, designing a turbo equalization scheme based on a soft-in soft-out ADMM detector would be particularly promising for practical applications.
- **Utilizing CD-NOMA to MEC enabled networks:** Mobile Edge Computing (MEC) is an emerging technology designed to alleviate the heavy computational demands on mobile and IoT devices in upcoming 6G networks [103]. Compared to traditional OMA-based MEC, NOMA-based MEC (NOMA-MEC) offers significant advantages in jointly addressing computation offloading and resource allocation. The optimization of resource allocation and computation offloading in NOMA-MEC has gained considerable attention in the research community and can be modeled as a Markov Decision Process (MDP) [104]. Consequently, reinforcement learning (RL) algorithms can be effectively employed to

solve this problem. Additionally, CD-NOMA-based MEC [105] provides further benefits over PD-NOMA-based MEC, thanks to the shaping and coding gains inherent in CD-NOMA systems. Exploring the joint optimization of resource allocation and computation offloading in CD-NOMA-based MEC, formulated as an MDP problem, presents a promising future research direction.

- **Design and analysis of NOMA aided heterogeneous networks:** One of the primary objectives of 6G is to accommodate users with diverse quality-of-service (QoS) and bandwidth requirements. This thesis focuses on the design, analysis, and optimization of code-domain NOMA (CD-NOMA) systems tailored for users with similar QoS and bandwidth demands. A promising direction for future research lies in extending this work to heterogeneous network scenarios. In particular, the design and analysis of NOMA systems based on irregular factor graphs (FGs) offer a strong foundation for addressing the challenges posed by such heterogeneous environments.
- **Design and analysis of NOMA for more practical channel scenarios:** This thesis assumes Rayleigh fading channel conditions for the design and analysis of CD-NOMA systems. However, analyzing CD-NOMA under more practical channel environments and developing corresponding codebook designs and detection algorithms is crucial. One of the widely accepted channel models for 5G and 6G communications is the Clustered Delay Line (CDL) model. Extending the current work to incorporate CDL-type channels represents a significant and promising direction for future research. In this direction, investigating codebook design for CDL type channel models along with diversity analysis and the derivation of new key performance indicators (KPIs) is essential. Additionally, it is important to evaluate the performance of ADMM-based detection algorithms under CDL-type channel conditions for CD-NOMA systems.

A

Appendix

A.1 Proof of Theorem 3.1

The following three lemmas are used to prove **Theorem 1** according to [106]. **Lemma 1** proves the objective function, $f(\mathbf{x}_{0j,k}) = \frac{1}{2}\|\mathbf{r}_k - \mathbf{H}_k\left(\sum_{j=1}^{N_u}\mathbf{x}_{0j,k}\right)\|_2^2$ is Lipschitz continuous gradient w.r.t. constant $\lambda_{\max}^2(\mathbf{H}_k^H\mathbf{H}_k)$ [65]. **Lemma 2** derives conditions on penalty parameters under which augmented Lagrangian function $\mathcal{L}$ from (3.30) always decreases. **Lemma 3** proves that there exists a lower bound on $\mathcal{L}$ with proper selection of penalty parameters. The proofs of Lemmas are provided below.

Lemma A.1. *The following inequality holds for the proposed Algorithm 2*

$$\|\mathbf{y}_{j,k}^{t+1} - \mathbf{y}_{j,k}^{t}\|^2 \leq \lambda_{\max}^2(\mathbf{H}_k^H\mathbf{H}_k)\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2, \quad \forall\, j,k.$$

Proof. Since $\mathbf{x}_{0j,k}^{t+1}$ is a minimizer of (3.32), the optimality condition is as follows

$$\nabla_{\mathbf{x}_{0j,k}} f(\mathbf{x}_{0j,k}^{t+1}) + (\mathbf{y}_{j,k}^{t} + \rho(\mathbf{x}_{0j,k}^{t+1} - \mathbf{z}_{0j,k}^{t+1})) = 0$$
$$\text{From (3.33)}, \quad \mathbf{y}_{j,k}^{t+1} = -\nabla_{\mathbf{x}_{0j,k}} f(\mathbf{x}_{0j,k}^{t+1}) \tag{A.1}$$

From the Lagrangian mean value theorem,

$$\frac{\nabla_{\mathbf{x}_{0j,k}} f(\mathbf{x}_{0j,k}^{t+1}) - \nabla_{\mathbf{x}_{0j,k}} f(\mathbf{x}_{0j,k}^{t})}{\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}} = \nabla_{\mathbf{x}_{0j,k}}^2 f(\mathbf{x}_{0j,k})$$
$$\nabla_{\mathbf{x}_{0j,k}}^2 f(\mathbf{x}_{0j,k}) = \mathbf{H}_k^H\mathbf{H}_k \leq \lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)I$$

$$\|\nabla_{\mathbf{x}_{0j,k}} f(\mathbf{x}_{0j,k}^{t+1}) - \nabla_{\mathbf{x}_{0j,k}} f(\mathbf{x}_{0j,k}^{t})\|^2 \leq \lambda_{\max}^2(\mathbf{H}_k^H \mathbf{H}_k)\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2, \forall j \tag{A.2}$$

From (B.6),

$$\|\mathbf{y}_{j,k}^{t+1} - \mathbf{y}_{j,k}^{t}\|^2 \leq \lambda_{\max}^2(\mathbf{H}_k^H \mathbf{H}_k)\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2, \forall j. \tag{A.3}$$

□

Lemma A.2. *For the augmented Lagrangian function (3.30), the following holds true*

$$\mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t+1}\}_{j=1}^{J}\right) \geq \sum_{j=1}^{N_u}\left(\frac{\rho+\gamma_j}{2} - \frac{\lambda_{\max}^2(\mathbf{H}_k^H\mathbf{H}_k)}{\rho}\right)\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2 + \sum_{j=1}^{N_u}\frac{\gamma_j - \rho}{2}\|\mathbf{z}_{0j,k}^{t+1} - \mathbf{z}_{0j,k}^{t}\|^2. \tag{A.4}$$

Proof.

$$\mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t+1}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) = \underbrace{\left(\mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t+1}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right)\right)}_{\text{term-1}} + \underbrace{\left(\mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right)\right)}_{\text{term-2}} \tag{A.5}$$

The term-1 in (A.5) is given by

$$\mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t+1}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) = \sum_{j=1}^{Nu}\frac{1}{\rho}\|\mathbf{y}_{j,k}^{t+1} - \mathbf{y}_{j,k}^{t}\|^2 \leq \sum_{j=1}^{Nu}\frac{1}{\rho}\lambda_{\max}^2(\mathbf{H}_k^H\mathbf{H}_k)\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2 \tag{A.6}$$

The term-2 in (A.5) is given by

$$\begin{aligned}
\mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) &= \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) \\
&\quad + \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) - \mathcal{L}\left(\{\mathbf{x}_{0j,k}^{t}, \mathbf{z}_{0j,k}^{t}, \mathbf{y}_{j,k}^{t}\}_{j=1}^{J}\right) \\
&\overset{(a)}{\leq} \sum_{j=1}^{N_u}\langle\nabla_{\mathbf{x}_{0j,k}}\mathcal{L}(\mathbf{x}_{01,k}^{t+1}, \ldots, \mathbf{x}_{0j-1,k}^{t+1}, \mathbf{x}_{0j,k}^{t+1}, \mathbf{x}_{0j+1,k}^{t}, \ldots, \mathbf{x}_{0J,k}^{t}, \{\mathbf{z}_{0j,k}, \mathbf{y}_{j,k}\}_{j=1}^{J}, (\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t})\rangle - \frac{\rho+\gamma_j}{2}\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2 \\
&\quad + \sum_{j=1}^{N_u}\langle\nabla_{\mathbf{z}_{0j,k}}\mathcal{L}(\mathbf{z}_{01,k}^{t+1}, \ldots, \mathbf{z}_{0j-1,k}^{t+1}, \mathbf{z}_{0j,k}^{t+1}, \mathbf{z}_{0j+1,k}^{t}, \ldots, \mathbf{z}_{0J,k}^{t}, \{\mathbf{x}_{0j,k}, \mathbf{y}_{j,k}\}_{j=1}^{J}), (\mathbf{z}_{0j,k}^{t+1} - \mathbf{z}_{0j,k}^{t})\rangle - \frac{\gamma_j - \rho}{2}\|\mathbf{z}_{0j,k}^{t+1} - \mathbf{z}_{0j,k}^{t}\|^2 \\
&\overset{(b)}{\leq} \sum_{j=1}^{N_u} -\frac{\rho+\gamma_j}{2}\|\mathbf{x}_{0j,k}^{t+1} - \mathbf{x}_{0j,k}^{t}\|^2 - \sum_{j-1}^{N_u}\frac{\gamma_j - \rho}{2}\|\mathbf{z}_{0j,k}^{t+1} - \mathbf{z}_{0j,k}^{t}\|^2
\end{aligned} \tag{A.7}$$

where (a) uses the fact that $\mathcal{L}\left(\{\mathbf{x}_{0j,k}, \mathbf{z}_{0j,k}, \mathbf{y}_{j,k}\}_{j=1}^{J}\right)$ is strongly convex [98] w.r.t. each $\mathbf{x}_{0j,k}$ and $\mathbf{z}_{0j,k}$, with modulus $\rho + \gamma_j$ and $\gamma_j - \rho$, respectively. In (b) we have used the optimality condition of subproblems (3.32) and (3.31), where $\mathbf{x}_{0j,k}^{t+1}$ and $\mathbf{z}_{0j,k}^{t+1}$ are the minimizers of (3.32) and (3.31),

respectively. By combining the inequalities (A.6) and (A.7) in (A.5), we obtain

$$\begin{aligned}\mathcal{L}\left(\{\mathbf{x}^{t+1}_{0j,k},\mathbf{z}^{t+1}_{0j,k},\mathbf{y}^{t+1}_{j,k}\}_{j=1}^{J}\right)-\mathcal{L}\left(\{\mathbf{x}^{t}_{0j,k},\mathbf{z}^{t}_{0j,k},\mathbf{y}^{t}_{j,k}\}_{j=1}^{J}\right)&\leq\sum_{j=1}^{Nu}-\frac{\rho+\gamma_j}{2}\|\mathbf{x}^{t+1}_{0j,k}-\mathbf{x}^{t}_{0j,k}\|^2-\sum_{j=1}^{Nu}\frac{\gamma_j-\rho}{2}\|\mathbf{z}^{t+1}_{0j,k}-\mathbf{z}^{t}_{0j,k}\|^2\\&+\sum_{j=1}^{Nu}\frac{\lambda^2_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{\rho}\|\mathbf{x}^{t+1}_{0j,k}-\mathbf{x}^{t}_{0j,k}\|^2\\&\leq\sum_{j=1}^{Nu}\left(\frac{\lambda^2_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{\rho}-\frac{\rho+\gamma_j}{2}\right)\|\mathbf{x}^{t+1}_{0j,k}-\mathbf{x}^{t}_{0j,k}\|^2-\sum_{j=1}^{Nu}\frac{\gamma_j-\rho}{2}\|\mathbf{z}^{t+1}_{0j,k}-\mathbf{z}^{t}_{0j,k}\|^2\end{aligned}\tag{A.8}$$

The above inequality implies that if $\rho(\rho+\gamma_j)\geq 2\lambda^2_{\max}(\mathbf{H}_k^H\mathbf{H}_k)$ and $\gamma_j-\rho>0,\forall j=1,\ldots,J$, then the augmented Lagrangian function always decreases. One can always find ρ and γ_k based on $\mathbf{H}_k$ such that the above condition is satisfied. □

Lemma A.3. *Let* $\{\mathbf{x}^t_{0j,k},\mathbf{z}^t_{0j,k},\mathbf{y}^t_{j,k}\}_{j=1}^J$ *be generated by Algorithm 2. Assume* $\gamma_j,\forall j=1,\ldots,J$, $\gamma_j>\rho$ *and* $\rho\geq\lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)$. *The following lower bound exists*

$$\mathcal{L}\left(\{\mathbf{x}^t_{0j,k},\mathbf{z}^t_{0j,k},\mathbf{y}^t_{j,k}\}_{j=1}^J\right)\geq f(\mathbf{z}^t_{0j,k})+\sum_{j=1}^{Nu}\frac{\gamma_j}{2}\|\mathbf{z}^t_{0j,k}\|^2.\tag{A.9}$$

Proof.

$$\mathcal{L}\left(\{\mathbf{x}^t_{0j,k},\mathbf{z}^t_{0j,k},\mathbf{y}^t_{j,k}\}_{j=1}^J\right)=f(\mathbf{x}^t_{0j,k})+\sum_{j=1}^{Nu}\frac{\gamma_j}{2}\|f(\mathbf{z}^t_{0j,k})\|^2+\sum_{j=1}^{Nu}\langle\nabla_{\mathbf{x}_{0j,k}}f(\mathbf{x}^t_{0j,k}),\mathbf{z}^t_{0j,k}-\mathbf{x}^t_{0j,k}\rangle+\sum_{j=1}^{Nu}\frac{\rho}{2}\|\mathbf{x}^t_{0j,k}-\mathbf{z}^t_{0j,k}\|^2\tag{A.10}$$

The objective function f is $\lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)$-gradient Lipschitz according to (A.2). The upper quadratic approximation for f is given by [97]

$$f(\mathbf{z}^t_{0j,k})\leq f(\mathbf{x}^t_{0j,k})+\sum_{j=1}^{Nu}\langle\nabla_{\mathbf{x}_{0j,k}}f(\mathbf{x}^t_{0j,k}),\mathbf{z}^t_{0j,k}-\mathbf{x}^t_{0j,k}\rangle+\sum_{j=1}^{Nu}\frac{\lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{2}\|\mathbf{z}^t_{0j,k}-\mathbf{x}^t_{0j,k}\|^2$$

which can be written as

$$f(\mathbf{x}^t_{0j,k})+\sum_{j=1}^{Nu}\langle\nabla_{\mathbf{x}_{0j,k}}f(\mathbf{x}^t_{0j,k}),\mathbf{z}^t_{0j,k}-\mathbf{x}^t_{0j,k}\rangle\geq f(\mathbf{z}^t_{0j,k})-\sum_{j=1}^{Nu}\frac{\lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{2}\|\mathbf{z}^t_{0j,k}-\mathbf{x}^t_{0j,k}\|^2\tag{A.11}$$

Plugging (A.11) into (A.10), we get

$$\mathcal{L}\left(\{\mathbf{x}^t_{0j,k},\mathbf{z}^t_{0j,k},\mathbf{y}^t_{j,k}\}_{j=1}^J\right)\geq f(\mathbf{z}^t_{0j,k})+\sum_{j=1}^{Nu}\frac{\gamma_j}{2}\|\mathbf{z}^t_{0j,k}\|^2+\sum_{j=1}^{Nu}\left(\frac{\rho}{2}-\frac{\lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{2}\right)\|\mathbf{z}^t_{0j,k}-\mathbf{x}^t_{0j,k}\|^2.$$

Suppose $\rho\geq\lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)$ and $f(\mathbf{z}^t_{0j,k})+\sum_{j=1}^{Nu}\frac{\gamma_j}{2}\|\mathbf{z}^t_{0j,k}\|^2$ is bounded over the real part $\mathbf{z}_{0j,k,R}\in[-\alpha_j,\alpha_j]$ and the imaginary part $\mathbf{z}_{0j,k,I}\in[\beta_j,\beta_j]$. Hence, $\mathcal{L}\left(\{\mathbf{x}^t_{0j,k},\mathbf{z}^t_{0j,k},\mathbf{y}^t_{j,k}\}_{j=1}^J\right)\geq f(\mathbf{z}^t_{0j,k})+\sum_{j=1}^{Nu}\frac{\gamma_j}{2}\|\mathbf{z}^t_{0j,k}\|^2$. □

Now the proof for **Theorem-1** is presented below.

Proof. According to Lemma-2, adding both sides of the (A.4) for $t = 1, \ldots, \infty$, we get

$$\begin{aligned}\mathcal{L}\left(\{\mathbf{x}^1_{0j,k}, \mathbf{z}^1_{0j,k}, \mathbf{y}^1_{j,k}\}_{j=1}^J\right) - \lim_{t\to\infty} \mathcal{L}\left(\{\mathbf{x}^t_{0j,k}, \mathbf{z}^t_{0j,k}, \mathbf{y}^t_{j,k}\}_{j=1}^J\right) \\ \geq \sum_{t=1}^{\infty}\sum_{j=1}^{Nu}\left(\frac{\rho+\gamma_j}{2} - \frac{\lambda^2_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{\rho}\right)\|\mathbf{x}^{t+1}_{0j,k} - \mathbf{x}^t_{0j,k}\|^2 \\ + \sum_{t=1}^{\infty}\sum_{j=1}^{Nu}\frac{\gamma_j-\rho}{2}\|\mathbf{z}^{t+1}_{0j,k} - \mathbf{z}^t_{0j,k}\|^2\end{aligned} \tag{A.12}$$

From Lemma-3, $\lim_{t\to\infty}\mathcal{L}\left(\{\mathbf{x}^t_{0j,k}, \mathbf{z}^t_{0j,k}, \mathbf{y}^t_{j,k}\}_{j=1}^J\right) > -\infty$. Furthermore, if $\frac{\rho+\gamma_j}{2} - \frac{\lambda^2_{\max}(\mathbf{H}_k^H\mathbf{H}_k)}{\rho} > 0$ and $\gamma_j - \rho > 0$, the following limit can be obtained:

$$\lim_{t\to\infty}\|\mathbf{x}^{t+1}_{0j,k} - \mathbf{x}^t_{0j,k}\| = 0$$

$$\lim_{t\to\infty}\|\mathbf{z}^{t+1}_{0j,k} - \mathbf{z}^t_{0j,k}\| = 0$$

From Lemma-1, $\|\mathbf{y}^{t+1}_{j,k} - \mathbf{y}^t_{j,k}\| \leq \lambda_{\max}(\mathbf{H}_k^H\mathbf{H}_k)\|\mathbf{x}^{t+1}_{0j,k} - \mathbf{x}^t_{0j,k}\|, \forall j,k$

$$\lim_{t\to\infty}\|\mathbf{y}^{t+1}_{j,k} - \mathbf{y}^t_{j,k}\| = 0.$$

As, $\mathbf{y}^{t+1}_{j,k} = \mathbf{y}^t_{j,k} + \rho(\mathbf{x}^{t+1}_{0j,k} - \mathbf{z}^{t+1}_{0j,k})$, the following limit exists

$$\lim_{t\to\infty}\|\mathbf{x}^{t+1}_{0j,k} - \mathbf{z}^{t+1}_{0j,k}\| = 0 \tag{A.13}$$

Since the real and imaginary parts of $\mathbf{x}_{0j,k}$ and $\mathbf{z}_{0j,k}$ are bounded over $[-\alpha_j, \alpha_j]$ and $[-\beta_j, \beta_j]$, respectively, the following limit point exists for $\mathbf{x}_{0j,k}$ and $\mathbf{z}_{0j,k}$

$$\lim_{t\to\infty}\mathbf{x}^t_{0j,k} = \mathbf{x}^\star_{0j,k}, \quad \lim_{t\to\infty}\mathbf{z}^t_{0j,k} = \mathbf{z}^\star_{0j,k} \tag{A.14}$$

Plugging (A.14) into (A.13), then

$$\lim_{t\to\infty}\mathbf{x}^t_{0j,k} = \mathbf{x}^\star_{0j,k} = \mathbf{z}^\star_{0j,k}$$

Since $\mathbf{x}_{0j,k}$ is bounded as mentioned above, from (B.10), the stationary point exists for $\mathbf{y}_{j,k}$ as follows

$$\lim_{t\to\infty}\mathbf{y}^t_{j,k} = \mathbf{y}^\star_{j,k}.$$

Hence, $\{\mathbf{x}^\star_{0j,k}, \mathbf{z}^\star_{0j,k}, \mathbf{y}^\star_{j,k}\}_{j=1}^J$ is the stationary solution. □

B

Appendix

B.1 Shrinkage operation

Shrinkage operation is used to solve the problems in (4.9b) and (4.9c), which are convex but non-differentiable. Typically, sub-gradient operation is applied on problems of this kind as given in (4.12) and (4.13). The shrinkage operation on problem (4.10) related to (4.13) is given by

Proof.

$$\min_{\mathbf{z}_j} \alpha_1 w_j \|\mathbf{z}_j\|_2 + \frac{\rho}{2}\|\mathbf{x}_j - \mathbf{z}_j + \mathbf{u}_{1j}\|_2^2, \tag{B.1}$$

where $\mathbf{z}_j, \mathbf{x}_j$, and $\mathbf{u}_{1j} \in \mathbb{C}^K$. The sub-gradient of $\|\mathbf{z}_j\|_2$ is given by

$$\mathbf{v} = \begin{cases} \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} & \text{iff} \mathbf{z}_j \neq 0, \\ \in \{\mathbf{v} : \|\mathbf{v}\|_2 \leq 1\} & \text{iff } \mathbf{z}_j = 0. \end{cases} \tag{B.2}$$

Taking the derivative of (B.1), we get

$$\alpha_1 w_j \mathbf{v} + \rho(\mathbf{z}_j - \mathbf{x}_j - \mathbf{u}_j) = 0. \tag{B.3}$$

By substituting (B.2) in (B.3), we obtained the optimal solution for the problem (B.1) is given by

$$\hat{\mathbf{z}}_j = \left(\|\mathbf{x}_j + \mathbf{u}_{1j}\|_2 - \frac{\alpha_1 w_j}{\rho}\right)_+ \frac{\mathbf{x}_j + \mathbf{u}_{1j}}{\|\mathbf{x}_j + \mathbf{u}_{1j}\|_2}.$$

The similar procedure is applicable to problem (4.9b), where the sub-gradient of $|\mathbf{q}||_1$ is given by

$$v_i = \begin{cases} \text{sign}(q_i) & \text{iff } q_i \neq 0, \\ \in \{v_i : |v_i| \leq 1\} & \text{iff } q_i = 0. \end{cases} \tag{B.4}$$

We obtain the element-wise optimal solution for the problem (4.9b) is given by

$$\hat{q}_i = \left(|x_i + u_{2i}| - \frac{\alpha_2 w_i}{\rho}\right)_+ \frac{x_i + u_{2i}}{|x_i + u_{2i}|}.$$

or equivalently

$$\hat{q}_i = \max(x_i + u_{2i} - \frac{\alpha_2 w_i}{\rho}, 0) - \max(-x_i - u_{2i} - \frac{\alpha_2 w_i}{\rho}, 0).$$

□

B.2 Proof of Theorem 4.1

The following assumptions are made to analyze the proposed ADMM-based detection algorithm and are subsequently validated through Lemmas..

Assumptions:

- The function f is Lipschitz continuous, then there exists a positive constant m such that

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}^{t+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^t)\| \leq m\|\mathbf{x}^{t+1} - \mathbf{x}^t\|,$$

  where $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r} - \mathbf{H}\mathbf{x}\|_2^2$.

- The parameter ρ is chosen positive and large enough such that the sub-problems related to x and z are strongly convex with modulus γ_1 and γ_2, respectively, where $\rho\gamma_1 > 2m^2$ and $\rho \geq m$. It implies that $\mathcal{L}\left(\{\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right)$ always decreases and lower bounded as follows

$$\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) > \mathcal{L}\left(\{\mathbf{x}^{t+1}, \{\mathbf{z}_j^{t+1}\}, \mathbf{y}^{t+1}\right),$$
$$\lim_{t\to\infty} \mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) > -\infty,$$

  where $\mathcal{L}\left(\{\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right)$ is the augmented Lagrangian function of problem (4.15) is given by

$$\mathcal{L}\left(\mathbf{x}, \{\mathbf{z}_j\}_{j=1}^J, \mathbf{y}_1\right) = \frac{1}{2}\|\mathbf{r} - \mathbf{H}\mathbf{x}\|_2^2 + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j\|_2 + \text{Re}\langle \mathbf{y}_1, \mathbf{x} - \mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|_2^2. \tag{B.5}$$

The following three lemmas are used to prove **Theorem 1** according to [106]. **Lemma 1** proves the function, $f(\mathbf{x})$ is Lipschitz continuous w.r.t. constant $m = \lambda_{\max}(\mathbf{H}^H\mathbf{H})$ [65], it shows the difference between the dual variables in two consecutive iterations bounded above by that of the primal variables. **Lemma 2** derives conditions on penalty parameters under which augmented Lagrangian function $\mathcal{L}$ (B.5) always decreases. **Lemma 3** proves that there exists a lower bound on $\mathcal{L}$ with proper selection of penalty parameters. The proofs of Lemmas are provided below.

Lemma B.1. *The following inequality holds for the proposed Algorithm 3*

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}^{t+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^t)\|^2 \leq \lambda_{\max}^2(\mathbf{H}^H\mathbf{H})\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2.$$

Proof. Since $\mathbf{x}^{t+1}$ is a minimizer of (4.16a), the optimality condition is as follows

$$\nabla_{\mathbf{x}} f(\mathbf{x}^{t+1}) + (\mathbf{y}^t + \rho(\mathbf{x}^{t+1} - \mathbf{z}^{t+1})) = 0,$$
$$\text{From (4.17}c\text{)}, \quad \nabla_{\mathbf{x}} f(\mathbf{x}^{t+1}) = -\mathbf{y}_1^{t+1}. \tag{B.6}$$

From the Lagrangian mean value theorem,

$$\frac{\nabla_{\mathbf{x}} f(\mathbf{x}^{t+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^t)}{\mathbf{x}^{t+1} - \mathbf{x}^t} = \nabla_{\mathbf{x}}^2 f(\mathbf{x}), \tag{B.7}$$
$$\nabla_{\mathbf{x}}^2 f(\mathbf{x}) = \mathbf{H}^H\mathbf{H} \leq \lambda_{\max}(\mathbf{H}^H\mathbf{H})I, \tag{B.8}$$

Substitute (B.8) in (B.7),

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}^{t+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^t)\|^2 \leq \lambda_{\max}^2(\mathbf{H}^H\mathbf{H})\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \tag{B.9}$$

From (B.6),

$$\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \leq \lambda_{\max}^2(\mathbf{H}^H\mathbf{H})\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2, \forall j. \tag{B.10}$$

□

Lemma B.2. *For the augmented Lagrangian function (B.5), the following holds true*

$$\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) - \mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^{t+1}\}, \mathbf{y}^{t+1}\right)$$
$$\geq \left(\frac{\gamma_1}{2} - \frac{\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})}{\rho}\right)\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \sum_{j=1}^{J}\frac{\gamma_2}{2}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2. \tag{B.11}$$

Proof.

$$\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^{t+1}\}, \mathbf{y}^{t+1}\right) - \mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) = \underbrace{\left(\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^{t+1}\}, \mathbf{y}^{t+1}\right) - \mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^{t+1}\}, \mathbf{y}^t\right)\right)}_{\text{term-1}}$$
$$+ \underbrace{\mathcal{L}\left(\mathbf{x}^{t+1}, \{\mathbf{z}_j^{t+1}\}, \mathbf{y}^t\right) - \mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right)}_{\text{term-2}}. \tag{B.12}$$

From (4.17c) and (B.10), the term-1 in (B.12) is simplified as

$$\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t+1}\},\mathbf{y}^{t+1}\right)-\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t+1}\},\mathbf{y}^{t}\right)=\frac{1}{\rho}\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\|^2\leq\frac{1}{\rho}\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\|^2. \tag{B.13}$$

The term-2 in (B.12) is given by

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t+1}\},\mathbf{y}^{t}\right)-\mathcal{L}\left(\mathbf{x}^{t},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)&=\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t+1}\},\mathbf{y}^{t}\right)-\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)+\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)-\mathcal{L}\left(\mathbf{x}^{t},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)\\&\overset{(a)}{\leq}\sum_{j=1}^{J}\langle\varsigma_{\mathbf{z}_j^{t+1}},\mathbf{z}_j^{t+1}-\mathbf{z}_j^{t}\rangle-\frac{\gamma_2}{2}\|\mathbf{z}_j^{t+1}-z_j^{t}\|_2^2+\langle\nabla_x\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right),\mathbf{x}^{t+1}-\mathbf{x}^{t}\rangle-\frac{\gamma_1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\|_2^2\\&\overset{(b)}{\leq}-\frac{\gamma_1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\|^2-\sum_{j=1}^{J}\frac{\gamma_2}{2}\|\mathbf{z}^{t+1}-\mathbf{z}^{t}\|^2,\end{aligned} \tag{B.14}$$

where (a) uses the fact that $\mathcal{L}\left(\mathbf{x},\{\mathbf{z_j}\},\mathbf{y}\right)$ is strongly convex [98] w.r.t. each $\mathbf{x}$ and $\mathbf{z}$, with modulus $\gamma_1=\rho+\lambda_{min}(\mathbf{H}^H\mathbf{H})$ and $\gamma_2=\rho$, respectively. It is neccesaty that $\rho>0$ to hold the strong convexity of $\mathcal{L}\left(\mathbf{x},\{\mathbf{z}_j\},\mathbf{y}\right)$ w.r.t variable $\mathbf{z}_j$. In (b) we have used the optimality conditions of sub-problems (4.16a) and (4.16b) given by $\varsigma_{\mathbf{z}_j^{t+1}}\in\partial_{\mathbf{z}_j}\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t+1}\},\mathbf{y}^{t}\right)\in 0$ and $\nabla_x\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)=0$, where $\mathbf{x}^{t+1}$ and $\mathbf{z}^{t+1}$ are the minimizers of (4.16a) and (4.16b), respectively, and $\varsigma_{\mathbf{z}_j^{t+1}}$ is a sub-gradient vector. $\varsigma_{\mathbf{z}_j^{t+1}}$ implies $\mathcal{L}\left(\mathbf{x},\{\mathbf{z_j}\},\mathbf{y}\right)$ is non-differentiable function w.r.t. $\mathbf{z}_j$ can be observed in (B.5).

By combining the inequalities (B.13) and (B.14), equation (B.12) can be written as

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}^{t+1},\{\mathbf{z}_j^{t+1}\},\mathbf{y}^{t+1}\right)-\mathcal{L}\left(\mathbf{x}^{t},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)&\leq-\frac{\gamma_1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\|^2-\sum_{j=1}^{J}\frac{\gamma_2}{2}\|\mathbf{z}^{t+1}-\mathbf{z}^{t}\|^2\\&\quad+\frac{\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})}{\rho}\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\|^2\\&\leq\left(\frac{\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})}{\rho}-\frac{\gamma_1}{2}\right)\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\|^2-\sum_{j=1}^{J}\frac{\gamma_2}{2}\|\mathbf{z}^{t+1}-\mathbf{z}^{t}\|^2.\end{aligned} \tag{B.15}$$

If $\rho\gamma_1\geq 2\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})$ and $\gamma_2>0$, then the augmented Lagrangian function always decreases. It is always possible to find a ρ which satisfy the conditions $\rho\gamma_1\geq 2\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})$ based on $\mathbf{H}$ along with $\gamma_2>0$. □

Lemma B.3. *Let* $\{\mathbf{x}^t,\{\mathbf{z}^t\}_{j=1}^J,\mathbf{y}^t\}$ *be generated by Algorithm 3. Assume* $\rho\geq\lambda_{\max}(\mathbf{H}^H\mathbf{H})$. *The following lower bound exists*

$$\mathcal{L}\left(\mathbf{x}^{t},\{\mathbf{z}_j^{t}\},\mathbf{y}^{t}\right)\geq f(\mathbf{z}^t)+\alpha_1\sum_{j=1}^{J}w_j\|\mathbf{z}_j^t\|^2. \tag{B.16}$$

Proof. Recall the augmented Lagrangian function of sparse group LASSO formulated for DCMA

detection

$$\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) = f(\mathbf{x}^t) + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j^t\|^2 + \langle \nabla_{\mathbf{x}} f(\mathbf{x}^t), \mathbf{z}^t - \mathbf{x}^t \rangle + \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{z}^t\|^2. \tag{B.17}$$

The (B.17) uses the equality in (B.6). The $\nabla_{\mathbf{x}} f(\mathbf{x})$ is Lipschitz continuous with constant $\lambda_{\max}(\mathbf{H}^H\mathbf{H})$ according to (B.9). Thus, according to descent Lemma [97] the upper quadratic approximation for f is given by

$$f(\mathbf{z}^t) \leq f(\mathbf{x}^t) + \langle \nabla_{\mathbf{x}} f(\mathbf{x}^t), \mathbf{z}^t - \mathbf{x}^t \rangle + \frac{\lambda_{\max}(\mathbf{H}^H\mathbf{H})}{2}\|\mathbf{z}^t - \mathbf{x}^t\|^2,$$

which can be written as

$$f(\mathbf{x}^t) + \langle \nabla_{\mathbf{x}} f(\mathbf{x}^t), \mathbf{z}^t - \mathbf{x}^t \rangle \geq f(\mathbf{z}^t) - \frac{\lambda_{\max}(\mathbf{H}^H\mathbf{H})}{2}\|\mathbf{z}^t - \mathbf{x}^t\|^2. \tag{B.18}$$

Plugging (B.18) into (B.5), we get

$$\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) \geq f(\mathbf{z}^t) + \alpha_1 \sum_{j=1}^{J} w_j \|\mathbf{z}_j^t\|^2 + \left(\frac{\rho}{2} - \frac{\lambda_{\max}(\mathbf{H}^H\mathbf{H})}{2}\right)\|\mathbf{z}^t - \mathbf{x}^t\|^2.$$

Suppose $\rho \geq \lambda_{\max}(\mathbf{H}^H\mathbf{H})$ and $f(\mathbf{z}^t) + \sum_{j=1}^{Nu} \frac{\gamma_j}{2}\|\mathbf{z}^t\|^2$ is bounded, due to the fact that $\mathbf{z}^t$ is belongs to finite energy codebooks. Hence, $\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) \geq f(\mathbf{z}^t) + \sum_{j=1}^{Nu} \frac{\gamma_j}{2}\|\mathbf{z}^t\|^2$, lower bounded. This combined with (B.11) in **Lemma** 2, whenever the penalty parameter ρ chosen sufficiently large then $\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right)$ is monotonically decreasing and lower bounded. The two assumptions made in the beginning of the Appendix are validated. □

Now the proof for **Theorem-1** is presented below.

Proof. According to Lemma-2, adding both sides of the (B.11) for $t = 1, \ldots, \infty$, we get

$$\mathcal{L}\left(\mathbf{x}^1, \{\mathbf{z}_j^1\}, \mathbf{y}^1\right) - \lim_{t\to\infty} \mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) \geq \left(\frac{(\rho + \lambda_{min}(\mathbf{H}^H\mathbf{H}))}{2} - \frac{\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})}{\rho}\right)\sum_{t=1}^{\infty}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \sum_{t=1}^{\infty}\frac{\rho}{2}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2 \tag{B.19}$$

From Lemma-3, $\mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right)$ is lower bounded i.e. $\lim_{t\to\infty} \mathcal{L}\left(\mathbf{x}^t, \{\mathbf{z}_j^t\}, \mathbf{y}^t\right) > -\infty$. Furthermore, if $\frac{(\rho+\lambda_{min}(\mathbf{H}^H\mathbf{H}))}{2} - \frac{\lambda_{\max}^2(\mathbf{H}^H\mathbf{H})}{\rho} > 0$ and $\rho > 0$, the following limit can be obtained:

$$\lim_{t\to\infty} \|\mathbf{x}^{t+1} - \mathbf{x}^t\| = 0 \tag{B.20}$$

$$\lim_{t\to\infty} \|\mathbf{z}^{t+1} - \mathbf{z}^t\| = 0 \tag{B.21}$$

From Lemma-1, $\|\mathbf{y}^{t+1} - \mathbf{y}^t\| \leq \lambda_{\max}(\mathbf{H}^H\mathbf{H})\|\mathbf{x}^{t+1} - \mathbf{x}^t\|, \forall j, k$

$$\lim_{t\to\infty} \|\mathbf{y}^{t+1} - \mathbf{y}^t\| = 0. \tag{B.22}$$

As, $\mathbf{y}^{t+1} = \mathbf{y}^t + \rho(\mathbf{x}^{t+1} - \mathbf{z}^{t+1})$, the following limit exists

$$\lim_{t\to\infty} \|\mathbf{x}^{t+1} - \mathbf{z}^{t+1}\| = 0 \tag{B.23}$$

Since the real and imaginary parts of $\mathbf{x}$ and $\mathbf{z}$ are bounded over $[\alpha_j, \alpha_j]$ and $[-\beta_j, \beta_j]$, respectively, then the limit point exists for $\mathbf{x}$ and $\mathbf{z}$ from (B.20) and (B.21) as follows

$$\lim_{t\to\infty} \mathbf{x}^t = \mathbf{x}^\star, \quad \lim_{t\to\infty} \mathbf{z}^t = \mathbf{z}^\star \tag{B.24}$$

Plugging (B.24) into (B.23), then

$$\lim_{t\to\infty} \mathbf{x}^t = \mathbf{x}^\star = \mathbf{z}^\star \tag{B.25}$$

Since $\mathbf{x}$ is bounded as mentioned above, from (B.10), the stationary point exists for $\mathbf{y}$ as follows

$$\lim_{t\to\infty} \mathbf{y}^t = \mathbf{y}^\star. \tag{B.26}$$

Hence, $\{\mathbf{x}^\star, \mathbf{z}^\star, \mathbf{y}^\star\}_{j=1}^J$ is the stationary solution. □

B.2.1 Convergence of ADMM algorithm on problem (4.26)

The above derivations on convergence of ADMM on (4.15) holds true for prior-information aided ADMM algorithm on (4.26). The augmented Lagrangian function for the problem (4.26) is given by

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}^{[l]}, \{\mathbf{z}_j^{[l]}\}_{j=1}^J, \mathbf{y}_1^{[l]}\right) = \frac{1}{2}\|\mathbf{r}^{[l]} - \mathbf{H}^{[l]}\mathbf{x}^{[l]}\|_2^2 + \alpha_1 \sum_{j=1}^{J} w_{j,p}^l w_{j,d} \|\mathbf{z}_j\|_2 \\ \frac{\mu^l}{2}\|\mathbf{x}^{[l]} - \hat{\beta}_{q^{[l]}}^{[l-1]}\|^2 + \mathrm{Re}\langle \mathbf{y}_1^{[l]}, \mathbf{x}^{[l]} - \mathbf{z}^{[l]}\rangle + \frac{\rho}{2}\|\mathbf{x}^{[l]} - \mathbf{z}^{[l]}\|_2^2\end{aligned} \tag{B.27}$$

In this case $f(\mathbf{x})$ in Subsection B.2 turns out to be $f(\mathbf{x}^{[l]})$ as given by

$$f(\mathbf{x}^{[l]}) = \frac{1}{2}\|\mathbf{r}^{[l]} - \mathbf{H}^{[l]}\mathbf{x}^{[l]}\|_2^2 + \frac{\mu^l}{2}\|\mathbf{x}^{[l]} - \hat{\beta}_{q^{[l]}}^{[l-1]}\|^2,$$

By following the similar procedure in Section B.2, the conditions on ρ in problem (B.27) are $\rho(\rho + \mu^l + \lambda_{\min}(\mathbf{H}^H\mathbf{H})) \geq 2(\lambda_{\max}(\mathbf{H}^H\mathbf{H}) + \mu^l)^2$ and $\rho > (\lambda_{\max}(\mathbf{H}^H\mathbf{H}) + \mu^l)$.

Bibliography

[1] M. Vaezi, Z. Ding, and H. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Springer International Publishing, Jan. 2018.

[2] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[3] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of NOMA toward next generation multiple access (NGMA) for 6g," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1037–1071, 2022.

[4] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for iot: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1805–1838, 2020.

[5] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-Orthogonal Multiple Access (NOMA) for cellular future radio access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5.

[6] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[7] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, 2013, pp. 770–774.

[8] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the performance of noma with hybrid ARQ," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10 033–10 038, 2018.

[9] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[10] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication. Cambridge, U.K,: Cambridge Univ.*, 2005.

[11] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.

[12] W. Lee, "Overview of cellular CDMA," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 2, pp. 291–302, 1991.

[13] J. van de Beek and B. M. Popovic, "Multiple access with low-density signatures," in *GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference*, 2009, pp. 1–6.

[14] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, 2008.

[15] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for internet of things," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016, pp. 1–5.

[16] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *2010 IEEE 71st Vehicular Technology Conference*, 2010, pp. 1–5.

[17] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th Annual Int. Symp. Per. Indoor Mobile Radio Commun (PIMRC)*, 2013, pp. 332–336.

[18] G. D. Forney and L. Wei, "Multidimensional constellations. i. introduction, figures of merit, and generalized cross constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 877–892, 1989.

[19] J. Boutros, E. Viterbo, C. Rastello, and J.-C. Belfiore, "Good lattice constellations for both rayleigh fading and gaussian channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 502–518, 1996.

[20] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2013, pp. 332–336.

[21] Y. L.-L. Liu Zilong, "Sparse or Dense: A comparative study of code-domain NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 4768–4780, 2021.

[22] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc, IEEE 80th Veh. Technol. Conf. (VTC-Fall)*, 2014, pp. 1–5.

[23] D. Cai, P. Fan, X. Lei, Y. Liu, and D. Chen, "Multi-dimensional SCMA codebook design based on constellation rotation and interleaving," in *Proc, IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, 2016, pp. 1–5.

[24] Z. Mheich, L. Wen, P. Xiao, and A. Maaref, "Design of SCMA codebooks based on golden angle modulation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1501–1509, 2018.

[25] Y.-M. Chen and J.-W. Chen, "On the design of near-optimal sparse code multiple access codebooks," *IEEE Trans. Commun*, vol. 68, no. 5, pp. 2950–2962, 2020.

[26] X. Zhang, D. Zhang, G. Yang, H.-H. Chen, and D. Zhang, "SCMA codebook design based on uniquely decomposable constellation groups," *IEEE Trans. Wireless Commun*, vol. 20, no. 8, pp. 4828–4842, 2021.

[27] L. Yu, P. Fan, D. Cai, and Z. Ma, "Design and analysis of SCMA codebook based on star-qam signaling constellations," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 543–10 553, 2018.

[28] X. Li, Z. Gao, Y. Gui, Z. Liu, P. Xiao, and L. Yu, "Design of power-imbalanced SCMA codebook," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2140–2145, 2022.

[29] Q. Luo, Z. Liu, G. Chen, and P. Xiao, "Enhancing signal space diversity for scma over rayleigh fading channels," *IEEE Transactions on Wireless Communications*, vol. 23, no. 4, pp. 3676–3690, 2024.

[30] C. Zhang, Y. Luo, and Y. Chen, "A low-complexity SCMA detector based on discretization," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2333–2345, 2018.

[31] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1473–1476, 2016.

[32] J. Boutros and E. Viterbo, "Signal space diversity: a power- and bandwidth-efficient diversity technique for the rayleigh fading channel," *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1453–1467, 1998.

[33] X. Li, Z. Gao, Y. Gui, Z. Liu, P. Xiao, and L. Yu, "Design of power-imbalanced SCMA codebook," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2140–2145, 2022.

[34] M. Vameghestahbanati, I. D. Marsland, R. H. Gohary, and H. Yanikomeroglu, "Multidimensional constellations for uplink SCMA systems—a comparative study," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2169–2194, 2019.

[35] ——, "Multidimensional constellations for uplink SCMA systems—a comparative study," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2169–2194, 2019.

[36] S. Zhang, K. Xiao, B. Xiao, Z. Chen, B. Xia, D. Chen, and S. Ma, "A capacity-based codebook design method for sparse code multiple access systems," in *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*, 2016, pp. 1–5.

[37] C. Yan, G. Kang, and N. Zhang, "A dimension distance-based SCMA codebook design," *IEEE Access*, vol. 5, pp. 5471–5479, 2017.

[38] J. Bao, Z. Ma, Z. Ding, G. K. Karagiannidis, and Z. Zhu, "On the design of multiuser codebooks for uplink SCMA systems," *IEEE Communications Letters*, vol. 20, no. 10, pp. 1920–1923, 2016.

[39] J. Bao, Z. Ma, M. A. Mahamadu, Z. Zhu, and D. Chen, "Spherical codes for SCMA codebook," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016, pp. 1–5.

[40] Y. Zhou, Q. Yu, W. Meng, and C. Li, "SCMA codebook design based on constellation rotation," in *Proc, IEEE International Conf. Commun (ICC)*, 2017, pp. 1–6.

[41] M. Alam and Q. Zhang, "Performance study of scma codebook design," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–5.

[42] ——, "Designing optimum mother constellation and codebooks for SCMA," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[43] K. Lai, J. Lei, L. Wen, G. Chen, W. Li, and P. Xiao, "Secure transmission with randomized constellation rotation for downlink sparse code multiple access system," *IEEE Access*, vol. 6, pp. 5049–5063, 2018.

[44] L. Yu, P. Fan, D. Cai, and Z. Ma, "Design and analysis of SCMA codebook based on Star-QAM signaling constellations," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 543–10 553, 2018.

[45] L. Li, Z. Ma, P. Z. Fan, and L. Hanzo, "High-dimensional codebook design for the SCMA down link," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10 118–10 122, 2018.

[46] X. Li, Z. Gao, Y. Gui, Z. Liu, P. Xiao, and L. Yu, "Design of power-imbalanced SCMA codebook," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2140–2145, 2022.

[47] T. Ericson and V. Zinoviev, "Spherical codes generated by binary partitions of symmetric pointsets," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 107–129, 1995.

[48] M. Beko and R. Dinis, "Designing good multi-dimensional constellations," *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 221–224, 2012.

[49] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[50] F. E. Oggier, E. Bayer-Flückiger, and E. Viterbo, "New algebraic constructions of rotated cubic lattice constellations for the rayleigh fading channel." in *ITW*, 2003, pp. 263–266.

[51] J. Boutros and E. Viterbo, "Signal space diversity: a power- and bandwidth-efficient diversity technique for the Rayleigh fading channel," *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1453–1467, 1998.

[52] X. Li, Z. Gao, Y. Gui, Z. Liu, P. Xiao, and L. Yu, "Design of power-imbalanced SCMA codebook," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2140–2145, 2022.

[53] A. Chockalingam and B. S. Rajan, *Large MIMO Systems*. Cambridge University Press, 2014.

[54] S.-C. Lim, N. Kim, and H. Park, "Uplink SCMA system with multiple antennas," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6982–6992, 2017.

[55] S. Elkawafi, A. Younis, and R. Mesleh, "Performance analysis of sparse code multiple access MIMO systems," in *2019 IEEE 30th Int. Symp. Per. Indoor Mobile Radio Commun (PIMRC)*, 2019, pp. 1–6.

[56] Z. Pan, J. Luo, J. Lei, L. Wen, and C. Tang, "Uplink spatial modulation SCMA system," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 184–187, 2019.

[57] C. Zhang, Y. Luo, and Y. Chen, "A Low-complexity SCMA Detector based on Discretization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2333–2345, 2018.

[58] L. Yang, Y. Liu, and Y. Siu, "Low complexity message passing algorithm for SCMA system," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2466–2469, 2016.

[59] J. Dai, G. Chen, K. Niu, and J. Lin, "Partially active message passing receiver for MIMO-SCMA systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 222–225, 2018.

[60] P. Wang, L. Liu, S. Zhou, G. Peng, S. Yin, and S. Wei, "Near-optimal MIMO-SCMA uplink detection with low-complexity expectation propagation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1025–1037, 2020.

[61] Z. Mheich, I. A. Hemadeh, Z. Liu, and P. Xiao, "Low-complexity expectation propagation detection for uplink MIMO-SCMA systems," in *2020 International Conference on UK-China Emerging Technologies (UCET)*, 2020, pp. 1–4.

[62] J. Dai, K. Niu, and J. Lin, "Iterative Gaussian-approximated message passing receiver for MIMO-SCMA system," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 753–765, 2019.

[63] F. Wei and W. Chen, "Low complexity iterative receiver design for sparse code multiple access," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 621–634, 2017.

[64] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 2015.

[65] Q. Zhang, J. Wang, and Y. Wang, "Efficient QAM signal detector for massive MIMO systems via PS/DPS-ADMM approaches," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8859–8871, 2022.

[66] 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.

[67] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[68] L. Liu, S. Huang, and B. M. Kurkoski, "Memory AMP," *IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 8015–8039, 2022.

[69] Y. Chen, L. Liu, Y. Chi, Y. Li, and Z. Zhang, "Memory AMP for Generalized MIMO: Coding Principle and Information-Theoretic Optimality," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[70] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Foundations and Trends, 2011.

[71] C. F. Zhouchen Lin, Huan Li, *Alternating Direction Method of Multipliers for Machine Learning*. Springer, 2022.

[72] S. Barman, X. Liu, S. C. Draper, and B. Recht, "Decomposition methods for large scale LP decoding," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7870–7886, 2013.

[73] H. Wei and A. H. Banihashemi, "ADMM check node penalized decoders for LDPC codes," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3528–3540, 2021.

[74] A. C. Cirik, N. Mysore Balasubramanya, and L. Lampe, "Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 46–49, 2018.

[75] S. Shahabuddin, I. Hautala, M. Juntti, and C. Studer, "ADMM-based infinity-norm detection for massive MIMO: Algorithm and VLSI architecture," *IEEE IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 4, pp. 747–759, 2021.

[76] H. Yaoyue, P. Zhiwen, L. Nan, and Y. Xiaohu, "Multidimensional constellation design for spatial modulated SCMA systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 8795–8810, 2021.

[77] C. Studer, A. Burg, and H. Bolcskei, "Soft-output sphere decoding: algorithms and VLSI implementation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 290–300, 2008.

[78] T. Cui and C. Tellambura, "An efficient generalized sphere decoder for rank-deficient MIMO systems," *IEEE Commun. Lett.*, vol. 9, no. 5, pp. 423–425, 2005.

[79] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, 2005.

[80] V. Vikas, A. Rajesh, K. Deka, and S. Sharma, "A comprehensive technique to design SCMA codebooks," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1735–1739, 2022.

[81] P. Som, T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Low-complexity detection in large-dimension MIMO-ISI channels using graphical models," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 8, pp. 1497–1511, 2011.

[82] C. Wang, E. K. Au, R. D. Murch, W. H. Mow, R. S. Cheng, and V. Lau, "On the performance of the MIMO zero-forcing receiver in the presence of channel estimation error," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 805–810, 2007.

[83] 3GPP, "Numbering, addressing and identification," no. 23.003. [Online]. Available: http://www.3gpp.org/DynaReport/23003.htm

[84] J. Chen, A. Dholakia, E. Eleftheriou, M. Fossorier, and X.-Y. Hu, "Reduced-complexity decoding of LDPC codes," *IEEE Trans. on Commun.*, vol. 53, no. 8, pp. 1288–1299, 2005.

[85] M. Mohammadkarimi, M. A. Raza, and O. A. Dobre, "Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications," *IEEE Vehicular Technology Magazine*, vol. 13, no. 4, pp. 40–50, 2018.

[86] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2320–2323, 2016.

[87] Y. Du, B. Dong, Z. Chen, X. Wang, Z. Liu, P. Gao, and S. Li, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2812–2828, 2017.

[88] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free noma," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 7894–7909, 2018.

[89] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Communications Letters*, vol. 21, no. 3, pp. 640–643, 2017.

[90] 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 Processing Magazine*, vol. 35, no. 5, pp. 88–99, 2018.

[91] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Communications Letters*, vol. 16, no. 7, pp. 972–974, 2012.

[92] Y. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. United Kingdom: Cambridge University Press, 2012.

[93] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[94] Y. Mei, Z. Gao, Y. Wu, W. Chen, J. Zhang, D. W. K. Ng, and M. Di Renzo, "Compressive sensing-based joint activity and data detection for grant-free massive iot access," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1851–1869, 2022.

[95] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[96] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[97] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999, vol. 4.

[98] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[99] A. H. Sayed, *Inference and Learning from Data: Foundations*. Cambridge University Press, 2022.

[100] Y. Wang and W. Yin, "Sparse signal reconstruction via iterative support detection," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 462–491, 2010. [Online]. Available: https://doi.org/10.1137/090772447

[101] W. Lu and N. Vaswani, "Regularized modified BPDN for noisy sparse reconstruction with partial erroneous support and signal value knowledge," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 182–196, 2012.

[102] R. Hayakawa, "Asymptotic performance prediction for ADMM-based compressed sensing," *IEEE Trans. Signal Process.*, vol. 70, pp. 5194–5207, 2022.

[103] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2018.

[104] C. Shang, Y. Sun, H. Luo, and M. Guizani, "Computation offloading and resource allocation in NOMA–MEC: A deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 10, no. 17, pp. 15 464–15 476, 2023.

[105] P. Liu, K. An, J. Lei, W. Liu, Y. Sun, G. Zheng, and S. Chatzinotas, "SCMA-enabled multi-cell edge computing networks: Design and optimization," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7987–8003, 2023.

[106] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.