NOMA Integrated with Enabling Technologies and Practical Challenges

by

Aditya Rajasekaran, B.Eng, M.Eng

A dissertation submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
August 2023

©Copyright
Aditya Rajasekaran, 2023
Abstract

Non-orthogonal multiple access (NOMA) schemes allow multiple users to share the same resource and separate the users in either the power-domain (PD-NOMA) or in the code domain (CD-NOMA). To make NOMA a reality in networks, several important practical challenges need to be considered. This thesis addresses some of these challenges in both the PD-NOMA and CD-NOMA space. In PD-NOMA systems integrated with mmWave technology, the thesis considers the practical constraint of the end user processing capabilities, modelled through an successive interference cancellation (SIC) decoding capability constraint that captures the number of other user’s signals any given user can decode in the SIC decoding procedure. 6G networks are expected to include a mix of users of varying processing capabilities, all needing to access the same spectrum. To solve the rate maximization problem when each user has different processing capabilities, the thesis proposes low-complexity heuristics to maximize the sum-rate after factoring in each individual users SIC decoding capability constraint. However, these algorithms are based on the instantaneous channel conditions of users and need to be run on a millisecond granularity. To address this, a machine learning based neural network approach is proposed that takes this complexity offline where the neural network is trained on simulated and past network data, and the trained network is directly applied to solve the clustering problem in live networks. The thesis also addresses the practical constraint around the availability of CSI by exploiting the growing field of integrated communication and sensing solutions using a camera equipped base station to aid the user clustering process in NOMA. In CD-NOMA systems using the widely promoted sparse code multiple access (SCMA) scheme for uplink (UL) NOMA, the thesis studies the PAPR problem in UL SCMA-OFDM systems. A novel link between the obtained PAPR statistics and the SCMA modulation scheme and the placement of the sub-carriers (SC’s) that carry the SCMA codewords is presented. The thesis highlights unique opportunities that SCMA-OFDM systems present to the widely studied PAPR problem due to the
statistical dependency between the OFDM SCs carrying the codewords as opposed to traditional OFDM systems where the SCs are independently modulated.

Non-orthogonal multiple access (NOMA) schemes allow multiple users to share the same resource (e.g., a time/frequency resource block) and separate the users in either the power-domain, called power-domain NOMA (PD-NOMA) or in the code domain, called code-domain NOMA (CD-NOMA). Interest has been growing in academia, industry and standardization bodies like the 3rd generation partnership project (3GPP) to adopt NOMA. However, to make NOMA a reality, several important practical constraints and limitations need to be considered. This thesis aims to address some of these challenges in both the PD-NOMA and CD-NOMA space.

The thesis starts by examining challenges in integrating PD-NOMA systems with mmWave technology when factoring in the important practical constraint of the end user processing capabilities. An SIC decoding capability constraint is introduced that captures the number of other user’s signals any given user can decode in the SIC decoding procedure of PD-NOMA. 6G networks are expected to include a large mix of users of varying processing capabilities, all needing to access the same spectrum. To solve the rate maximization problem when each user can have different processing capabilities, the thesis proposes low-complexity heuristics to maximize the sum-rate after factoring in each individual users SIC decoding capability constraint. However, these algorithms are based on the instantaneous channel conditions of users and so need to be run on a millisecond granularity. To address this, a machine learning based neural network approach is proposed that takes this complexity offline where the neural network is trained on simulated and past network data, and the trained network can be directly applied to solve the clustering problem in live networks. Another important practical constraint that can limit the feasibility of these clustering schemes is the availability of CSI. As a result, the thesis also proposes a vision based approach to the user clustering problem enabled by camera equipped base stations that allow the user clustering to be done exclusively through images or jointly in conjunction with partially available CSI.

Switching to CD-NOMA systems using the widely promoted sparse code multiple access (SCMA) scheme for uplink (UL) NOMA, the thesis studies the important PAPR problem and how that can be tackled in UL SCMA-OFDM systems. Concretely, a novel link between the obtained PAPR statistics and the SCMA modulation scheme and the placement of the SC’s that carry the SCMA codewords is presented.
This chapter of the thesis highlights unique opportunities that SCMA-OFDM systems present to the widely studied PAPR problem due to the statistical dependency between the OFDM SCs carrying the codewords as opposed to traditional OFDM systems where the SCs are independently modulated.
Dedicated to my parents (Sethuramalingam Rajasekaran and Saraswathy Rajasekaran), my wife (Megha) and my daughter (Aanya).
Acknowledgments

I would like to express sincere gratitude to my supervisor, Prof. Halim Yanikomeroglu, for (a) believing in me and giving me the opportunity to pursue this Ph.D. degree while studying part-time, and (b) motivating me and providing continuous support throughout my Ph.D. study and research. I would like to acknowledge the people I got to collaborate with at different points in this journey. Particularly, I want to thank Monirosharhe Vameghestahbanati who provided me guidance during my early days and also Omar Maraqa, with whom I got to collaborate extensively on a number of different papers during my Ph.D. I would also like to acknowledge Ericsson Canada, to whom I was employed throughout this degree, for providing tuition assistance and encouragement throughout the years.

I am deeply grateful to my parents for all their support and encouragement over many years to get me to this thesis. In particular, I would like to dedicate this thesis to my late father who always encouraged me towards higher studies throughout my life. I would like to thank my wife, Megha, for her patience and support through these final years while I completed this degree in parallel with my work and family; this would not have been possible if not for this. Last but not least, I would like to thank my daughter, Aanya, who was born somewhere in the middle of my Ph.D. study. At the time of writing, you are only just shy of two years old, but somehow you provided me that final push to get me over the line.
# Table of Contents

Abstract ii

Acknowledgments vi

Table of Contents vii

List of Tables x

List of Figures xi

Nomenclature xv

1 Introduction 1

1.1 Motivation ........................................... 1
1.2 Thesis Contributions ................................... 4
1.3 List of Publications .................................... 6
1.3.1 Papers Included in the Thesis ......................... 6
1.3.2 Papers Not Included in the Thesis ....................... 6
1.3.3 Patents Granted as a Result of Work in the Thesis .... 7
1.4 Organization of Thesis .................................. 7

2 PD-NOMA Integrated with MIMO, mmWave and ML: Background and Survey 8

2.1 Enabling Technologies ................................... 10
2.1.1 MISO, MIMO, and mMIMO Communications ................ 10
2.1.2 Advanced Multi-antenna Architectures .................. 11
2.1.3 mmWave and THz Communications ..................... 11
2.1.4 Machine Learning .................................... 12
2.2 Survey of rate optimization schemes with PD-NOMA integrated with enabling technologies ........................................... 14
  2.2.1 Rate-optimal NOMA with MISO, MIMO, and mMIMO Communications .................................................. 14
  2.2.2 Rate-optimal NOMA with Advanced Multi-antenna Architectures ...................................................... 19
  2.2.3 Rate-optimal NOMA with mmWave and THz Communications ......................................................... 23
  2.2.4 Rate Optimization of NOMA-enabled Systems with Machine Learning .................................................. 27

3 User Clustering in mmWave-NOMA Systems with User Decoding Capability Constraints for B5G Networks 31
  3.1 Introduction .................................................................................................................................................. 32
  3.1.1 Related Work ......................................................................................................................................... 34
  3.1.2 Motivation and Contributions ............................................................................................................. 34
  3.2 System Model and Problem Formulation .................................................................................................... 37
  3.3 Proposed Algorithm(s) ............................................................................................................................... 48
  3.4 Simulation Results and Discussion ............................................................................................................ 53
  3.5 Conclusion .................................................................................................................................................. 63

4 Neural Network Aided User Clustering in mmWave-NOMA Systems with User Decoding Capability Constraints 64
  4.1 Introduction .................................................................................................................................................. 64
  4.1.1 Related Work ......................................................................................................................................... 65
  4.1.2 Motivation and Contributions ............................................................................................................. 66
  4.2 System Model and Problem Formulation .................................................................................................... 69
  4.3 Proposed Algorithm(s) ............................................................................................................................... 74
    4.3.1 Neural Network Structure ................................................................................................................. 74
    4.3.2 Training and Testing ............................................................................................................................ 77
  4.4 Simulation Results and Discussion ............................................................................................................ 86
  4.5 Conclusion .................................................................................................................................................. 95

5 Vision-Assisted User Clustering for Robust mmWave-NOMA 97
  5.1 Introduction .................................................................................................................................................. 98
  5.2 Why Camera-equipped BS in mmWave-NOMA Clustering? ..................................................................... 100
5.2.1 Dimension Space for Possible Clustering Approaches in mmWave-NOMA Systems ........................................ 100
5.2.2 Role of the Vision Dimension in Clustering ................. 102
5.3 A Case Study: Robust User Clustering in mmWave-NOMA via Vision-Based Deep Learning .................................. 103
5.4 Challenges and Future Directions ..................................... 108
5.5 Conclusion ..................................................................... 109

6 PAPR in Uplink SCMA-OFDM Systems: The Impact of Modulation Schemes and Sub-carrier Allocation 111
6.1 Introduction ..................................................................... 112
6.2 System Model: Uplink SCMA-OFDM Transmission ............ 116
6.3 PAPR in Traditional OFDM Systems vs. PAPR in SCMA-OFDM Systems .............................................................. 118
6.3.1 PAPR in OFDM systems .............................................. 120
6.3.2 PAPR in SCMA-OFDM Systems .................................... 122
6.4 Resource Allocation Schemes ........................................... 125
6.4.1 Constructing the Diversity Allocation Scheme ............... 126
6.5 Simulation Results .......................................................... 127
6.6 Exploiting Statistical Dependency in PAPR Reduction Schemes Based on Multiple Signalling .............................. 131
6.7 Appendix: Zero dB PAPR Constellation Design .................. 137

7 Summary and Future Research Directions 139
7.1 Summary ......................................................................... 139
7.2 Future Research Directions .............................................. 141
7.2.1 User Clustering in PD-NOMA Considering SIC Decoding Capabilities in Networks Using Conventional BSs without Camera 141
7.2.2 User Clustering in PD-NOMA Considering SIC Decoding Capabilities in Networks Using CBSs Exploiting nonCSI Feedback 142
7.2.3 PAPR Analysis and Reduction in SCMA-OFDM Systems .... 143

List of References ................................................................. 145
## List of Tables

2.1 MISO-NOMA communications systems: current status of rate optimization schemes. ................................................................. 15
2.2 MIMO-NOMA communications systems: current status of rate optimization schemes. ................................................................. 16
2.3 mMIMO-NOMA, advanced multi-antenna architectures (cell-free mMIMO-NOMA, and reconfigurable antenna NOMA-enabled scheme): current status of rate optimization schemes. ................................................................. 20
2.4 mmWave-NOMA and THz-NOMA systems: current status of rate optimization schemes. ................................................................. 24
3.1 List of parameters. ........................................................................................................................................................................... 37
3.2 Simulation parameters ..................................................................................................................................................................... 53
4.1 Comparison of this work with existing literature .......................................................................................................................... 65
4.2 Simulation parameters ..................................................................................................................................................................... 87
List of Figures

2.1 An illustration of a two-user downlink power-domain NOMA scheme with superposition coding and successive interference cancellation decoding [17]. ................................. 9
2.2 The different categories of machine learning algorithms with the main tasks/features, some example algorithms, and some example applications in wireless communications [32]. ....................... 13
3.1 The ABF scheme used by the BS is illustrated on the left with a fixed set of $B+1$ precoding weights, creating $B+1$ candidate beam directions to choose from when serving a NOMA cluster of users in one orthogonal channel. On the right, the $N$ users, each with their own SIC decoding capability, is illustrated. ............................. 40
3.2 Performance of NOMA-BB for a homogeneous system where each users SIC decoding capability constraint $d = d_{\text{max}}$, effectively making $d_{\text{max}}$ the maximum number of users that can be placed in a cluster. (a) Overall spectral efficiency compared to OMA (b) Number of clusters, $K$, required to serve the $N$ users. ................................. 55
3.3 Performance of NOMA-MEC for a heterogeneous system where each user has its own decoding capability constraints. NOMA-MEC is compared against OMA and NOMA-BB-Het, a modified version of NOMA-BB to account for the heterogeneous user decoding capabilities. .......................... 56
3.4 NOMA-MEC run in a heterogeneous system with users having their SIC decoding capability, $d$, randomly distributed in $[0, d_{\text{max}}]$, is compared against NOMA-BB and NOMA-MEC run for a homogeneous system where all users have $d = d_{\text{max}}$ ($d_{\text{max}} = 5$). ................................. 57
3.5 Analyzing NOMA-MEC in terms of the number of clusters, $K$, outputted by the algorithm, as the number of heterogeneous users in the system increases for different values of $b$. ................................. 58
3.6 Analyzing the impact of $b$ on the performance of NOMA-MEC as system parameters $B$ and $M$ vary. .................................................. 60
3.7 Impact of the number of beams per user set, $b$, on the complexity of NOMA-MEC in terms of (a) $I_1$ which represents the number of cluster combinations examined in step-1 of NOMA-MEC and (b) $I_2$, which represents the number of valid combinations in $C_v$ to be considered by the greedy algorithm to solve the MEC problem. ............................ 62
4.1 Illustrating the proposed ANN-NOMA-MEC and ANN-NOMA-BB networks learning offline from labelled training data obtained using the NOMA-MEC and NOMA-BB algorithms and then applied to real user inputs in a live network. ................................. 67
4.2 Proposed neural network structure for ANN-NOMA-MEC and ANN-NOMA-BB. ................................................................. 74
4.3 Illustrating the fully connected structure of the hidden neurons in the proposed neural network. .................................................. 76
4.4 Effect of the number of neurons in the hidden layer, $H$, on the performance of ANN-NOMA-MEC for different numbers of users, $N$, with $b = 2$. ................................................................. 88
4.5 Impact of the learning rate, $\alpha$, on MSE and $R_{\text{sum}}$ for different numbers of users, $N$, at (a) $n_{\text{epochs}} = 1$ and (b) $n_{\text{epochs}} = 10$. ANN-NOMA-MEC with $b = 2$. ................................................................. 89
4.6 $R_{\text{sum}}$ at different learning epochs, $e$, and different training data samples, $T_{\text{train}}$. ANN-NOMA-MEC with $N = 200$ users and $b = 2$. ................................. 90
4.7 Analyzing the effect of the number of training data samples $T_{\text{train}}$ on MSE and $R_{\text{sum}}$ considering different numbers of users, $N$. ANN-NOMA-MEC with $b = 2$. ................................................................. 91
4.8 Evolution of ANN-NOMA-MEC learning through 20 epochs with $N = 200$ and $b = 2$. Training in offline mode. ......................... 92
4.9 Evaluation of different ANN-NOMA-MEC models trained offline with different configurations of $N$ and $b$, benchmarked with the NOMA-MEC heuristic. (a) Performance $R_{\text{sum}}$ and (b) Number of user clusters, $K$. ................................................................. 93
4.10 Evaluation of different ANN-NOMA-BB models trained offline with different number of users, $N$, benchmarked with the NOMA-BB heuristic for different values of the SIC decoding capability, $d_{\text{max}}$. (a) Performance $R_{\text{sum}}$ and (b) Number of user clusters, $K$. .......................... 94

4.11 Comparison of a brute force search (BFS) with NOMA-MEC and ANNs trained on both the NOMA-MEC heuristic and the BFS. .................. 95

5.1 Illustrating the dimensions of feedback that a CBS can exploit for NOMA clustering decisions, categorized into CSI-based feedback and non-CSI-based feedback, comprised of UE location and BS camera feed images. .......................... 100

5.2 High-level illustration of the proposed NOMA clustering scheme based on non-CSI-based feedback involving the images captured by the camera feed of the BS and UE location input. .......................... 101

5.3 Illustrating the NOMA-BB algorithm, where from a codebook of beams each users best beam is determined and users that share the same best beam are clustered together. .......................... 104

5.4 Simulation results highlighting the close performance of the proposed DL approach using camera and location feed (non-CSI feedback) with the clustering schemes that use CSI feedback. .......................... 107

6.1 The transmitter for User-1 in an SCMA-OFDM system where 6 users share 4 SCs, with 2 non-zero SCs allocated to each user. .......................... 117

6.2 The 4-point SCMA constellations used in this study. .......................... 119

6.3 Illustration of the power peaks produced when summing sinusoids of evenly spaced frequencies as is the case in an OFDM signal. .......................... 121

6.4 Illustration of the diversity allocation mapping from the virtual view of SCMA blocks (also the regular allocation) to the physical OFDM SCs. 125

6.5 Comparing 4-LDS, 4-Bao and 4-OPP with different SC allocation schemes. For illustration, the curve when independent 4-QAM symbols (non-SCMA symbols) are transmitted in the same data carrying SCs is included. .......................... 129

6.6 Comparing the PAPR statistics with different choices of SCMA modulation schemes and SC allocation strategies. These options are static configuration parameters that impact the PAPR statistics in SCMA-OFDM systems. .......................... 132
6.7 Block diagrams highlighting the PAPR reduction techniques of SLM and IL that are described in the context of SCMA-OFDM systems.  . 133

6.8 Comparing the PAPR reduction achieved with SLM in an SCMA-OFDM system with 4-Bao constellation and different SC allocation strategies and different values of \( R \) for the SLM reduction.  . . . . . . . 135

6.9 Comparing the PAPR reduction achieved with interleaving in an SCMA-OFDM system with 4-Bao constellation and different SC allocation strategies.  . . . . . . . . . . . . . . . . . . . . . . . . . . . 136

6.10 Proposed M-0PAPR scheme that has zero dB constellation PAPR.  . 138
# Nomenclature

## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>B5G</td>
<td>Beyond 5G</td>
</tr>
<tr>
<td>BS</td>
<td>Base Station</td>
</tr>
<tr>
<td>CBS</td>
<td>Camera equipped base station</td>
</tr>
<tr>
<td>CD-NOMA</td>
<td>Code Domain - NOMA</td>
</tr>
<tr>
<td>DL</td>
<td>Downlink</td>
</tr>
<tr>
<td>eMBB</td>
<td>Enhanced mobile broadband</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>mMIMO</td>
<td>Massive Multiple Input Multiple Output</td>
</tr>
<tr>
<td>mMTC</td>
<td>massive machine type connectivity</td>
</tr>
<tr>
<td>mmWave</td>
<td>millimeter wave</td>
</tr>
<tr>
<td>mmWave-NOMA</td>
<td>A NOMA system operated in the mmWave spectrum</td>
</tr>
<tr>
<td>NOMA</td>
<td>Non-Orthogonal Multiple Access</td>
</tr>
<tr>
<td>NOMA-MEC</td>
<td>NOMA-Minimum Exact Cover</td>
</tr>
<tr>
<td>NOMA-BB</td>
<td>NOMA-Best Beam</td>
</tr>
<tr>
<td>NR</td>
<td>New Radio</td>
</tr>
<tr>
<td>OFDM</td>
<td>Orthogonal Frequency Division Multiplexing</td>
</tr>
<tr>
<td>PAPR</td>
<td>Peak-to-Average Power Ratio</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PD-NOMA</td>
<td>Power Domain - NOMA</td>
</tr>
<tr>
<td>SC</td>
<td>Sub-carrier</td>
</tr>
<tr>
<td>SCMA</td>
<td>Sparse Code Multiple Access</td>
</tr>
<tr>
<td>SCMA-OFDM</td>
<td>An SCMA system combined with OFDM</td>
</tr>
<tr>
<td>SIC</td>
<td>Successive Interference Cancellation</td>
</tr>
<tr>
<td>UL</td>
<td>Uplink</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

With the first phase of 5G still targeting the enhanced mobile broadband (eMBB), research is now focused on the promising new use-cases for B5G systems. The massive machine-type connectivity (mMTC) is among the most exciting of these new use-cases, that is also referred to as the Internet of Things (IoT). This paradigm is associated with a many fold increase in the number of connected devices in the network. Further, each of these connected users come with different processing capabilities and requirements. B5G systems need to support a very large number of low-cost devices for the IoT connections in addition to the traditional high data rate mobile broadband connections that are also growing exponentially.

In order to meet the spectral efficiency requirements placed on B5G wireless communication systems from needing to support all these users and their varied requirements, non-orthogonal multiple access (NOMA) has been proposed as a key enabler. NOMA allows multiple users to share the same resource (e.g., a time/frequency resource block) and separate the users in other domains with some additional receiver complexity [1]. When the power domain is used to separate the users, it is referred to as the power-domain NOMA (PD-NOMA) [2] scheme. Alternatively, if the users are separated through non-orthogonal codes, it is referred to as code-domain NOMA (CD-NOMA) [3,4]. PD-NOMA has also been studied in conjunction with CD-NOMA recently [5]. Also, NOMA is a technology that can be combined with several other technology enablers for B5G systems like massive MIMO systems, operating in high frequency mmWave and THz bands, coordinated multi-point etc. In the thesis, we survey the integration of PD-NOMA with some of these technologies and explore
practical challenges that come up with realizing these deployments.

NOMA has been considered as a candidate multiple access scheme in various standardization activities. For long term evolution advanced (LTE-A) systems (3GPP Release 13), NOMA was considered under the name of multi-user superposition transmission (MUST) [6]. Furthermore, in LTE-A Pro (3GPP Release 14), the standardization body recognized that at least uplink NOMA schemes should be considered especially for mMTC [7]. In 5G new radio (NR) phases (1 & 2) (3GPP Release 15 and Release 16) [8,9], multiple studies on the advancements needed in the transmitter and receiver sides for adopting NOMA schemes have been proposed. However, NOMA has not yet found widespread adoption because many practical challenges still exist for deploying it on a large scale, whether that be PD-NOMA or CD-NOMA. The focus of this thesis then is to define and solve a few key practical problems with deploying NOMA.

For PD-NOMA systems, the thesis has a particular focus on the mmWave spectrum which offers a large amount of bandwidth to scale up the capacity from the cellular networks that operate today in the sub-6 GHz range. As described before, using PD-NOMA, multiple users can be served in the same orthogonal resource, e.g., time, frequency, orthogonal frequency division multiplexing (OFDM) resource block (RB), etc., by separating the users in the power domain instead. Hence, when combined, mmWave-NOMA has the potential to serve the high rates and massive connectivity demands of B5G networks. In downlink (DL) PD-NOMA systems operated in the mmWave band, user channels get highly correlated which can be exploited in mmWave-NOMA systems to cluster a set of “correlated” users together [10–12]. Identifying the set of users to cluster greatly affects the viability of DL NOMA systems. As we group users in NOMA clusters, the user with the weakest DL channel only has to decode its own signal, while the user with the strongest DL channel (least pathloss) has to decode the signals of all other users in the successive interference cancellation (SIC) procedure. The decoding of other users’ signals requires significant additional processing capability, in terms of hardware capability, energy consumption, etc. [13,14]. The authors in [15] identified this SIC decoding complexity as the first major practical implementation issue for NOMA. NOMA is expected to support a wide variety of end-user devices in B5G systems, each with different signal processing capabilities [15,16]. Hence, each user has its own limitations on the number of other users signals that it can decode. In this thesis, we define these limitations through
an "SIC decoding capability" parameter. We envision a system where users report their SIC decoding capability to the base station (BS), i.e., the number of other users signals a user is capable of decoding in the SIC procedure. The BS can use this information to make clustering decisions using low-complexity heuristics that are proposed in the thesis.

The problem with applying optimization techniques for PD-NOMA clustering decisions, even low-complexity heuristics, in live mmWave-NOMA networks for user clustering is that they require a very large number of computation steps to make a clustering decision. If these clustering decisions are based on the instantaneous channel of hundreds of users, it becomes prohibitively complex to implement in practical systems on a millisecond granularity as required by beyond 5G (B5G) systems. To address this issue, the thesis also proposes a computationally efficient two-stage machine learning based approach using neural networks to solve the cluster assignment problem in a millimeter wave-non orthogonal multiple access (mmWave-NOMA) system where each user’s individual SIC decoding capabilities are taken into consideration. An artificial neural network (ANN) is applied in real time to assign users to clusters taking each user’s instantaneous channel state information (CSI) and SIC decoding capabilities as inputs. The algorithm is trained offline on cloud resources, i.e., not using the base station (BS) compute resources. This training is done using a dataset obtained by offline computation of input parameters using the heuristics proposed in this thesis.

Typically, in mmWave-NOMA systems, only channel state information (CSI) is used to make these clustering decisions. When any problem arises in accessing up-to-date and accurate CSI, user clustering will not properly function due to its hard-dependency on CSI, and obviously, this will negatively affect the robustness of the NOMA systems. To improve the robustness of the NOMA systems, the thesis also proposes utilizing emerging trends such as location-aware and camera-equipped base stations (CBSs) which do not require any extra radio frequency resource consumption. Three different dimensions of feedback that a CBS can benefit from to solve the user clustering problem are explored, categorized as CSI-based feedback and nonCSI-based feedback, where the non-CSI-based feedback is comprised of user equipment (UE) location and the CBS camera feed. The thesis investigates how the vision assistance of a CBS can be used in conjunction with other dimensions of feedback to make clustering decisions in various scenarios.
Finally, switching to the code-domain NOMA, sparse code multiple access (SCMA) is a code-domain NOMA uplink solution that overloads resource elements (RE’s) with more than one user. Given the success of orthogonal frequency division multiplexing (OFDM) systems, SCMA will likely be deployed as a multiple access scheme over OFDM, called an SCMA-OFDM system. One of the major challenges with OFDM systems is the high peak-to-average power ratio (PAPR) problem, which is typically studied through the PAPR statistics for a system with a large number of independently modulated sub-carriers (SCs). This chapter of the thesis proposes novel aspects to studying the PAPR statistics for SCMA-OFDM systems that is different from the vast body of existing PAPR literature in the context of traditional OFDM systems. The main difference lies in the fact that the SCs are not independently modulated in SCMA-OFDM systems. Instead, the SCMA codebook uses multi-dimensional constellations, leading to a statistical dependency between the data carrying SCs. Further, the SCMA codebook dictates that an UL user can only transmit on a subset of the available SCs. We highlight the joint effect of the two major factors that influence the PAPR statistics - the phase bias in the multi-dimensional constellation design along with the resource allocation strategy. The choice of modulation scheme and SC allocation strategy are static configuration options, thus allowing for PAPR reduction opportunities in SCMA-OFDM systems through the setting of static configuration parameters. Compared to the class of PAPR reduction techniques in the OFDM literature that rely on multiple signalling and probabilistic techniques, these gains come with no computational overhead.

1.2 Thesis Contributions

The contributions of this thesis can be summarized as follows:

- **User clustering in mmWave PD-NOMA systems**
  The thesis address two important practical challenges related to the user clustering problem in mmWave PD-NOMA systems, namely incorporating the end-user SIC processing capabilities into the problem and secondly, addressing the availability of channel state information for user clustering.

  - In Chapter 3, an SIC decoding capability constraint is introduced that captures the number of other user’s signals any given user can decode in the
SIC decoding procedure of PD-NOMA. Incorporating the SIC decoding capability, a rate maximization problem is framed for mmWave-NOMA systems. Two low-complexity heuristics, namely NOMA-MEC and NOMA-BB, are respectively proposed to solve the user clustering aspect of the rate maximization problem for heterogeneous systems where each user can be modeled with its own SIC decoding capability constraint and homogeneous systems where all users have the same SIC decoding capability constraint.

- The algorithms presented in Chapter 3, while successfully incorporating the SIC decoding constraint into a low-complexity user clustering heuristic, are based on the instantaneous channel conditions of users and so need to be run on a millisecond granularity as the users channels change. To address this, in Chapter 4, a machine learning based artificial neural network (ANN) approach is proposed that takes this complexity offline where the neural network is trained on simulated and past data, and the trained network is directly applied to solve the clustering problem in live networks. These ANNs are trained offline using data obtained from the heuristics presented in Chapter 3, and so are termed ANN-NOMA-MEC and ANN-NOMA-BB, respectively.

- In Chapter 5, we address the issues related to CSI in user clustering problems in mmWave-NOMA systems and propose utilizing emerging trends such as location-aware and camera-equipped base stations (CBSs) to aid the user clustering step instead of, or, in conjunction with CSI.

**PAPR problem in SCMA CD-NOMA systems**

Here, the thesis focuses on CD-NOMA systems and concretely the sparse code multiple access that has been promoted in the literature for UL systems. Concretely, we focus on the PAPR problem:

- In Chapter 6, we highlight some novel aspects to the PAPR statistics for SCMA-OFDM systems that is different from the vast body of existing PAPR literature in the context of traditional OFDM systems. The thesis shows that the use of SMCA-OFDM systems allows for new opportunities for PAPR reduction through the choice of the SCMA modulation scheme and SC allocation strategy.
1.3 List of Publications

1.3.1 Papers Included in the Thesis


*This was a joint work and only the relevant sections from the paper where I was the primary driver are included in the thesis.


1.3.2 Papers Not Included in the Thesis


1.3.3 Patents Granted as a Result of Work in the Thesis


1.4 Organization of Thesis

The rest of the thesis is organized as follows. We do a survey of rate optimal PD-NOMA works from the literature when combined with MIMO, mMIMO, mmWave and ML technologies in Chapter 2. Following that, we investigate the user clustering aspect in mmWave-NOMA systems when including the practical constraint of SIC decoding capabilities of the user in Chapter 3. An artificial neural network (ANN) to address the same system model for mmWave-NOMA systems is then proposed in Chapter 4. In Chapter 5, we propose how camera equipped base stations can use the vision dimension to aid user clustering in mmWave-NOMA systems. We then switch to CD-NOMA systems and discuss novel aspects to the well studied PAPR problem when they are applied to SCMA based uplink transmissions in Chapter 6. Finally, a brief summary along with ideas on extending the research done in this thesis is presented in Chapter 7.
Chapter 2

PD-NOMA Integrated with MIMO, mmWave and ML: Background and Survey

In this chapter, we describe the basic workings of PD-NOMA and conduct a literature survey on its integration with MIMO, mmWave and machine learning technologies. These are all enablers we use in subsequent chapters of the thesis in the proposed schemes. In the literature survey, we focus on the types of rate optimization problems framed and how they are broken down in the literature into user clustering and power allocation sub-problems.

The power-domain NOMA (PD-NOMA) concept was first introduced in [18] to improve the spectral efficiency of wireless networks by allowing multiple users to simultaneously share both the time and frequency resources. The theoretical roots of PD-NOMA scheme lie in multi-user information theory, built on the concepts of superposition coding and successive interference cancellation (SIC). A basic two-user downlink NOMA scheme is shown in Fig. 2.1, where User 1 is close to the BS with strong channel gain and User 2 is farther away from the BS with a weaker channel gain. At the transmitter, both signals for the weak and the strong users are superimposed upon each other with different power allocation. The transmitter tends to allocate more power to the weak user as it has a larger path loss, as compared to the strong user. At the strong user receiver, the signal of the weak user has a high signal-to-noise ratio (SNR) which implies that the strong user can successfully decode and subtract the weak user signal before decoding its own signal (i.e., performing SIC). On the other hand, at the weak user receiver, the strong user signal is considered as noise.
as its transmission power is lower than the weak user signal. Subsequently, the weak user can decode its signal directly without SIC [19].

The grouping or clustering of the users to be served in the same resource block is essential in PD-NOMA and is typically carried out as two users per cluster (also known as user pairing) or multiple users per cluster. The selection in basic NOMA scheme, with a single-antenna BS and single-antenna users, is usually based on the instantaneous scalar channel gains and the users are ranked accordingly to allow proper SIC decoding which tends to improve as the channel gain disparity increases unless the channel gain of the weaker user is very small which may render NOMA inefficient. In general, the optimal user clustering requires an exhaustive search and may not be affordable for practical systems and networks with a large number of users [20]. For this reason, researchers resort to low complexity solutions to solve the user clustering problem through heuristic algorithms which may lead to unpredictable results.

In the survey in this chapter, we focus on the characteristics of rate optimization problems from the literature for NOMA systems integrated with MIMO, mmWave and machine learning technologies. The considered system models, the optimization

**Figure 2.1:** An illustration of a two-user downlink power-domain NOMA scheme with superposition coding and successive interference cancellation decoding [17].
methods utilized to maximize the achievable rates, and the main lessons learnt on
the optimization and the performance of these NOMA-enabled schemes and tech-
nologies are discussed in detail along with the future research directions for these
combined schemes. The rest of this chapter is structured as follows. In Section 2.1,
we briefly describe the enabling technologies and then in Section 2.2, we describe
the rate optimization schemes from the literature along with possible future research
directions.

2.1 Enabling Technologies

In this section, we briefly describe the enabling technologies that are later integrated
with PD-NOMA in rate optimization problems in the literature survey that was
performed.

2.1.1 MISO, MIMO, and mMIMO Communications

The use of multiple antennas at the transmitter, receiver, or both has helped signif-
icantly drive up data rates and been an integral part of cellular systems right from
4G [13]. When multiple antennas are available at the BS, but only a single antenna
at the user, it is referred to as a multiple-input-single-output (MISO) system; while
if multiple antennas are also available at the receiver, it is referred to as a multiple-
input-multiple-output (MIMO) system. These multiple antennas at the transmitter
and receiver are used to realize beamforming or spatial multiplexing gains in single-
user scenarios. In a multi-user environment, the multiple antennas can be used to
separate the users in the space domain, creating the so-called spatial division multiple
access [13]. The use of large-scale antenna arrays at the BS, where the number of
transmit antennas far exceeds the number of users in the system, is referred to as
a massive-MIMO (mMIMO) system. In such mMIMO systems, when the number
of transmit antennas approach infinity, it creates favorable propagation conditions
whereby a unique beam can be formed for each user and perfect separation in the
space domain is possible [21]. When a very large number of antennas are used for
MIMO, it is sometimes referred to as ultra mMIMO.
2.1.2 Advanced Multi-antenna Architectures

The high spectral efficiency demands in B5G networks (100 b/s/Hz or more) require advanced multi-antenna configurations beyond the conventional MIMO and mMIMO in 4G and 5G networks, respectively. Some of these advanced multi-antenna architectural schemes are outlined in the subsequent sub-sections.

Cell-free mMIMO (CF-mMIMO)

With B5G required to support different use-cases, large scale antenna arrays can be used for other engineering challenges such as providing cost and energy-efficient solutions. One such low-cost solution that is an extension of the current mMIMO, is the concept of cell-free mMIMO (CF-mMIMO) [22]. With CF-mMIMO, a subset of UEs are served by a wide geographic distribution of a large number of individually controllable antenna elements. The key idea with CF-mMIMO is to use all the antennas available in the network to serve a subset of UEs, without the typical constraints of having a defined cell coverage area. This allows the served UEs to access a much larger pool of antennas, moving closer to theoretical mMIMO of having infinite antennas at the BS.

3-D MIMO

In 5G NR systems, 3-D separation of users is possible in both the azimuth and vertical direction, sometimes called 3-D MIMO [23]. In [24], many future directions for mMIMO systems are outlined for B5G networks, including extremely large aperture arrays, 3-D MIMO, holographic mMIMO, 6-D positioning, large-scale MIMO radar and smart mMIMO integrated with artificial intelligence.

2.1.3 mmWave and THz Communications

In contrast to the sub-6 GHz frequency range, a large amount of bandwidth is available in the mmWave frequency bands and beyond. Hence, they are seen as a key enabler of high data rates in 5G and beyond cellular networks. However, unlike in the sub-6 GHz spectrum, the propagation of electromagnetic waves in the mmWave bands exhibits high path loss and is highly directional in nature [25]. As a result, the beam gain offered by large-scale antenna arrays, i.e., mMIMO systems, is typically
seen as a requirement to enable successful deployment of networks in these mmWave-frequency bands [26]. Additionally, the short wavelengths of the electromagnetic waves in the mmWave frequency band make it practical to deploy large antenna arrays in a compact area, as required by a mMIMO system [27].

While mmWave communications in the 30 GHz spectrum is already part of the 5G standard, the academic literature is growing in even higher frequency bands such as mmWave for 100-300 GHz and THz bands [28]. The sheer amount of bandwidth available in this region of the frequency spectrum makes it very appealing to meet the large data requirements of 100 Gbps and more for B5G networks, but it comes at the cost of very high propagation loss. The study by Rappaport et al. in [29] highlights several promising applications for communications above 100 GHz in 6G networks, including wireless fiber for back-haul and information showers where a blast of data is sent to the user in a short amount of time as the user passes through an area of coverage at these high frequencies. The same study in [29] also highlights the additional challenges for wave propagation at these frequencies, particularly that the free-space path loss and penetration loss increases significantly as we go to higher frequencies, leading to shorter coverage areas. However, at these frequency bands, the wavelength and consequently the size of the antennas are very small, allowing for the use of highly directional antennas for ultra-precise beamforming, multi-reflector antennas, lens-integrated antenna arrays and other such innovative solutions [28]. Hence, B5G networks will likely employ ultra mMIMO systems to solve the propagation challenges at these bands.

### 2.1.4 Machine Learning

Another developing trend for next-generation wireless networks, such as B5G cellular networks, is the advent of a data-driven approach using machine learning (ML) to complement or even replace the traditional model-driven approach [30, 31] used in communications systems. Relevant to the discussion in this chapter, applying ML algorithms to resource optimization problems has certain advantages over traditional optimization techniques. Primarily, optimization approaches suffer from high cost and complexity when the number of parameters to be configured becomes large. Optimization algorithms are often sensitive to the parameter selection and heuristics have to be re-run from scratch every time there is a small change in the system model, e.g., the arrival of new users. In other words, the entire algorithm has to be
run every time there is a small change in any system parameter. All these limitations of the traditional optimization tools have motivated researchers to explore the use of ML techniques for resource optimization in communications systems [32]. Additionally, when multi-objective optimization problems are framed, the goal is to find Pareto-optimal solutions, i.e., a solution space where the improvement of one metric necessarily degrades some other metric. Due to the large search space involved, ML is an attractive solution for finding such Pareto-optimal solutions [32].

**Figure 2.2:** The different categories of machine learning algorithms with the main tasks/features, some example algorithms, and some example applications in wireless communications [32].

ML algorithms can be broadly classified into three main categories - supervised, unsupervised and reinforcement learning, as illustrated in Fig. 2.2. Supervised learning algorithms use labeled training data for tasks like regression and classification. Unsupervised ML algorithms do not use training data and can be used for tasks like clustering and dimensionality reduction. Reinforcement learning algorithms refer to the set of ML algorithms that have an agent that learns an optimal set of actions by interacting with the environment. Recently, a more powerful form of learning algorithms called deep learning (DL) has emerged. DL algorithms aim to mimic the neurons in the brain, by forming artificial neural networks that comprise multiple hidden layers between the input and output. Unlike classic supervised ML algorithms, DL is more powerful as it can first extract a set of features from the data and then use that for classification or prediction. Such neural networks can also be used as an agent...
in an RL system, forming a Deep Reinforcement Learning algorithm (DRL). DRL algorithms can be used to solve complex non-convex network optimization problems, particularly in modern communications systems where ad-hoc, autonomous decisions need to be taken [37]. As illustrated in Fig. 2.2, all these ML algorithms have been used to solve different problems in communications systems. For example, resource allocation is one such problem often tackled by all classes of these ML algorithms, each tackling the problem from a different angle. A detailed description of this and its potential applications to NOMA communications systems are provided in Section 2.2.4 of the future research directions.

2.2 Survey of rate optimization schemes with PD-NOMA integrated with enabling technologies

2.2.1 Rate-optimal NOMA with MISO, MIMO, and mMIMO Communications

In this section, we first present the representative work for conventional multi-antenna NOMA systems, sometimes referred to as MIMO-NOMA systems in Table 2.2 as it forms the base for rate optimization works for all other multi-antenna NOMA-enabled technologies. The representative work on MISO and MIMO integrated NOMA systems are presented in Table 2.2 for both the single-cell and multi-cell scenarios. Compared to the basic single-input-single-output NOMA (SISO-NOMA) system, the use of multiple antennas at either the transmitter, receiver, or both means that the channel is now described by a vector or a matrix. Hence, unlike in a SISO-NOMA system, the ordering of users according to their channel conditions is no longer trivial. Hence, the MIMO-NOMA system is usually broken down into a SISO-NOMA form that is easier to work with [65]. This is typically achieved by using the multiple antennas at the transmitter for beamforming, such that users can be grouped into beams, often referred to as clusters. Within each cluster, the problem breaks down into a typical SISO-NOMA setting and the traditional PD-NOMA scheme can be used. This approach is classified as NOMA-BF in [66] and cluster-based MIMO-NOMA in [67]. Alternatively, the multiple antennas at the base station can also be used to form one
Table 2.1: MISO-NOMA communications systems: current status of rate optimization schemes.

<table>
<thead>
<tr>
<th>#</th>
<th>Classification</th>
<th>System Model</th>
<th>Design Objective</th>
<th>Optimization Method</th>
<th>Main Finding(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>Two users only</td>
<td>One BS + two users</td>
<td>Max. ergodic sum-rate</td>
<td>Gradient projection algorithm (GPA) and Bisection algorithm (BA)</td>
<td>The proposed schemes outperform OMA and SDMA in the two-user scenario</td>
</tr>
<tr>
<td>39</td>
<td>Multiple users with clustering</td>
<td>One BS + K users (Two users per cluster)</td>
<td>Max. sum rate</td>
<td>Iterative clustering and power allocation algorithm</td>
<td>The proposed scheme outperforms traditional MU-MIMO for the setting where two correlated users per cluster are available</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>One BS + K users (Two users per cluster)</td>
<td>Max. sum rate</td>
<td>Geometric programming (GP), then successive optimization to find Karush-Kuhn-Tucker (KKT) point</td>
<td>Applying ZF-BF in a MIMO-NOMA setup with two users per cluster maximizes the sum rate of the users in the cluster</td>
</tr>
<tr>
<td>41</td>
<td></td>
<td>One BS + K users (Two users per cluster)</td>
<td>Maxmin user-rate</td>
<td>Pareto-boundary computed using reformulation and convex-concave procedure</td>
<td>The proposed scheme allows for rate control between strong and weak users, offering flexibility compared to other MISO-NOMA schemes</td>
</tr>
<tr>
<td>42</td>
<td></td>
<td>One BS + K users (Two users per cluster)</td>
<td>Max. sum rate</td>
<td>Branch-and-bound (BnB) method</td>
<td>NOMA with the proposed BnB technique for beam design outperforms NOMA with ZF</td>
</tr>
<tr>
<td>43</td>
<td></td>
<td>One BS + K users uniformly distributed around the BS</td>
<td>Max. sum rate</td>
<td>One dimensional search</td>
<td>Proposed scheme optimizes the number of feedback hits, but a performance gap always exists compared to the perfect CSI case</td>
</tr>
<tr>
<td>44</td>
<td></td>
<td>One BS + K users (Two users per cluster)</td>
<td>Max. sum rate</td>
<td>Geometric programming (GP)</td>
<td>The proposed scheme optimizes the worst-case achievable sum-rate through robust BF design &amp; outperforms OMA</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>One BS + K users (user ordering assumed)</td>
<td>Max. sum rate</td>
<td>Minorization maximization algorithm (MMA)</td>
<td>The proposed algorithm applies user-specific precoding and the algorithm is shown to converge in a few iterations</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td>One BS + K users (user ordering assumed)</td>
<td>Max. weighted sum rate (WSR)</td>
<td>KKT optimality conditions (hidden convexity for homogeneous channels)</td>
<td>The proposed algorithm designs user-specific beam weights by exploiting favorable complexity in the NOMA WSR problem</td>
</tr>
<tr>
<td>47</td>
<td>Other</td>
<td>Multiple groups each getting multi-cast data</td>
<td>Max. sum rate</td>
<td>Majorization-Minimization algorithm</td>
<td>The proposed scheme always performs better than OMA, but performs better than SDMA only in select scenarios</td>
</tr>
<tr>
<td>48</td>
<td>An UL system with one BS + K users, SIC is applied at the codeword-level</td>
<td>Multiple groups each getting multi-cast data</td>
<td>Max. weighted average SINR</td>
<td>Geometric programming (GP)</td>
<td>The proposed scheme mitigates the error propagation problem specific to UL MIMO-NOMA systems</td>
</tr>
<tr>
<td>49</td>
<td>K users in G groups (Groups based on channel gain)</td>
<td>Multiple groups each getting multi-cast data</td>
<td>Max. sum rate</td>
<td>Singular value decomposition-based multi-user scheme that exploits CSI</td>
<td>The proposed scheme has lower complexity than the MMA scheme in [45] and the duality scheme in [50]</td>
</tr>
</tbody>
</table>
Table 2.2: MIMO-NOMA communications systems: current status of rate optimization schemes.

<table>
<thead>
<tr>
<th>#</th>
<th>Classification</th>
<th>System Model</th>
<th>Design Objective</th>
<th>Optimization Method</th>
<th>Main Finding(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[51]</td>
<td>Two users only</td>
<td>One BS + two users</td>
<td>Max. ergodic capacity</td>
<td>Bisection algorithm (with suboptimal bounds)</td>
<td>The proposed scheme is superior to MIMO-OMA scheme for each user</td>
</tr>
<tr>
<td>[52]</td>
<td>One BS + two users</td>
<td>Layered transmissions to each user</td>
<td>Max. sum rate and Max. average sum rate</td>
<td>Alternating optimization method</td>
<td>Sum rate in MIMO-NOMA system with layered transmission is concave in allocated powers to the multiple layers</td>
</tr>
<tr>
<td>[53]</td>
<td>One BS + two users</td>
<td>Max. average sum rate</td>
<td>Lagrangian dual decomposition</td>
<td>Alternating optimization method</td>
<td>The proposed scheme adapted power and rate allocation to the channel fading state</td>
</tr>
<tr>
<td>[54]</td>
<td>One BS + two users</td>
<td>Max. sum rate</td>
<td>Bisection algorithm</td>
<td>Alternating optimization method</td>
<td>Sum rate in the proposed scheme is better than TDMA MIMO and MU-MIMO for the 2-user scenario studied</td>
</tr>
<tr>
<td>[55]</td>
<td>Multiple users with clustering</td>
<td>One BS + K users</td>
<td>Maxmin user-rate</td>
<td>Bisection and heuristic algorithms</td>
<td>The sum rate - complexity tradeoff was analyzed and the proposed heuristics achieved a good balance</td>
</tr>
<tr>
<td>[56]</td>
<td>Downlink system with unicast and multicast data (Two users per cluster)</td>
<td>Max. weighted sum rate</td>
<td>Iterative weighted minimum mean square error (WMSE) algorithm</td>
<td>The proposed scheme outperforms a ZF-precoder for this system model</td>
<td></td>
</tr>
<tr>
<td>[57]</td>
<td>One BS + K users (num. antennas at UE &gt; num. antennas at BS)</td>
<td>Max. sum rate</td>
<td>Suboptimal heuristics</td>
<td>The proposed scheme performs better than OMA scheme and other MIMO-NOMA schemes for this antenna configuration</td>
<td></td>
</tr>
<tr>
<td>[58]</td>
<td>One BS + K users (short + long-term channel feedback)</td>
<td>Max. weighted sum rate</td>
<td>Iterative WMSE algorithm</td>
<td>The proposed scheme uses long-term channel feedback for US, but the instantaneous CSI for beam design and this outperforms traditional MU-MIMO</td>
<td></td>
</tr>
<tr>
<td>[59]</td>
<td>One BS + K users</td>
<td>Max. both sum rate and the number of admitted users</td>
<td>An iterative algorithm</td>
<td>The authors proved that the more users are admitted to the same cluster, the lower is the achieved sum rate. Hence, a user admission scheme has been proposed that achieves a good balance between the sum rate and number of admitted users</td>
<td></td>
</tr>
<tr>
<td>[60]</td>
<td>One beam per user</td>
<td>One BS + K users (with channel uncertainties)</td>
<td>Max. worst case achievable rate</td>
<td>Cutting-set method WMSE formulation</td>
<td>Proposed robust design is shown to handle channel uncertainties better than a non-robust design</td>
</tr>
<tr>
<td>[61]</td>
<td>Other</td>
<td>One BS + K users (using a specialized superposition scheme)</td>
<td>Max. sum rate</td>
<td>Convex optimization problem solved with interior-point method</td>
<td>The proposed superposition scheme does not affect the decoding of signalling info, allowing for sum-rate gains</td>
</tr>
<tr>
<td>[62]</td>
<td>Multi-cell</td>
<td>Multi-antenna users (Two users per cluster)</td>
<td>Max. sum rate</td>
<td>Convex quadratic programming and semi-definite programming (SDP)</td>
<td>The proposed path-following algorithms increase the overall sum-rate in the system compared to any OMA scheme</td>
</tr>
<tr>
<td>[63]</td>
<td>Multi-cell</td>
<td>Multi-antenna users and a two-cell setup with no CoMP</td>
<td>Max. weighted sum rate of strong user in each cell</td>
<td>Sequential convex approximation (SCA) and MMA techniques</td>
<td>The proposed scheme for the two-cell setup is superior to ZF-NOMA, orthogonal-NOMA and OMA schemes</td>
</tr>
<tr>
<td>[64]</td>
<td>Single-antenna users (No clustering)</td>
<td>Max. sum rate</td>
<td>Iterative scheme based on local optimum till KKT optimality conditions are satisfied</td>
<td>NOMA with the proposed PA scheme achieves higher sum rate than NOMA schemes with basic PA</td>
<td></td>
</tr>
<tr>
<td>[65]</td>
<td>A downlink HetNets MIMO-NOMA scheme</td>
<td>Max. sum rate</td>
<td>A game theory-based approach</td>
<td>The proposed scheme achieves better sum-rate performance compared to conventional MIMO-OMA and MIMO-NOMA-based HetNets schemes</td>
<td></td>
</tr>
</tbody>
</table>
beam per user like in MU-MIMO, and then the beamforming weights are designed to create enough difference between the users’ channel conditions such that the NOMA principles can be applied [45].

With the clustering-based approach, the two users should have sufficient difference in the magnitude of the channel gain coefficients for the PD-NOMA scheme to work well [2]. Thus, user selection (US) is a key parameter in such schemes. The objective here is to pair users who have sufficiently different channel conditions and also fit within a cluster, i.e., a beam. This design variable (or parameter) is often optimized in such schemes along with the power allocation (PA) coefficients assigned to each user in the cluster, e.g., [39,68]. However, with randomly located users, good user pairing algorithms are not guaranteed to find the ideal user clustering even with an exhaustive search of all possible solutions [69] because acceptable solutions may not exist. Hence, user-specific beamforming weights are also considered as a parameter in these sum-rate optimization problems [41,44,56,69,70]. The addition of the optimization of the per-user beamforming (BF) weights to the clustering-based MIMO-NOMA rate optimization problem allows the schemes to either create more channel gain difference among the users in a cluster, or to better separate the different clusters from each other.

With the one-beam-per-user model, the user ordering is predefined and the optimization schemes design a set of precoder weights to meet the given user ordering [45,46]. The sum rate of the users is maximized through a minorization maximization algorithm (MMA) for the non-convex problem proposed by Hanif et al. [45]; while in the technique suggested by Zhu et al. [46], it is shown that forming a weighted sum-rate maximization problem instead has hidden convexity that can be solved using the KKT optimality conditions. While this one-beam-per-user approach generally requires full channel state information at the transmitter (CSIT) for the required beamforming design, the study in [59] framed an optimization problem to maximize the worst-case achievable rate in order to offer robustness to channel uncertainties.

The availability of CSIT also affects the formulation of the MIMO-NOMA rate optimization problems considered in the literature. If the full CSIT is available, the design objective is to maximize one of the overall sum-rate [39,45,56], weighted sum-rate [55,57], or, the worst user-rate [41,54]. On the other hand, if only statistical CSIT is available, then the design objective is to maximize the ergodic sum-rate [38,51]
rather than the actual sum-rate. While obtaining full CSIT helps maximize the system throughput, it comes with additional cost and overhead, especially in frequency division duplex (FDD) systems where reciprocity cannot be used. This motivates the need for system design assuming only limited feedback or partial CSIT [71]. In the work by Yang et al. [43], the number of feedback bits for CSIT is optimized and it is demonstrated that the system throughput (i.e., sum rate) can be increased as the number of feedback bits for CSIT is increased.

Further, the representative works on mMIMO integrated NOMA systems are presented in Table 2.3. While the favorable propagation conditions in mMIMO systems say that an individual beam can be formed per user when the number of antennas at the BS is infinite, users with highly correlated user channels are hard to separate through SDMA with a finite number of antennas. However, as with the cluster-based MIMO-NOMA systems, such users are ideal to be grouped in a cluster if there is sufficient difference in their large-scale fading coefficients [72]. This leads to similar US and PA optimization algorithms as the cluster-based MIMO-NOMA systems. As such, sum-rate optimization problems in such a mMIMO-NOMA setting have been studied in single-cell [73,74] and multi-cell [75] frameworks using this clustering-based approach.

• Lessons Learnt

– The literature of NOMA integrated with conventional MIMO forms the base for integrating NOMA with other advanced technologies, as multiple antennas are typically involved at either end of the communication system. The typical approach in most of the literature is to form effective clusters of users after considering the effective channel gains that account for the beamforming weight multiplication. Typically, once the clusters are formed and user ordering established, optimal power allocation coefficients are derived to maximize the sum rate of the users in the cluster.

– In mMIMO systems that serve multiple users, cluster formation for NOMA and SDMA can be seen as complementary approaches to jointly serve multiple users and improve the overall spectral efficiency. Users with channels that are highly correlated are easy to separate through NOMA but hard to separate through SDMA and vice-versa. This property of SDMA and NOMA to favor opposing characteristics in the channel conditions, make
them well suited to be used in conjunction with each other.

**Future Work**

A lot of work already exists in this area including in mMIMO-NOMA systems. However, there are still a few interesting research directions to pursue in these fields that we outline in this section, followed by a more detailed discussion of future work in NOMA integrated with the other advanced multi-antenna architectures operating in the higher end of the mmWave spectrum as well as THz frequencies in subsequent sections.

For the conventional MIMO-NOMA systems, investigating rate optimization schemes in multi-carrier settings like OFDM requires further consideration as the user ordering will be different across the different blocks of OFDM sub-carriers, depending on the coherence bandwidth of the channel. Considering channel assignment, along with PA and US adds a significant level of complexity to the rate-optimization problem, e.g., [76,77]. Secondly, uplink MIMO-NOMA communications systems offer interesting research directions when the power budget of the UL transmitters is considered in the problem, especially the low-cost ones typical of the massive machine type connectivity scenarios in 5G [78]. As a result, the PA and BF strategy has to consider the end-user capabilities in the problem, particularly in multi-carrier settings like OFDM, where the peak-to-average-power-ratio (PAPR) constraints come into play [79].

**2.2.2 Rate-optimal NOMA with Advanced Multi-antenna Architectures**

NOMA has recently been integrated with some of the other advanced multi-antenna architectures, presented in Section 2.1.2 and the representative work in this area is captured in Table 2.3. Only those technologies from the list in Section 2.1.2 that have representative NOMA-integrated rate optimization works are included here, while the integration of NOMA with other advanced multi-architecture schemes, namely, LIS and 3-D MIMO are discussed in the future work.
Table 2.3: mMIMO-NOMA, advanced multi-antenna architectures (cell-free mMIMO-NOMA, and reconfigurable antenna NOMA-enabled scheme): current status of rate optimization schemes.

<table>
<thead>
<tr>
<th>[#]</th>
<th>Classification</th>
<th>System Model</th>
<th>Design Objective</th>
<th>Optimization Method</th>
<th>Main Finding(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[72]</td>
<td>mMIMO</td>
<td>mMIMO BS, single-antenna users</td>
<td>Max. sum rate</td>
<td>Heuristic PA algorithm</td>
<td>A user pairing algorithm with the proposed hybrid approach outperforms traditional MU-MIMO in scenarios with correlated users</td>
</tr>
<tr>
<td>[73]</td>
<td>mMIMO BS</td>
<td>A hybrid NOMA clustering + traditional ZF-BF MU-MIMO</td>
<td>Max. sum rate</td>
<td>SCA technique</td>
<td>With the proposed PA scheme, NOMA achieves significant spectral efficiency (SE) gain in massive connectivity scenarios</td>
</tr>
<tr>
<td>[74]</td>
<td>mMIMO BS</td>
<td>single-antenna users Clustering-based model</td>
<td>Max. weighted sum rate</td>
<td>Heuristic US algorithm</td>
<td>The proposed US algorithm enhances the sum rate compared to other US algorithms that group pairs of users</td>
</tr>
<tr>
<td>[75]</td>
<td>mMIMO BS</td>
<td>multi-antenna UEs</td>
<td>Max. sum rate</td>
<td>Geometric programming, then successive approximation</td>
<td>The proposed scheme works better than OMA and other MIMO-NOMA schemes in massive connectivity scenarios</td>
</tr>
<tr>
<td>[76]</td>
<td>Cell-free mMIMO</td>
<td>An uplink CP-mMIMO NOMA system, $K$ single-antenna users</td>
<td>Max. sum rate</td>
<td>SCA-based GP algorithm</td>
<td>The rate-performance of the proposed NOMA scheme is superior/inferior to the OMA scheme for medium-high/low user deployment because of the NOMA severe intra-cluster interference</td>
</tr>
<tr>
<td>[77]</td>
<td>Reconfigurable Antenna Systems</td>
<td>A single-cell lens-based reconfigurable antenna mmWave-NOMA scheme, $K$ single-antenna users</td>
<td>Max. sum rate</td>
<td>Solving the dual convex min. total transmission power problem via KKT optimality conditions</td>
<td>The rate-performance of the proposed scheme outperforms both conventional OMA and reconfigurable antenna multiple access schemes</td>
</tr>
<tr>
<td>[78]</td>
<td>A semi-blind interference aligned MISO-NOMA system, $K$ single reconfigurable antenna users</td>
<td>Max. sum rate</td>
<td>An alternative methodology based on some approximations</td>
<td>The sum-rate performance of the proposed scheme outperforms both the MISO-OMA as well as MISO-NOMA based on regularized zero forcing beamforming schemes</td>
<td></td>
</tr>
</tbody>
</table>
Cell-free mMIMO (CF-mMIMO)

The integration of NOMA in CF-mMIMO is an emerging trend in the literature as it captures two promising B5G technologies that offer high spectral efficiency [85]. Instead of considering cells, the system models in such a setting involve distributed access points (AP’s) in a geographical area trying to serve a set of users. The AP’s are connected to a central processing unit that does the bulk of the baseband processing. Like in other MIMO-NOMA settings, the integration of NOMA means that the distributed antennas are used to form clusters of users that can be jointly served in the same orthogonal resource by PD-NOMA. Hence, compared to CF-mMIMO-OMA, the additional design complexity comes in the user pairing problem as the right sets of users have to be identified to be grouped in the cluster. The introduction of NOMA however brings spectral efficiency gains to a CF-mMIMO system, both in the DL [86] and UL [82]. In [82], a sum-rate optimization problem similar to the theme of this chapter for the UL direction is solved using the SCA method.

Reconfigurable Antenna Systems

The integration of NOMA with the concept of reconfigurable antennas has recently been explored in the literature [83, 84]. In [83], reconfigurable antennas and NOMA are used as complementary multiple access schemes and decisions are made based on the grouping of users as to which scheme is selected for which users. In NOMA, particularly in mmWave systems where the LoS path dominates, the best clusters of users are those that have a similar angle of departure (AoD), but different channel gains. The scheme in [83] then serves users who meet such criteria through PD-NOMA, while users who have not only a similar AoD, but also similar channel gains are served with a reconfigurable antenna based multiple access scheme. On the other hand, in [84], reconfigurable antennas are used to mitigate the inter-cluster interference in multi-user MISO-NOMA settings. Through only knowledge of the large scale channel properties, the proposed scheme is shown to effectively cancel out the inter-cluster interference. This allows a large number of users to be served in the same orthogonal resource element, through a combination of PD-NOMA SIC for intra-cluster interference management and reconfigurable antennas for inter-cluster interference.

- Lessons Learnt
While work in this area is new and still limited, the common theme of integrating NOMA with these advanced multi-antenna architecture schemes is that NOMA adds an additional degree of flexibility that enables further spectral efficiency gains compared to when these techniques are used in conjunction with OMA. The main design challenge compared to OMA in each of these techniques is the user clustering problem, as an effective beam to cover a set of users in a NOMA cluster has to be formed with distributed AP’s in CF-mMIMO, reconfigurable antennas or with the extremely thin pencil beams in ultra mMIMO systems operating at higher frequencies.

**Future Work**

Compared to the integration of MIMO with NOMA which is a well-studied area, the combination of NOMA with other advanced multi-antenna architectures is still in its infancy with a large room for contributions. In what follows, we outline some promising research directions for the integration of NOMA with each of these architectures.

The integration of NOMA and CF-mMIMO is in its infancy [82, 86] and much work remains to be done in this area. The user pairing problem brings some unique flavors in such CF-mMIMO-NOMA settings that are different from other NOMA settings. In particular, due to the lack of definition of a cell, the users can be served by a large number of geographically distributed access points (AP’s). To identify good NOMA pairs, i.e., correlated in the angle domain but sufficiently different in channel gains, can be quite challenging in such settings. On the contrary, compared to a regular OMA-based CF-mMIMO setting, NOMA brings tremendous advantages as it avoids the need to form unique beams for correlated users with the distributed AP’s, instead allowing them to be served through PD-NOMA in one cluster. This tradeoff between the advantages of OMA and NOMA in CF-mMIMO, in terms of complexity and spectral efficiency gains, is an area of future research.

When it comes to reconfigurable antenna systems, the advantage is that it allows for changing physical attributes of the antennas to favor the desired transmit beamforming. The integration of NOMA to such a setting means that the reconfigurable properties of the antennas can be exploited for desired cluster formations. As described in Section 2.2.2, two flavors of these techniques were proposed in [83, 84]. However, the question of how best to exploit the reconfigurable properties of the antennas to best suit NOMA cluster formation still has lots of room for contribution in
the literature. For example, when a large number of users are clustered in an area using mmWave systems with largely LoS paths, reconfigurable antennas can be used to form two clusters with different polarizations, to even out the distribution of users between the two clusters.

The large intelligent surfaces (LIS), also described in Section 2.1.2 can be combined with NOMA to further enhance the spectral efficiency. In traditional LIS systems, despite the use of a large surface full of transmitting antennas, beams that are narrow enough to distinguish the different users are challenging to form [87]. This is particularly true for massive connectivity scenarios. A NOMA-LIS system that groups clusters of users through the LIS and then serves the users within a cluster through NOMA is an area worthy of future work. The LIS by nature is an expensive solution, so processors capable of a large amount of signal processing can be expected. As a result, it is a promising solution in the uplink as it might be feasible to combine a large number of users in a cluster. The large number of antennas on the surface can be exploited for advanced inter-cluster interference mitigation purposes.

Finally, a system that integrates 3-D MIMO with NOMA can solve the typical challenges 3-D MIMO systems face on their own. To separate users on different floors of a building through vertical BF with a limited antenna array size, the challenge is to form narrow enough beams that keep the inter-beam interference under control [23]. The introduction of PD-NOMA to such a setting can help alleviate this problem, as users within a beam can be separated using NOMA, thereby allowing for much wider beams to be formed without affecting the overall sum-rate.

2.2.3 Rate-optimal NOMA with mmWave and THz Communications

The works capturing the integration of NOMA with mmWave and THz bandwidth technology is shown in Table 2.4. When operating in these higher frequency bands, the system model needs to consider the low-rank channels and spatial correlation characteristics that are typical for the mmWave communications [25, 90]. While the high spatial correlation of users is typically a constraint for user separation in MU-MIMO systems, MIMO-NOMA systems operating in the mmWave band can use the clustering approach to exploit this correlation to form a cluster and separate the users in the power domain [12, 25]. Another practical limitation of systems operating in the
<table>
<thead>
<tr>
<th>[#]</th>
<th>Classification</th>
<th>System Model</th>
<th>Design Objective</th>
<th>Optimization Method</th>
<th>Main Finding(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[88]</td>
<td>mmWave Communications</td>
<td>A two-tier HetNet cooperative multicast mmWave NOMA scheme, ( K ) multi-antenna users</td>
<td>Max. sum multicast rate</td>
<td>A golden section search algorithm</td>
<td>The proposed cooperative NOMA multicast outperforms NOMA multicast and multicast in terms of the sum multicast rate.</td>
</tr>
<tr>
<td>[89]</td>
<td>A beamspace mMIMO-NOMA scheme in mmWave band, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>An iterative optimization algorithm</td>
<td>The proposed scheme can achieve higher spectrum efficiency than that of beamspace MIMO scheme.</td>
<td></td>
</tr>
<tr>
<td>[90]</td>
<td>A random beamforming mmWave NOMA scheme, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>A suboptimal algorithm based on matching theory</td>
<td>The proposed scheme outperforms the conventional mmWave OMA scheme in terms of the sum rate.</td>
<td></td>
</tr>
<tr>
<td>[91]</td>
<td>An uplink mmWave NOMA scheme, two single-antenna users</td>
<td>Max. sum rate</td>
<td>Decomposition and relaxation</td>
<td>Proposed scheme achieves better sum-rate performance compared to OMA scheme.</td>
<td></td>
</tr>
<tr>
<td>[92]</td>
<td>A downlink mmWave-NOMA scheme, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>A game theory-based algorithm</td>
<td>The proposed scheme outperforms the conventional mmWave-OMA scheme in terms of the sum rate.</td>
<td></td>
</tr>
<tr>
<td>[93]</td>
<td>A hybrid precoding-based MIMO-NOMA scheme with SWIPT, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>Iterative optimization algorithm</td>
<td>The proposed scheme achieves higher sum-rate than a hybrid-based MIMO-OMA scheme with SWIPT.</td>
<td></td>
</tr>
<tr>
<td>[94]</td>
<td>A multi-beam mmWave NOMA scheme, ( K ) multi-antenna users</td>
<td>Max. sum rate</td>
<td>Difference of convex (D.C.) programming transformation</td>
<td>The proposed scheme provide a higher sum-rate performance as compared to the single-beam mmWave-NOMA and the mmWave-OMA schemes</td>
<td></td>
</tr>
<tr>
<td>[95]</td>
<td>A downlink mmWave mMIMO-NOMA scheme with hybrid architecture, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>A matching theory-based algorithm and an iterative optimization algorithm</td>
<td>The proposed scheme provide a higher sum-rate performance as compared to the conventional beamspace MIMO scheme</td>
<td></td>
</tr>
<tr>
<td>[96]</td>
<td>A downlink mmWave-NOMA scheme with analog beamforming, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>A suboptimal solution via decomposing and relaxing the original non-convex problem</td>
<td>The proposed scheme provide a higher minimal-user rate performance as compared to the conventional mmWave-OMA scheme</td>
<td></td>
</tr>
<tr>
<td>[97]</td>
<td>A downlink mmWave-NOMA scheme, ( K ) single-antenna users, Hybrid BF and user clusters formed</td>
<td>Max. sum rate</td>
<td>Suboptimal heuristics</td>
<td>The proposed joint UP, Hybrid BF &amp; PA scheme outperforms conventional mmWave-OMA systems and the mmWave-NOMA scheme in [90].</td>
<td></td>
</tr>
<tr>
<td>[98]</td>
<td>A downlink mmWave-NOMA scheme, ( K ) users, analog phased arrays at BS &amp; users</td>
<td>Max. sum rate</td>
<td>Suboptimal heuristics</td>
<td>The proposed mmWave-NOMA scheme outperforms mmWave-OMA scheme</td>
<td></td>
</tr>
<tr>
<td>[99]</td>
<td>A downlink mmWave-NOMA scheme, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>A K-means-based machine learning algorithm</td>
<td>The proposed scheme outperforms the mmWave-NOMA scheme with random-user clustering and its counterpart mmWave-OMA scheme in terms of the sum rate</td>
<td></td>
</tr>
<tr>
<td>[100]</td>
<td>A downlink mmWave-NOMA scheme, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>An algorithm based on unsupervised machine learning method</td>
<td>The sum-rate performance of the proposed scheme outperforms OMA scheme when the number of users exceeds the number of resource blocks</td>
<td></td>
</tr>
<tr>
<td>[101]</td>
<td>A downlink multi-cell mmWave mMIMO-NOMA enabled system, ( 2K ) multi-antenna users</td>
<td>Max. sum rate</td>
<td>An alternate optimization algorithm based on the constrained convex-concave procedure (CCCP)</td>
<td>The proposed schemes to optimize the precoder and decoder design, as well as a cooperative scheme outperform baseline mmWave-NOMA schemes in multi-cell settings</td>
<td></td>
</tr>
<tr>
<td>[102]</td>
<td>A downlink mmWave-NOMA scheme in dense network, ( K ) single-antenna users</td>
<td>Max. sum rate</td>
<td>Generic algorithm, particle swarm optimization, and simulated annealing</td>
<td>The proposed PSO-based/two-stage PA algorithms perform better/slightly worse than the benchmark scheme in [99] with low computational complexity</td>
<td></td>
</tr>
<tr>
<td>[103]</td>
<td>THz Communications</td>
<td>A downlink THz-NOMA scheme with hybrid BF, ( K ) multiple-antenna users</td>
<td>Max. sum rate</td>
<td>Dual-decomposition method with iterative sub-gradient algorithm</td>
<td>The sum-rate performance of the proposed THz-NOMA scheme is superior to OMA scheme</td>
</tr>
</tbody>
</table>
mmWave bands with mMIMO large scale antenna arrays is that scaling the number of transceivers with the number of antennas is often unfeasible. Hence, unlike with regular MIMO-NOMA systems, the system models studied in the literature often use either analog BF with a single RF chain [27] or a hybrid BF design with a reduced number of RF chains [25,93,94]. With these additional constraints in the system model, sum-rate optimization schemes in the mmWave bands have similar US and PA design objectives as those with other MIMO-NOMA schemes, e.g., [90]. In multi-cell mmWave-NOMA settings, the authors in [98] propose an angle-domain NOMA scheme that schedules one cell-center and one cell-edge user in a NOMA pair, for each beam in each cell. The sum rate is then maximized by optimizing the precoder and decoder BF design along with the selection of paired users, in settings where the cells co-operate as well as when they do not. For the US sub-problem, while it is typically tackled through traditional optimization schemes, in [12] and [97], the US problem is solved using an unsupervised clustering ML approach. Both these works exploit the high correlation amongst users’ channels and the fact that mmWave propagation is dominated by the LoS path to effectively employ K-means clustering. Since the effects of multipath propagation are limited in mmWave spectrum, the user clustering in mmWave-NOMA systems comes down to finding spatially correlated users with the available CSI at the BS. This is precisely what unsupervised clustering algorithms are capable of achieving without any labeled training data.

The appeal of integrating NOMA in even higher parts of the spectrum, e.g., THz communications, that offer orders of magnitude more bandwidth than even the mmWave spectrum is obvious for 5G networks, as it offers opportunities to cluster more users in a very wide band. Through the joint optimization of beam, bandwidth, and power allocation, a THz-NOMA system allows opportunities for meeting both massive connectivity and extremely high data rates in B5G systems [28]. In [100], the authors design a THz-NOMA system that accounts for the distance and frequency selective properties of the THz band. In traditional THz bands, the Long-User-Central-Window (LUCW) principle is usually used to assign the central sub-band to long users, while the side sub-bands are allocated to the short users [28]. Through the integration of NOMA, the authors in [100] enhance this principle to NOMA clusters of four users each, rather than individual users. Through a hybrid BF design in an ultra mMIMO system that forms 4-user clusters, followed by power allocation and sub-band assignment, a rate optimization that outperforms THz-OMA is demonstrated.
in [100].

- Lessons Learnt

  - In mmWave systems, due to the high correlation amongst user channels, particularly when users are geographically clustered, it offers great opportunities for integrating NOMA. Users that are hard to separate with individual beams using MU-MIMO techniques can now easily be grouped together in a NOMA cluster. Through the use of mMIMO, beams are formed to serve clusters of users and separated from other beams (clusters) using BF techniques. Since the mmWave channels are usually dominated by the most dominant path, usually the LoS path, the US problem breaks down to finding users with highly correlated channels in the angle domain, often captured through the cosine similarity metric. Unsupervised machine learning techniques that automatically identify such clusters of correlated users is an emerging trend in the mmWave-NOMA literature.

Future Work

Most of the existing work in mmWave communications is studied assuming the one-path model where the LoS path or non-dominant LoS path dominates. The performance of the proposed user clustering and power allocation schemes in more generic mmWave models that might apply in scenarios where a LoS path is absent is a possible future work. As discussed in the literature survey on mmWave-NOMA, due to the high correlation in users’ channels in the angle domain, ML techniques that identify clusters of highly correlated users is an emerging trend. There is still lots of room for more innovative solutions that exploit these features of mmWave channels for the benefit of NOMA cluster formation, through both ML and traditional optimization.

Further, since B5G systems are likely to use frequency bands above 100 GHz [29], where the main challenge is the high path loss experienced during signal propagation. Through the use of extremely large antenna arrays, beamforming solutions help increase the coverage area. The integration of NOMA serves as a great complement to such systems, as it helps cover a large set of users within such a beam. In this way, a large number of users can be supported at very high data rates, a key requirement on B5G networks. Particularly, since analog or hybrid BF is often employed in very large antenna systems to reduce the cost [27], the BS is only able to produce one
beam at a time with analog BF. The beam and user selection problems in such a setting are promising research topics in the context of rate optimal NOMA-enabled schemes operating beyond 100 GHz with very large antenna arrays.

In fact, the study by Rappaport et al. in [29] highlights some peculiarities in bands above 100 GHz, where some bands are prone to high attenuation, while some other bands suffer surprisingly little loss compared to sub-6 GHz bands. In [29], the authors advocate that such bands above 100 GHz that suffer less loss and so can provide good coverage are candidates for deployment in high-speed 6G networks. The integration of NOMA to such bands is an area of future research work.

2.2.4 Rate Optimization of NOMA-enabled Systems with Machine Learning

For NOMA-enabled next-generation wireless systems, a common theme to all categories of the rate optimization problems surveyed in this chapter is that the number of design variables becomes prohibitively large to configure as the complexity of the system model grows. The multiple design variables are hard to jointly optimize due to the combinatorial complexity. Hence, there is a large potential to apply machine learning (ML) to solve the types of optimization problems for NOMA-enabled systems surveyed in this chapter. In [101], applications of ML and DL are discussed to resource optimization problems in IoT and other cellular networks, including a brief description of the applications to NOMA systems. Similarly, the survey by Vaezi et al. in [102] presents some discussion on NOMA integrated with ML and deep learning (DL).

To address the problem of combinatorial complexity, the most common approach in the NOMA literature is to divide the problem into a set of sub-problems. ML techniques can be used to aid one or more of these sub-problems as appropriate. For example, in [12] and [97], the user selection sub-problem is tackled through an unsupervised clustering algorithm and the power allocation problem is addressed through conventional optimization. In [76], where a multi-carrier setting is studied, the channel assignment is tackled through a DRL algorithm while the power allocation is again addressed through conventional optimization. When further variables are added to the problem, ML techniques can be applied to the power optimization too. For example, the authors in [103] use an RL algorithm for power allocation when an intentional jammer is present. The takeaway message is that as the complexity of
the model grows introducing a prohibitively large number of design variables, as is typical for NOMA-enabled systems, ML techniques can be used to solve a subset of these problems and can be used in tandem with traditional optimization techniques. For example, in [104], an artificial neural network is used in conjunction with a traditional optimization approach to solve a joint problem of power allocation and UAV’s placement to maximize the sum rate of all users. As complexity is introduced to the NOMA enabled systems through multi-cell, multi-carrier, cooperative settings, etc., the number of design variables can even grow too large for ML algorithms. In such scenarios, an interesting research direction can be to investigate a deep learning neural network such as the one proposed in [105] to identify the parameters that have the largest impact on the sum-rate performance. The selected optimization variables can then be set using conventional optimization approaches or through other ML techniques.

While applying ML to NOMA-enabled systems comes with many attractive advantages, it has its challenges also. One of the big concerns with ML algorithms is the computational power required to run some of these ML algorithms. However, emerging trends such as quantum machine learning (QML) for 6G systems are being studied to aid this. The authors in [106] specifically discuss how QML can significantly speed up multi-objective optimization problems that involve tweaking a large number of parameters and their constraints, a typical setting for all NOMA-enabled systems. Another challenge with applying ML algorithms is the large amount of data required. However, communications systems collect and discard a large amount of data today, e.g., CSI, user locations, etc., using them only for instantaneous scheduling decisions. With big data processing developing rapidly, these can be fed to ML algorithms in NOMA-enabled systems.

Another challenge with applying ML at the physical layer is that the channel changes so fast that an ML algorithm does not have enough time to collect meaningful data to learn from. In particular, this makes applying supervised algorithms that learn from past data challenging at the physical layer. However, in mMIMO systems, supervised learning algorithms have been studied for channel feedback and estimation, MIMO detection, and other related problems [24,107]. However, in what follows, we focus on how the other ML techniques, namely unsupervised learning, reinforcement learning, and deep learning, outlined in Section 2.1.4, can be applied to rate optimization problems in the NOMA-enabled system models surveyed in this
chapter. These discussions offer several future research avenues for applying ML in the context of NOMA enabled systems.

Unsupervised ML algorithms do not rely on past training data. In particular, clustering algorithms are a natural fit for NOMA-enabled systems due to user selection sub-problem. As we discussed earlier, due to the combinatorial complexity of the joint optimization of a large number of design variables, the typical approach in NOMA literature is to divide the problem into several sub-problems. The user clustering or user pairing is a typical first sub-problem that researchers tackle. Clustering algorithms such as K-means clustering can be used to tackle this sub-problem. As described in Section 2.2.3, the works in [12] and [97] study a mmWave-NOMA system and exploit the high correlation amongst users’ channels and the fact that mmWave propagation is dominated by the LoS path to effectively employ K-means clustering. System models that are dominated by a LoS path favor the use of K-means clustering as the problem breaks down to finding spatially correlated users. This type of channel model appears in several NOMA-enabled systems. For example, in UAV or satellite communications, the link between the UAV/satellite and ground users is LoS-dominated and offers an opportunity for user clustering based on unsupervised ML algorithms like K-means clustering. For example, in [108], K-means clustering is used to find an initial clustering of spatially correlated users after which a Q-learning algorithm (RL algorithm) is used for the 3-D placement of the UAV BS. It is, however, more challenging to apply such a K-means clustering in a rich multipath environment such as in lower frequency bands because good user pairs are not necessarily correlated users in space in such a setting. An area worth investigating is if techniques that infer the user location from the reported CSI such as the channel charting proposed in [109] can be fed to a K-means clustering algorithm to form good user clusters for NOMA-enabled systems operating in a rich multipath environment.

Deep Learning (DL) is the more powerful form of machine learning as it involves multiple layers and can extract a set of features in the data, before performing tasks like classification [32]. However, due to the fast-changing nature of the physical channel, it becomes difficult to implement a DL approach that first extracts the relevant features from the channel and then applies it to NOMA enabled systems. However, the power of deep learning can still be extracted in several ways in NOMA-enabled systems. In [110], a deep recurrent neural network is constructed to provide optimal resource allocation results for the NOMA heterogeneous IoT with fast convergence
and low computational complexity. We discussed earlier in this section how a neural network such as the one in [105] can be used to extract the most important parameters when the number of design variables and objectives grows very large. However, the most promising use of DL for rate optimization problems in NOMA-enabled systems is when used in conjunction with reinforcement learning, such that the agent employs a multi-layered neural network to make decisions when interacting with the environment. This is termed as deep reinforcement learning (DRL) to highlight the joint use of DL techniques with RL.

DRL algorithms have been studied for several resource allocation problems in next-gen wireless communications systems [37]. Such ideas can easily be extended to the NOMA-enabled systems surveyed in this chapter. We described the two works of [76] and [103] earlier, where DRL agents are used for channel assignment and power allocation respectively in NOMA systems. For the NOMA-enabled B5G technologies surveyed in this chapter, the DRL agent can be either the BS, UAV, users, relay nodes, etc. that need to make autonomous decisions based on their interaction with the other nodes in the system. For example, in [111], a RL agent is employed for UAV positioning. Similar to the idea of ML clustering being used to solve the user selection sub-problem, the DRL can be used to solve certain sub-problems for the overall rate optimization objective. Potential avenues worth investigating include using a RL agent for the sub-problems of relay selection in NOMA relay networks, spectrum selection in CR-NOMA networks, or UAV placement in UAV-NOMA networks. In this way, DRL can be applied to the NOMA-enabled versions of each of these systems to complement NOMA-specific optimization algorithms. Another interesting problem where a RL algorithm can be used is for systems employing a hybrid MU-MIMO and NOMA approach such as [72]. Here, a RL agent can be used to switch between the two spectrum sharing schemes, depending on the favorable conditions it detects based on its past experience of interacting with the system.

Another important class of ML algorithms, called online learning, can help address the issue of system flexibility in NOMA-enabled systems, i.e., to be able to adapt to new users or slight changes in the system without too much overhead. For example, the authors in [12] designed an online ML clustering algorithm that can handle new users entering the system up to a certain threshold. Applying online ML algorithms to these NOMA-enabled rate optimization problems is an important direction of future work, as it helps make the systems practically implementable.
Chapter 3

User Clustering in mmWave-NOMA Systems with User Decoding Capability Constraints for B5G Networks

This chapter proposes a millimeter wave-NOMA (mmWave-NOMA) system that takes into account the end-user signal processing capabilities, an important practical consideration. The implementation of NOMA in the downlink (DL) direction requires successive interference cancellation (SIC) to be performed at the user terminals, which comes at the cost of additional complexity. In NOMA, the weakest user only has to decode its own signal, while the strongest user has to decode the signals of all other users in the SIC procedure. Hence, the additional implementation complexity required of the user to perform SIC for DL NOMA depends on its position in the SIC decoding order. Beyond fifth-generation (B5G) communication systems are expected to support a wide variety of end-user devices, each with their own processing capabilities. We envision a system where users report their SIC decoding capability to the base station (BS), i.e., the number of other users signals a user is capable of decoding in the SIC procedure. We investigate the rate maximization problem in such a system, by breaking it down into a user clustering and ordering problem (UCOP), followed by a power allocation problem. We propose a NOMA-minimum exact cover (NOMA-MEC) heuristic algorithm that converts the UCOP into a cluster minimization problem from a derived set of valid cluster combinations after factoring in the SIC decoding capability. The complexity of NOMA-MEC is analyzed for various algorithm and system parameters. For a homogeneous system of users that all have the same decoding capabilities, we show that this equates to a simple maximum number of users per cluster constraint and propose a lower complexity NOMA-best
beam (NOMA-BB) algorithm. Simulation results demonstrate the performance superiority in terms of sum rate compared to orthogonal multiple access (OMA) and traditional NOMA clustering schemes that do not incorporate individual users’ SIC decoding capability constraints.

3.1 Introduction

Beyond fifth-generation (B5G) communication systems are expected to support a large number of connected users at a time, each with different processing capabilities and requirements. The massive machine-type connectivity (mMTC), also called the Internet of Things (IoT), as well as the ultra-reliable low latency communication (URLLC) use-cases, are expected to bring many different types of connected users into the system compared to traditional mobile broadband users [112]. Thus, B5G systems need to support a very large number of low-cost devices for the IoT connections in addition to the traditional high data rate mobile broadband connections that are also growing exponentially. This puts enormous spectral efficiency requirements on the B5G wireless communication systems.

The mmWave spectrum offers a large amount of bandwidth to scale up the capacity from the cellular networks that operate today in the sub-6 GHz range. Further, non-orthogonal multiple access (NOMA) techniques offer a way to serve multiple users in the same orthogonal resource, e.g., time, frequency, orthogonal frequency division multiplexing (OFDM) resource block (RB), etc., by separating the users in the power domain instead (PD-NOMA). Hence, when combined, mmWave-NOMA has the potential to serve the high rates and massive connectivity demands of B5G networks. Additionally, the high level of correlation amongst users channels in mmWave, makes them ideal for the formation of user clusters to be served by a single beam and separated in the power domain through NOMA [10–12].

The survey in [113] shows that the key aspects of achieving a good performance in NOMA systems are user clustering, user ordering, beamforming, and power allocation. User clustering refers to the selection of users to serve in a NOMA cluster, typically in a beam via beamforming techniques. User ordering refers to the order in which successive interference cancellation (SIC) is applied at the users in the downlink. Power allocation techniques are then used to allocate the right amount of power to each user in the cluster, so that SIC decoding can be successful and each users target
rates are met. The focus of this work is user clustering and user ordering.

As we group users in NOMA clusters, the weakest user only has to decode its own signal, while the strongest user has to decode the signals of all other users in the SIC procedure. The decoding of other users’ signals requires significant additional processing capability, in terms of hardware capability, energy consumption, etc. [13, 14]. The authors in [15] identified this SIC decoding complexity as the first major practical implementation issue for NOMA. NOMA is expected to support a wide variety of end-user devices in B5G systems, each with different signal processing capabilities [15,16]. Hence, each user has its own limitations on the number of other users signals that it can decode. We term this the SIC decoding capability of the user. For NOMA, this SIC decoding capability translates to the number of other users signals a DL user can decode before decoding its own signal. For IoT devices, this could be as low as zero or one, while for high-end smartphones this can be a much higher value due to the differences in hardware processing capability. Hence, when implementing NOMA in the DL, the BS needs to respect this SIC decoding capability limit of the user when it orders users to be served in a NOMA cluster as we discuss further in the motivation in Section 3.1.2.

The SIC decoding capability can be communicated to the BS during connection setup. It can be communicated directly as a value that is applicable as long as the user is connected or it could be a dynamic parameter that is reported periodically as a function of its battery life etc. Alternatively, users could report certain relevant parameters about its hardware capability and current battery condition that the BS can use to determine the SIC decoding capability of the users. For instance, a BS algorithm can take into account the user input along with the rate of the other users, e.g., the QAM order modulation schemes used by the users in a NOMA cluster and determine the appropriate SIC decoding capability of each user for the scheduling instance. For the work in this thesis, the important thing is that the BS has complete knowledge of how many other users signals each user can decode in the SIC decoding procedure for the scheduling instance it is running the algorithm for. It can be seen as a future extension on how to obtain the said SIC decoding capability of each user for a given instant of time.
3.1.1 Related Work

In [114], the authors highlight the tight coupling between user clustering, cluster sizes, and user ordering on the performance of NOMA systems. In typical NOMA works from the literature, the user pairing or user clustering schemes have been designed to group two users per cluster [115,116], or some fixed number of users per cluster [117], respectively. In [117], the optimum cluster sizes from a performance perspective is analyzed. However, we focus in this work on the cluster size as a constraint. More importantly, it is not just a generic constraint that limits the number of users in the cluster, but there is a constraint from each user in the cluster on how many other users signals it can decode in the SIC decoding order. When it comes to the SIC decoding order within a cluster, as highlighted in [118], users are typically ordered either based on their effective channel gains, i.e., channel gains after considering the beamforming weights, or based on their quality of service (QoS) using a cognitive radio concept. In this chapter, we focus on the effective channel gain strategy as we assume all users have the same QoS.

Unlike multi-user MIMO (MU-MIMO), where correlated users are difficult to separate by individual beams, such correlated users can easily be grouped together in a NOMA cluster [113]. In mmWave systems, the users’ channels are highly correlated due to the highly directional nature of mmWave transmission [119,120]. The user clustering schemes in mmWave-NOMA systems typically exploit the high correlation amongst users channels to cluster correlated users together, e.g., [10,121]. In [122], the authors use an angle-domain NOMA scheme that schedules one cell-center and one cell-edge user in a NOMA pair, for each beam in each cell. Recent works in mmWave-NOMA systems have also used machine learning clustering techniques to identify correlated users and group them in NOMA clusters [12,123,124]. Further, in mmWave systems, since it is often infeasible to scale up the number of transceivers with the number of antennas, studies in mmWave-NOMA systems often use either analog beamforming (BF) with a single RF chain [27,119,125] or a hybrid BF design with a reduced number of RF chains [25,93].

3.1.2 Motivation and Contributions

In practical deployments, the typical clustering approaches described above in the related work in mmWave-NOMA have two important limitations. First, they can
lead to arbitrarily large and uneven cluster sizes. If we have a system model where one cluster is served on one channel, this could lead to over-use on one channel and under-use on other channels. More than the imbalance in resource usage, large cluster sizes mean the users at the end of the SIC decoding orders need to decode a very large number of users. This is particularly an issue in dense deployments, where large clusters of correlated users can exist. The second important limitation with these algorithms is there is no flexibility incorporated to account for the SIC decoding capability limitations of each individual user. Concretely, just finding groups of correlated users, can lead to cluster formations where the individual user decoding capability limits of some users are not respected, i.e., users are placed in SIC decoding positions in a cluster that require them to decode the signals of a greater number of users than their indicated SIC decoding capability.

The clustering schemes from the mmWave-NOMA literature that focus on finding correlated users, e.g., [12], can be modified to arbitrarily divide the groups of correlated users the algorithm identified into different clusters, served on different channels. Users then need to be decoded in the order of their decoding capabilities, rather than the effective channel gains. Even when users are decoded in the order of SIC decoding capability, further orthogonal channels might be needed if some users’ constraints are not met. All such workarounds to meet the practical SIC decoding capability constraints of users in real deployments would erode the gains from the clustering algorithms that strived to find good sets of correlated users, with sufficient separation between the clusters. Instead, if the clustering algorithm was able to consider the user decoding capability requirements as part of its input, it would be better able to construct clusters that maximize the overall spectral efficiency, while taking into account these individual user decoding capability constraints. This is the motivation for this chapter of the thesis.

Against this background, in this chapter, we investigate a rate maximization problem for a mmWave-NOMA system that takes into consideration the SIC decoding capability of each individual user in the system. We break down the problem into a user clustering and ordering problem (UCOP), which is the focus of this chapter, followed by a power allocation (PA) problem. We consider a single-cell mmWave-NOMA equipped base station (BS) that applies analog beamforming (ABF) in a fixed set of directions uniformly distributed around the cell coverage area. A NOMA cluster of users will be served on an orthogonal channel, e.g., a time channel or OFDM resource
block, using one of these pre-defined beams. In this way, the UCOP can be framed as a cluster minimization problem in order to minimize the number of orthogonal channels used to serve the required number of users, while respecting each individual user’s SIC decoding capability constraints. We propose two algorithms to solve this UCOP. The first one we term NOMA-minimum exact cover (NOMA-MEC), as we decompose the problem into a MEC problem, a known NP-complete problem [126]. For a homogeneous system where all users have the same SIC decoding capability, we propose a less complex NOMA-best beam (NOMA-BB) algorithm. The key aspects of the two algorithms are outlined next.

In both algorithms, the BS uses the cosine similarity metric that aligns a user’s channel with the set of possible beam directions to rank the best beams for each user. The BS then chooses the best beams to form a candidate beam list for each user, with the number of beams in this list a configurable parameter that can be tuned for a complexity-performance trade-off as we discuss in-depth in this chapter. This step identifies users that can potentially cluster with each other. User ordering in any cluster is done in the order of the effective channel gains. NOMA-MEC then takes the SIC decoding capability of the users into account, and builds a list of valid cluster combinations such that the SIC decoding is done in the order of the users’ channel gains and each user SIC decoding capability is respected. Using this candidate list, NOMA-MEC is able to frame the problem as a MEC problem where the goal is to serve all the users in the least number of channels from the designed set of valid cluster combinations.

In a homogeneous system, the user decoding capability constraints of the users translate to limiting the number of users per cluster, as any user ordering within that cluster will satisfy each user’s decoding constraints since they are all equal. Such a homogeneous system with a restriction on the maximum number of users per cluster is what is typically considered in user clustering algorithms in the literature, e.g., [117]. In our case, the homogeneous system is just a special case of the heterogeneous system where all users have the same SIC decoding capabilities, and so the NOMA-MEC algorithm can still be run. However, for this simpler homogeneous system, we also propose a simpler NOMA-best beam (NOMA-BB) algorithm that demonstrates comparable performance to NOMA-MEC when we have the special setting of all users having the same decoding capability. Finally, we demonstrate the performance superiority of NOMA-MEC compared to orthogonal multiple access (OMA) as well
as the additional flexibility the NOMA-MEC scheme offers in heterogeneous systems compared to other NOMA clustering schemes like NOMA-BB that target a fixed number of users per cluster.

The contributions of this chapter can thus be summarized as follows:

- We design a joint user-clustering, user ordering, and beamforming scheme in a mmWave-NOMA ABF system with a fixed set of candidate beams that minimizes the number of clusters required to serve all the users, subject to each individual users decoding capability and beamforming constraints. Each NOMA cluster is served on one orthogonal channel, so minimizing the number of clusters also minimizes the number of channel uses. Together with a power allocation scheme per cluster, we maximize the sum rate of the system.

- This is the first NOMA work that considers the individual SIC decoding capability of each user when doing NOMA clustering and ordering. The proposed scheme is ideally suited for a mmWave-NOMA deployment involving a low-cost small-cell BS with only one RF chain, supporting a large number and variety of connected users, from low-cost IoT devices with limited processing capabilities to high-end smartphones with much larger processing capabilities. From the perspective of NOMA in the downlink, the processing capability of the user primarily impacts the SIC decoding capability, i.e., the number of other users signals a user can decode every channel use.

### 3.2 System Model and Problem Formulation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h_u, \alpha_u, L, r_u^\parallel )</td>
<td>mmWave channel parameters</td>
</tr>
<tr>
<td>( a(\theta_u), \phi_u, D, \lambda )</td>
<td>Antenna array related parameters</td>
</tr>
<tr>
<td>( M )</td>
<td>Number of antennas at BS</td>
</tr>
<tr>
<td>( N )</td>
<td>Total number of users in the system</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of clusters</td>
</tr>
<tr>
<td>$C = {C_1, ..., C_K}$</td>
<td>The set of $K$ clusters</td>
</tr>
<tr>
<td>$(B, \bar{\theta})$</td>
<td>Coverage area $\bar{\theta}$ is divided into $B$ parts, which the BS uses to generate $B + 1$ candidate beams</td>
</tr>
<tr>
<td>$B_c = {\text{Beam-0, ..., Beam-}B}$</td>
<td>The set of $B + 1$ candidate beams</td>
</tr>
<tr>
<td>$W_c = {w_0, ..., w_B}$</td>
<td>The set of $B + 1$ candidate precoding vectors</td>
</tr>
<tr>
<td>$B_u, b$</td>
<td>User-beam set for user-$u$ containing $b$ beams</td>
</tr>
<tr>
<td>$B_u^*$</td>
<td>Best beam in $B_u$, i.e., $B_u$ with $b = 1$</td>
</tr>
<tr>
<td>$(C_k, N_{C_k}, b_k, w_{b_k})$</td>
<td>The cluster $C_k$, containing $N_{C_k}$ users, served on beam-$b_k$ with a precoding vector $w_{b_k}$</td>
</tr>
<tr>
<td>$s_k$</td>
<td>The transmitted signal to cluster-$k$</td>
</tr>
<tr>
<td>$y_u$</td>
<td>The received signal at user-$u$</td>
</tr>
<tr>
<td>$p_u$</td>
<td>The power allocated to user-$u$</td>
</tr>
<tr>
<td>$P$</td>
<td>The power available per channel</td>
</tr>
<tr>
<td>$\xi_u$</td>
<td>The noise at user-$u$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>The noise power</td>
</tr>
<tr>
<td>$\pi_k(j)$</td>
<td>The user index for the $j$-th decoded user in the $k$-th cluster</td>
</tr>
<tr>
<td>$\Gamma_{\pi_k(j)}^{\pi_k(j')}</td>
<td>The SINR when decoding user $\pi_k(j)$ at user $\pi_k(j')$, where $j' &gt; j$</td>
</tr>
<tr>
<td>$\Gamma_{\pi_k(u)}^{\pi_k(u)}$</td>
<td>The SINR when decoding user $\pi_k(u)$ own signal</td>
</tr>
<tr>
<td>$R_{\text{sum}} / R_{\text{OMA}}$</td>
<td>The effective sum rate of the system that adopted NOMA/OMA scheme</td>
</tr>
<tr>
<td>$R_k$</td>
<td>The sum rate of the users within cluster-$k$</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>$\Gamma_{\text{min}}$</td>
<td>Users minimum QoS SINR</td>
</tr>
<tr>
<td>$d_{\text{max}}$</td>
<td>The maximum decoding capability among the $N$ users in the system</td>
</tr>
<tr>
<td>$d_u$</td>
<td>Decoding capability of user-$u$, in the range $[0, d_{\text{max}}]$</td>
</tr>
<tr>
<td>$m$</td>
<td>A NOMA-BB algorithm parameter between 1 and $d_{\text{max}}$</td>
</tr>
<tr>
<td>$d_{\pi_k(j)}$</td>
<td>The decoding capability of the $j$-th decoded user in the $k$-th cluster</td>
</tr>
<tr>
<td>$C_v$</td>
<td>A list of viable candidate cluster options</td>
</tr>
<tr>
<td>$n_b$</td>
<td>The number of users that have beam-$b$</td>
</tr>
<tr>
<td></td>
<td>in their user-beam set</td>
</tr>
<tr>
<td>$b_{\text{th}}$</td>
<td>User is in the coverage area of $b_{\text{th}}$ beams from the list of candidate beams in $B_c$</td>
</tr>
<tr>
<td>$I_1$</td>
<td>Number of iterations examined in step-1 of NOMA-MEC</td>
</tr>
<tr>
<td>$I_2$</td>
<td>Number of possible combinations for step-2 of NOMA-MEC to examine, i.e., number of elements in $C_v$</td>
</tr>
<tr>
<td>$u_{i,k}$, $z_{u,c}$, $x_c$</td>
<td>Binary variables defined as part of NOMA-MEC algorithm</td>
</tr>
<tr>
<td>$l$, $c_s$, $C_t$, $C_{\text{sorted}}$</td>
<td>Variables defined as part of NOMA-MEC algorithm</td>
</tr>
<tr>
<td>$(.,)^T$, $(.,)^H$</td>
<td>Transpose and Hermitian transpose</td>
</tr>
<tr>
<td>$\lceil . \rceil$</td>
<td>Ceiling function</td>
</tr>
</tbody>
</table>

Consider a mmWave-NOMA single-cell BS equipped with $M$ antennas serving $N$ single-antenna users, each with a minimum QoS constraint. We use the single path mmWave channel model used in several mmWave-NOMA works [10, 12, 120] to model the mmWave channel between the BS and user-$u$ as follows:
Figure 3.1: The ABF scheme used by the BS is illustrated on the left with a fixed set of $B+1$ precoding weights, creating $B+1$ candidate beam directions to choose from when serving a NOMA cluster of users in one orthogonal channel. On the right, the $N$ users, each with their own SIC decoding capability, is illustrated.

\[ h_u = a(\theta_u) \frac{\alpha_u}{\sqrt{L(1 + r_u^\eta)}} \]  \hspace{1cm} (3.1)

where $L$ denotes the number of paths, $r_u$ denotes the distance between the BS and user-$u$, $\eta$ denotes the path loss exponent and $\alpha_u$ denotes the complex channel gain for user-$u$. The parameter $\theta_u$ represents the physical angle of departure and for a uniform linear array (ULA), the normalized angle is defined as $\phi_u = \frac{2D}{\lambda} \sin(\theta_u)$, where $D$ is the separation between elements of the antenna and $\lambda$ is the wavelength of the carrier.
signal [120]. The term \( a(\theta_u) \) represents the steering vector and for a ULA can be represented as

\[
a(\theta_u) = [1, e^{-j2\pi \frac{D}{\lambda} \sin(\theta_u)}, \ldots, e^{-j2\pi (M-1) \frac{D}{\lambda} \sin(\theta_u)}]_T
\]

\[
= [1, e^{-j\pi \phi_u}, \ldots, e^{-j\pi (M-1) \phi_u}]_T.
\] (3.2)

Analog beamforming is used since only one radio frequency (RF) chain is available at the BS, typical of small-cell deployments where low hardware cost and power consumption is essential, e.g., [27,120]. Hence, only one beam can be transmitted at a time, which we equate to forming one beam to serve one cluster of NOMA users per channel use. Since we use ABF that can only generate one beam at a time, we use a time-division strategy to alternate between the different clusters.

As the left part of Fig. 3.1 illustrates, the entire coverage region, \( \bar{\theta} \), from \(-\pi/2\) to \(\pi/2\) is covered by a set of \( B + 1 \) candidate beams, with significant overlap between the candidate beams. A NOMA cluster of users will be served on an orthogonal channel using one of these candidate beams. Each beam-\( b \) in this candidate list has the following precoding vector,

\[
w_b = a(\bar{\theta}_b), \forall b \in [0, B]
\] (3.3)

where the parameter \( \bar{\theta}_b \) is

\[
\bar{\theta}_b = -\pi/2 + (b \times \pi/B).
\] (3.4)

In this way, we uniformly divide this entire coverage region, \( \bar{\theta} \), into \( B \) equal angles, effectively forming a set of \( B + 1 \) candidate beams, as illustrated in the left part of Fig. 3.1. The \( B + 1 \) beams can be thought of as a choice of \( B + 1 \) different precoding vectors based on (3.3), such that collectively, the steering vectors of the \( B \) candidate precoding vectors uniformly cover the entire region of \( \bar{\theta} = -\pi/2 \) to \( \pi/2 \) or \( \bar{\phi} = -1 \) to \( 0 \). We let \( B_c \) represent this list of candidate beams, such that \( B_c = \{\text{Beam-0}, \ldots, \text{Beam-B}\} \), with their respective list of candidate precoding vectors being \( W_c = \{w_0, \ldots, w_B\} \), as illustrated in Fig. 3.1.

A NOMA cluster of users will be served on an orthogonal channel using one of the \( B + 1 \) precoding vectors in the candidate list. In our system model, the orthogonal
channel is a time slice, hence each cluster will be served in one time slice. A total of $K$ such clusters are formed to serve the $N$ users. This equates to requiring $K$ channel uses, i.e., $K$ time-slices, to serve the $N$ users, $K \leq N$. Let $C = \{C_1, \ldots, C_K\}$ represent the $K$ clusters required to serve the $N$ users, where $C_k$ refers to the $N_{C_k}$ users, $N_{C_k} \leq N$, selected to serve in the cluster with index-$k$. For each cluster $C_k$, a beam-$b_k$ with a precoding vector $w_{b_k}$ is selected from the set of $B + 1$ possible beam options in $B_c$.

We exploit the high correlation in the mmWave channels as follows. We assume the BS has access to the full channel state information (CSI) vector of each user, $h_u$, from (3.1) [11, 12]. Additionally, the BS knows the precoding vectors of each beam, $w_b$, in the candidate set. The BS can use the cosine similarity metric between the user’s channel vector and the precoding vector of each beam to determine the level of correlation between the user and the beam. This metric has been used in several mmWave-NOMA works for user clustering to determine the correlation between users in [12, 123], and between users and random beams in [120]. Using similar steps as these works, we can derive the cosine similarity between a user-$u$ with channel $h_u$ and a beam-$b$ with precoding vector $w_b$ here as follows:

$$\cos(h_u, w_b) = \left| \frac{a(\phi_u)^H a(\phi_b)}{M} \right|$$
$$= \left| \frac{\sum_{i=0}^{M-1} e^{-j\pi i (\phi_u - \phi_b)}}{M} \right|$$
$$= \left| \frac{\sin\left(\frac{\pi M (\phi_u - \phi_b)}{2}\right)}{M \sin\left(\frac{\pi (\phi_u - \phi_b)}{2}\right)} \right|$$
$$= F_M\left(\pi [\phi_u - \phi_b]\right),$$

(3.5)

where $\phi_u$ and $\phi_b$ are the normalized directions of the user and beam respectively, $H$ represents the Hermitian transpose and $F_M$ represents the Fejer Kernel, whose properties dictate that as $|\phi_u - \phi_b|$ increases, $\cos(h_u, w_b) \to 0$. In other words, if the beam and users directions are well aligned, the cosine similarity metric is high and it reflects that it is suitable to schedule the user on a cluster served by the beam-$b$. In this way, the BS builds a user-beam set, $B_u$, for each user-$u$. This user-beam set, $B_u$ consists of $b$ beams each, by selecting the best $b$ beams for each user using the cosine similarity metric from (3.5). The parameter $b$ is a tunable parameter, as we discuss later in Section 3.3. We note that based on the choice of $M$ and $B$, the beams are
highly overlapping in nature and so a single user can be served by more than one beam, while still benefiting from a good beamforming gain. Alternatively, the user can select its best beams using the typical ABF approach for the mmWave in new radio (NR) standard [127], but that is beyond the scope of this work and is a topic for future work. It is also worth mentioning that our BF scheme is different from the random BF scheme in [120], where a random beam is generated with precoding vector $\mathbf{w} = \mathbf{a}(\theta)$, $\theta \in [-\pi/2, \pi/2]$ and all users with a high cosine similarity with that beam are then scheduled. In our scheme, while also ABF with a similar precoding vector, we are not randomly generating the beams, but instead selectively choosing an appropriate beam for a NOMA cluster from a given set of candidate options, $B_c$.

For cluster $C_k$, the BS applies superposition coding (SC) for the selected $N_{C_k}$ users as follows:

$$s_k = \sum_{u=1}^{N_{C_k}} \sqrt{p_u} s_{k,u}, \quad (3.6)$$

where $p_u$ represents the power allocated to user-$u$ with $\sum_{u=1}^{N_{C_k}} p_u \leq P$, where $P$ denotes the power available to the BS per-channel use. The received signal at user-$u$ in cluster $C_k$ is

$$y_u = \mathbf{h}_u^H \mathbf{w}_{b_k} s_k + \xi_u,$$

$$= \mathbf{h}_u^H \mathbf{w}_{b_k} \sqrt{p_u} s_{k,u} + \mathbf{h}_u^H \mathbf{w}_{b_k} \sum_{u \neq u, v=1}^{n_k} \sqrt{p_v} s_{k,v} + \xi_u. \quad (3.7)$$

In the SIC procedure, let $\pi_k(j)$ denote the user index for the $j$-th decoded user in the cluster $C_k$ serving $N_{C_k}$ users, $j \leq N_{C_k}$. This $j$-th user then needs to decode and subtract all the messages for all users $\{\pi_k(1), \ldots, \pi_k(j)\}$. The signal-to-interference-plus-noise ratio (SINR) when decoding user $\pi_k(j)$ at user $\pi_k(j')$, $j' > j$ can be represented as

$$\Gamma_{\pi_k(j') \pi_k(j)} = \frac{p_j |\mathbf{h}_{(j')}^H \mathbf{w}_{b_k}|^2}{|\mathbf{h}_{(j')}^H \mathbf{w}_{b_k}|^2 \sum_{v>j}^{N_{C_k}} p_v + \sigma^2}, \quad (3.8)$$

where $\sigma^2$ represents the noise power. Let $R_k$ denote the rate achieved in NOMA cluster $C_k$. The effective sum rate of the system, $R_{\text{sum}}$ can then be expressed as the
The sum of the rates, $R_k$, achieved in each of the $K$ clusters over which all $N$ users are served divided by the number of clusters, since each cluster is served by one channel. The effective sum rate can thus be represented as

$$R_{\text{sum}} = \frac{\sum_{k=1}^{K} R_k}{K} = \frac{\sum_{k=1}^{K} \sum_{u \in C_k} \log_2 \left(1 + \frac{\Gamma_{\pi_k(u)}}{\pi_k(u)}\right)}{K},$$

(3.9)

expressed in bits per second (bps) per channel-use and the term $\Gamma_{\pi_k(u)}$ refers to the SINR when decoding the $u$-th user’s own signal in the SIC decoding procedure and can be expressed as

$$\Gamma_{\pi_k(u)} = \frac{p_u |h_{(u)}^H w_{b_k}|^2}{|h_{(u)}^H w_{b_k}|^2 \sum_{v>j} p_v + \sigma^2}.$$  

(3.10)

For OMA, where each user has to be served in an individual channel, $K$ channels are required to serve the $N$ users, i.e., $K = N$. Each user will be served in its best beam from $B_u$ with a precoding vector $w_u$. This gives us an effective sum rate of

$$R_{\text{OMA}} = \frac{\sum_{u=1}^{N} \log_2 \left(1 + \frac{P|h_{u}^H w_u|^2}{\sigma^2}\right)}{N},$$

(3.11)

where $P$ is the power available per channel. For NOMA, since one cluster is formed per channel, $P$ represents the power available per cluster as we describe later in this section.

To model the user decoding capability constraints, we consider that each user-$u$ is associated with a decoding capability constraint $d$, represented as $d_u$. To illustrate the decoding capability constraint, the right side of Fig. 3.1 shows the distribution of $N$ users to be served by the BS. Using the cosine similarity metric, each user will find the $b$ beams it is best aligned with, forming the user-beam set, $B_u$, for the user. As Fig. 3.1 then highlights, each user has its SIC decoding capability associated with it. For example, user-1 with SIC capability of 0 ($d_1 = 0$) indicates it needs to be either served as an OMA user in an orthogonal channel of its own or in a NOMA cluster as the weakest user where it is not required to decode any other users’ signals. User-4 with SIC decoding capability of 4 ($d_4 = 4$) indicates it is capable of decoding four other users’ signals. This means that if user-4 is scheduled in cluster-$k$ at position $j$, then the maximum value of $j$ is 5 for this user since that would involve decoding 4
other users’ signals, i.e., $\max(j) = 5$.

Let $d_{\text{max}} = \max(d_u), \forall u = [1, ..., N]$, represent the maximum decoding capability among the $N$ users in the system. If all users have the same decoding capability, i.e., $d_u = d_{\text{max}}, \forall u = [1, ..., N]$, we refer to this as homogeneous user decoding capabilities, or just a homogeneous system for short. In a homogeneous system, since any user-$u$ has the same decoding capability $d_u = d_{\text{max}}$, this is equivalent to designing a user clustering scheme such that there are a maximum of $d_{\text{max}}$ users per cluster. On the other hand, in a heterogeneous system, user clustering must be done in tandem with user ordering, such that each user in the cluster needs to decode at most $d$ other user’s signals, where each user has its own value of $d$, $1 \leq d \leq d_{\text{max}}$. This means that for every user-$u$ with decoding capability $d_u$ at SIC decoding position $j$ in cluster $C_k$, i.e., $\pi_k(j)$, it must satisfy that $d \geq j - 1$. Using our nomenclature, $d_{\pi_k(j)}$ denotes the decoding capability of the $j$-th user in the $k$-th cluster.

In this chapter, the objective is to utilize NOMA to maximize the effective sum-rate of the system, such that each user’s QoS is met and all user decoding capability constraints are satisfied. Let $\Gamma_{\text{min}}$ denote the minimum SINR with which each user needs to be served, i.e., $\Gamma_{\pi_k(u)} \geq \Gamma_{\text{min}}, \forall u = [1, .., N]$. The overall objective function to maximize $R_{\text{sum}}$ can be stated as

$$
\max_{\{C_k\},\{w_k\},\{\pi_k\},\{p_u\}} R_{\text{sum}},
$$

subject to

$$
R_u \geq \log_2(1 + \Gamma_{\text{min}}), \forall u = 1, .., N
$$

$$
d_{\pi_k(j)} \geq j - 1, j = 1, .., N_{C_k}, \forall k = 1, .., K
$$

$$
\sum_{i=1}^{N_{C_k}} p_i \leq P, \forall k = 1, .., K
$$

where (3.12b) represents the QoS constraint, (3.12c) represents the decoding capability constraint, and (3.12d) represents the power per channel constraint. It is worth noting that the term $d_{\pi_k(j)}$ and its associated constraint in (3.12c) controls the ordering of users in a cluster, i.e., affects the optimization variable, $\pi_k$.

In order to solve the optimization problem in (3.12a), we break down the problem into two steps. First, we jointly tackle the user clustering, user ordering, and beamforming aspects, where we aim to minimize the number of clusters required to serve all the users while satisfying the beamforming and user decoding constraints.
Second, once we have clusters of users, we do a power allocation step for the users in each cluster. We describe each of these steps next.

In the first step, the goal is two-fold: a) to build clusters of SIC ordered users that satisfy the SIC decoding constraints and b) to identify which beam each of these clusters will be served by, such that the selected beam is in the user-beam set of each of the users selected to be in the cluster. The objective in this step is to serve all the users in the minimum number of clusters, while respecting the aforementioned constraints. Since each cluster is served on one orthogonal channel, the \( N \) users being served on \( K \) clusters, is equivalent to requiring \( K \) orthogonal channel uses to serve the \( N \) users. Hence, reducing \( K \) improves the channel re-use, and in doing so, in general, contributes to an increased spectral efficiency as illustrated by the presence of \( K \) in the denominator of (3.9). However, \( R_{\text{sum}} \) also depends on the SINR for each user in (3.9). This SINR for each user is affected by the other users they are clustered with, the order in which the users are decoded and finally the beamforming gain from the choice of beam from \( B_c \) to serve each cluster with. Along with this, each users SIC decoding capability constraints need to be respected. Hence, we tackle these aspects jointly as a cluster minimization problem subject to several constraints as discussed in what follows next.

For a single-cell NOMA deployment with no inter-cluster interference to consider since each cluster is served in an orthogonal channel, it is known that NOMA performance is significantly improved by decoding users in the order of their channel gains [114, 128]. Hence, given that we have the full CSI of each user along with the precoding vectors, for every cluster \( C_k \), we only allow the users to be decoded in the order of their effective channel gains. This means that the SIC decoding position-\( j \) of user-\( u \) with decoding capability \( d_u \) in cluster-\( k \), \( \pi_k(j) \), is determined by the effective channel gain of the user-\( u \) in relation to other users also selected to be in cluster \( C_k \). Since we have the constraint that \( d_{\pi_k(j)} \geq j - 1 \), we need to design clusters such that the users when ordered according to their effective channel gains, satisfy their SIC decoding constraints. Formally, the user clustering, user ordering, and BF optimization problem can be written as follows. Let \( C = \{C_1, ..., C_K\} \) represent the \( K \) clusters required to serve the \( N \) users. At most, each user is served in its own cluster or channel (equivalent to OMA), hence \( K \leq N \). Each cluster \( C_k \) in \( C \) represents a set of users ordered according to their effective channel gains when served by beam-\( b_k \) from \( B_c \) with precoding vector \( w_{b_k} \). Let \( u_{i,k} \) be a binary variable that represents
whether user-$i$ belongs to cluster-$C_k$, served by beam-$b_k$. Let $d_{\pi_k(j)}$ represents the user decoding capability of the $j^{th}$ decoded user in cluster-$C_k$. The objective of our user clustering, ordering, and BF scheme is to minimize $K$, as follows:

$$\min_{\{u_{i,k}\},\{b_k\},\{\pi_k\}} K,$$  \hspace{1cm} (3.13a)

subject to

$$\sum_{k=1}^{K} u_{i,k} = 1, \forall i = 1, \ldots, N,$$  \hspace{1cm} (3.13b)

$$b_k \in B_u, \ u_{i,k} = 1, \forall k = 1, \ldots, K,$$  \hspace{1cm} (3.13c)

$$d_{\pi_k(j)} \geq j - 1, j = 1, \ldots, N_{C_k}, \forall k = 1, \ldots, K,$$  \hspace{1cm} (3.13d)

$$u_{i,k} \in \{0, 1\},$$  \hspace{1cm} (3.13e)

$$b_k \in B_c, \forall k = 1, \ldots, K,$$  \hspace{1cm} (3.13f)

where constraint (3.13b) ensures each user is placed in exactly one cluster. Constraint (3.13c) is to ensure that the beam-$b_k$ chosen for cluster-$C_k$ in $C$ belongs to the user-beam list of each of the users selected to be served in that NOMA cluster. Constraint (3.13d) ensures the decoding capability constraints of each user in the system is adhered to. For the homogeneous system, since all users have the same decoding capability, we only need to limit the number of users per cluster. In other words, for a homogeneous system, within a cluster-$k$ of size $N_{C_k}$, if $N_{C_k} \leq d_{\text{max}}$, then any decoding order within that cluster is feasible since all users have the same decoding capability of $d_u = d_{\text{max}}$. Hence, for a homogeneous system, constraint (3.13d) can be simplified down to:

$$\sum_{i=1}^{N} u_{i,k} \leq d_{\text{max}}, \forall k = 1, \ldots, K.$$  \hspace{1cm} (3.14)

The second step is power allocation (PA), which is not the focus of this chapter but briefly described here for completeness. Since only one cluster is served on one channel in our model, the channel power budget, $P$, is equivalent to the cluster power budget. Hence, the goal is to divide the power $P$, among the $N_{C_k}$ users in each cluster $C_k \in C$. The objective in this step is to maximize the rate $R_k$ in each cluster. Since the users in the cluster are already ordered based on their effective channel gains, we iterate through the first $j = \{1, \ldots, N_{C_k} - 1\}$ users in the cluster at position $\pi_k(j)$ and
assign it as much power as it needs to satisfy $\Gamma_{\min}$ and ensure successful SIC decoding, based on (3.8). The strongest user is assigned the remaining power. We assume $P$ is always sufficient to meet each user’s QoS, including the remaining power left over for the strongest user. This is similar to the QoS-based PA schemes described in [129].

### 3.3 Proposed Algorithm(s)

In this section, we outline our two proposed algorithms, namely, the NOMA-MEC algorithm for heterogeneous systems in Algorithm 1 and the NOMA-BB algorithm for homogeneous systems in Algorithm 2.

We begin with the NOMA-MEC algorithm to solve the cluster minimization problem in (3.13a) for heterogeneous systems in Algorithm 1. The goal is to minimize the number of clusters used while respecting the beamforming and user decoding capability constraints of each user in each cluster, as captured in (3.13c) and (3.13d), respectively. To do this, we break down the NOMA-MEC into two steps. In step-1, we find all possible valid cluster combinations, $C_v$, that respect both the constraints, (3.13c) and (3.13d). We refer to $C_v$, which is a set of valid user combinations, as the candidate list of clusters. Then, in step-2, from $C_v$, we find the minimum number of clusters that cover all the users exactly once. This is a MEC problem [126], hence we term the algorithm NOMA-MEC.

Step-1 of NOMA-MEC begins by building a list of users that can potentially be served on a NOMA cluster by each of the $B + 1$ candidate beams in $B_c$. This is obtained by iterating through the user-beam set, $B_u$, of all users, $u = \{1, ..., N\}$. Through this step, we get a list of users that can potentially cluster with each other. Let $n_b$ represent the number of users that have beam-$b$ in their user-beam set. Clusters can be of size $l = \{2, ..., d_{\max}\}$. We treat clusters of one separately as described later in this section. Hence, we form all $\binom{n_b}{1}$ groups of users, for all $B + 1$ beams. These are all potential clusters to be served in an orthogonal channel with a beam using precoding vector, $w_b$. Along with $w_b$, each user’s channel vectors are known. Thus, we can order the users according to their effective channel gains from smallest to largest in each of these potential clusters. Since we only allow users to be decoded in this order, if any cluster has a user at position $\pi(j)$, such that $d_{\pi(j)} < j - 1$, that cluster is invalid. Only those clusters that satisfy the decoding capability constraint.
Algorithm 1: NOMA-MEC

**Input:** Beam-list \( B_u \) of \( b \) beams for user \( u \) with channel \( h_u \) and decoding capability \( d_u, \forall u = [1, \ldots, N] \). Also, precoding vectors of candidate beams, \( w_b, \forall b = [1, \ldots, B] \).

**Output:** \( K \) clusters of ordered users such that each user-\( u \), is served in cluster-\( k \) (with beam \( b_k \)) at position-\( j \), such that \( d_{\pi_k(j)} \geq j - 1 \) and \( b_k \in B_u, \forall u = [1, \ldots, N] \)

**Step-1:** Build candidate list \( C_v \);

for (beam-\( b : B_c \)) do
  Find all \( n_b \) users that have beam-\( b \) in \( B_u \);
  Form set \( C_t \) by computing all possible \( \binom{n_b}{l} \) combinations, \( \forall l = [2, \ldots, d_{\text{max}}] \);
  for (\( c : C_t \)) do
    Order the users in \( c \) according to the effective channel gains, creating set of ordered users in \( c \) as \( \{u_1, \ldots, u_n\} \), such that
    \[ |w_b^H h_{u_1}|^2 \leq \cdots \leq |w_b^H h_{u_n}|^2 \]
    for \( u_1 \) to \( u_n \) do
      if \( d_{\pi_c(j)} < j - 1 \) then
        combination \( c \) is invalid, skip it.
        break; (outer for loop)
      end
    end
    Add \( c \) to candidate list \( C_v \). (If we did not break, combination \( c \) satisfies \( d \geq j - 1 \) for all users in \( c \)).
  end
end
Add each users- \( u \), as a cluster of one with their best beam from \( B_u \) to candidate list \( C, \forall u = [1, \ldots, N] \);

**Step-2:** Run greedy MEC on \( C_v \) to obtain \( C \);

\[ C_{\text{sorted}} = \text{sort } C_v \text{ in descending order}; \]
\[ x_c = 0 \forall c \in C_v; \]
\[ C = \emptyset, K = 0; \]
for \( c_o : C_{\text{sorted}} \) do
  if \( z_{u,c}x_c \neq 0, \forall c \in C \) then
    \[ C = \{C, c_o\}; \]
    \[ K = K + 1; \]
  end
end
return \( K, C; \)

(3.13d) for all the users in the cluster are added to the candidate list, \( C_v \). Finally, all users can be in a cluster of their own and be served like they would be with OMA.
Algorithm 2: NOMA-BB

**Input:** Best beam $B^*_u$ for user $u$ with channel $h_u \ \forall u = [1,..,N]$, common decoding capability $d_{\text{max}}$ for all users, precoding vectors $w_b$, $\forall b = [1,..,B]$.

**Output:** $K$ clusters of ordered users such that each user-$u$, is placed in a cluster served by its best beam, $B^*_u$ and $N_{C_k} \leq d_{\text{max}}, \forall k = [1..K]$.

$C = \{\}$, $K = 0$;

$m = d_{\text{max}}$;

for (beam-$b : B_c$) do

Find all $n_b$ users that have $B^*_u = b$ and group them in set $c : \{u_1,..,u_{n_b}\}$;

if $n_b > m$ then

for $i = 1 : \left\lceil \frac{n_b}{m} \right\rceil$ do

$j = \min(m \times (i+1), n_b)$;

$C = \{C, \{u_{m \times i+1},..,u_j\}\}$;

$K = K + 1$;

end

end

else

$C = \{C, c\}$;

$K = K + 1$;

end

end

return $K, C$;

Hence, we add $N$ elements to $C$, each being a cluster of one where user-$u$ is served on its best beam from $B^*_u, u = \{1,..,N\}$.

In step-2 of NOMA-MEC, from the list of viable candidate cluster options in $C_v$, we want to select the minimum number of elements that would cover every user exactly once. Since we added clusters of one for each user in the last part of step-1, we are guaranteed the existence of a solution. Let $x_c$ be a binary variable that represents if element-$c$ from set $C_v$ is selected and $z_{u,c}$ be a binary variable that represents if user-$u$ belongs to element-$c$ in $C_v$. The optimization problem can be stated as follows:
\[ \begin{aligned}
\min_{x_c} & \quad \sum_{c \in C_v} x_c, \\
\text{s.t.} & \quad \sum_{u=1}^{N} z_{u,c} x_c = 1, \ \forall c : C_v
\end{aligned} \tag{3.15a} \]
\[ x_c, z_{u,c} \in \{0, 1\}, \quad \tag{3.15b} \]

where (3.15a) represents the objective of the problem that minimizes the number of clusters and constraint (3.15b) ensures that all users occur exactly once in the final cluster set. This is a minimum exact cover problem, a known NP-complete problem [126]. We solve this problem using a greedy algorithm as follows. The first step is to sort the clusters in \( C_v \) in descending order of the number of users they contain, since using clusters in \( C_v \) that cover the most number of users allows us to minimize the number of clusters we need to cover all the users. We then go through the list of cluster combinations, adding cluster-\( c \) to \( C \) only if all users in the cluster have not been covered by clusters already in \( C \). The algorithm stops when all users have been covered exactly once, as highlighted in Algorithm 1.

The complexity of the algorithm is influenced by the following parameters - 1) the number of beams each user picks in its beam set, \( b \), which is an algorithm specific parameter, 2) the number of candidate beams, \( B \), which is a system level design parameter that we can control and 3) the number of users, \( N \), that need to be served along with their respective decoding capabilities, \( d_u, u \in [1, N] \). In step-1 of NOMA-MEC, building \( C_v \) involves the construction of \( I_1 \) clusters as follows:

\[ I_1 = \binom{n_b}{l} \times (B + 1), \quad \tag{3.16} \]

where \( l = \{2, ..., d_{\text{max}}\} \). In each of these \( I_1 \) clusters, the users have to be ordered and then analyzed to check whether each users decoding capability criteria are satisfied. The parameter \( n_b \), the number of users that have beam-\( b \) in their user-beam set, scales with \( N \) and \( b \). The second step is the minimum exact cover problem. Let \( I_2 \) represent the number of valid combinations in \( C_v \) that the greedy algorithm in MEC needs to explore to find \( C \). In a homogeneous system, all the \( Z \) clusters are valid clusters, which means \( I_2 = I_1 \). Thus, a homogeneous system represents the worst-case complexity for the MEC part of the NOMA-MEC algorithm. However, in general, a
large number of the original cluster combinations will be rejected due to them being unable to meet the user decoding capabilities, resulting in \( I_2 < I_1 \). This in turn controls the complexity of the MEC part of the algorithm. This is discussed further in Section 3.4, supported by simulation figures.

The choice of \( b \) is the most important design parameter for the NOMA-MEC algorithm. From a performance perspective, a larger \( b \) gives the algorithm the ability to find a larger number of cluster combinations that satisfy the decoding capability constraint. So, strictly from the perspective of minimizing \( K \) in (3.13a), a large \( b \) is good. However, as \( b \) increases, we add beams that are less and less aligned with the user direction to the user-beam set \( B_u \), reducing the beamforming gain with each increment of \( b \). Due to the overlapping nature of the beams as seen in Fig. 3.1, there is a value of \( b \) such that the user is within the coverage area of all of its best \( b \) beams in \( B_u \). Let this value of \( b \) be \( b_{th} \). However, as \( b \) is further increased beyond \( b_{th} \), the user gets out of the coverage area of the beam-(\( b_{th}+1 \)) and if NOMA-MEC schedules the user on beam-(\( b_{th}+1 \)), it will have poor spectral efficiency, \( R_u \), bringing down the overall spectral efficiency, \( R_{sum} \), in the process. The exact value of \( b_{th} \) depends on system-level parameters \( M \) and \( B \). The number of antennas, \( M \), determines the width of the beam and together with the number of candidate beams, \( B + 1 \), determines the level of overlap between the beams and hence also determines the value of \( b_{th} \). Additionally, \( b \) is also an important parameter to control the complexity of NOMA-MEC. If the number of users is large or \( d \) is large for most users, \( b \) can be reduced to lower \( I_1 \). For a homogeneous system with a large \( d_{max} \), the number of combinations can be very large and so we need to scale back \( b \). If \( b = 1 \), it is equivalent to having each user pick its best beam. Hence, for the homogeneous system with large \( d_{max} \), we propose a low-complexity clustering algorithm called NOMA-best beam (NOMA-BB) that has each user served by its best beam in \( B_u \), as we outline next.

In the NOMA-BB algorithm, like NOMA-MEC, we iterate through each beam and build the list of \( n_b \) users that picked each beam-\( b \). However, compared to NOMA-MEC, the difference is that in this case, users have to belong to that beam since we follow the best beam strategy where each user picked only one beam in their user-beam set. Hence, these groups of users are effectively our clusters except that we might have beams that have more than \( d_{max} \) users in it, leading to some users needing to decode more than \( d_{max} \) users signals, which violates the SIC decoding capability constraint. Hence, for all beams where \( n_b > d_{max} \), we break up the one
cluster of \( n_b \) users into \( \lceil \frac{n_b m}{m} \rceil \) clusters, where \( m \) is an integer between 1 and \( d_{\text{max}} \), i.e., \( m \in [1, d_{\text{max}}] \), that controls the maximum number of users per cluster and \( \lceil . \rceil \) is the ceiling function. Since the goal is to minimize the number of clusters, we set \( m = d_{\text{max}} \). Setting \( m = d_{\text{max}} \) is feasible because all users have the same decoding capability, \( d = d_{\text{max}} \), and so any user ordering among the \( N_{C_k} \) users in some cluster-\( C_k \) formed by NOMA-BB would be valid, as long as \( N_{C_k} \leq d_{\text{max}} \). In this chapter, when we need to split \( n_b \) users in a beam into multiple clusters, i.e., when \( n_b > d_{\text{max}} \), we arbitrarily split the users into different clusters. As a future work, a more advanced NOMA-BB clustering schemes could aim to maximize the channel disparity between the users in the cluster when doing this split, a condition known to improve the rate in NOMA systems [130].

### 3.4 Simulation Results and Discussion

The performance of the proposed NOMA-MEC and NOMA-BB algorithms are evaluated using MATLAB simulations, with the system parameters described in Table 3.2. The mmWave channel model in (3.1) is considered, where \( L = 1 \), \( \eta = 2 \) and \( D/\lambda = 1/2 \) for the ULA steering vector. The BS is equipped with \( M = 8 \) antennas.

**Table 3.2:** Simulation parameters

<table>
<thead>
<tr>
<th>Parameter name, notation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of paths, ( L )</td>
<td>1 [10, 12, 120]</td>
</tr>
<tr>
<td>Path loss exponent, ( \eta )</td>
<td>2 [12]</td>
</tr>
<tr>
<td>BS antenna spacing</td>
<td>( \frac{D}{\lambda} ) [12]</td>
</tr>
<tr>
<td>Wavelength of the carrier signal</td>
<td>( \frac{D}{\lambda} ) [12]</td>
</tr>
<tr>
<td>Noise power, ( \sigma^2 )</td>
<td>( 7.962 \times 10^{-11} ) [12]</td>
</tr>
<tr>
<td>Users minimum QoS SINR, ( \Gamma_{\text{min}} )</td>
<td>0.02 [12]</td>
</tr>
<tr>
<td>Number of antennas at BS, ( M )</td>
<td>{2, 4, 8, 16, 32}</td>
</tr>
<tr>
<td>Number of candidate beams, ( B )</td>
<td>{10, 20, 30, 40}</td>
</tr>
<tr>
<td>Number of users in the system, ( N )</td>
<td>{50, 100, 150, 200}</td>
</tr>
<tr>
<td>Max. user decoding capability, ( d_{\text{max}} )</td>
<td>{0, 10}</td>
</tr>
<tr>
<td>User distribution</td>
<td>Randomly distributed around 5 meters radius from BS. [10]</td>
</tr>
</tbody>
</table>
unless specified otherwise. The noise power is $\sigma^2 = -174 + 10\log_{10}(W) + N_f$ dBm, where $W = 2$ GHz is the system bandwidth and the noise floor $N_f = 10$ dB. The users are randomly distributed around the BS within a 5 meter radius, i.e., $r_u \leq 5$. We consider the minimum user QoS to be an average of $N \times \Gamma_{\text{min}}$ bps per channel-use. Since the users are scheduled in $K \leq N$ channels ($K$ NOMA clusters), we can simplify this requirement by just considering a minimum user rate of $\Gamma_{\text{min}}$ for each user in every cluster. In the simulations, $\Gamma_{\text{min}} = 0.02$. Finally, the number of candidate beams is $B = 20$. In the simulations for heterogeneous systems, we set $d_{\text{max}} = 5$ and generate each user’s decoding capability, $d$, as a random integer in the range $[0, d_{\text{max}}]$. For the homogeneous systems, we vary $d_{\text{max}}$ from 1 to 10.

We start by evaluating the NOMA-BB algorithm for a homogeneous system, where each users decoding capability is $d = d_{\text{max}}$ that we vary from 1 to 10 in this simulation. In Fig. 3.2a, we compare the spectral efficiency, measured in bps per channel use, for the NOMA-BB algorithm against OMA for a system with 50, 100, 150, and 200 users. OMA is not influenced by the value of $d_{\text{max}}$ as it has to serve one user per cluster, irrespective. As seen in Fig. 3.2a, a NOMA setting with $d_{\text{max}} = 1$ is equivalent to OMA. As $d_{\text{max}}$ increases beyond one, we start to see the gain of NOMA. A higher value of $d_{\text{max}}$ means all users are capable of decoding more number of other users signals, i.e., we can serve more users per-cluster. Looking at the NOMA-BB algorithm, for each beam-$b$, we split the $n_b$ users who picked beam-$b$ into $\lceil \frac{n_b}{m} \rceil$ clusters, where $m = d_{\text{max}}$. Clearly, as $d_{\text{max}}$ increases, $m$ increases, and so NOMA-BB needs to form fewer clusters in this splitting step. This is illustrated in Fig. 3.2b, where the number of clusters, $K$ required to serve the $N$ users decreases as $d_{\text{max}}$ increases. Further, as the number of users in the system, $N$, increases, the likelihood of having beams with more than $d_{\text{max}}$ users increases in the first step of the NOMA-BB algorithm. As a result, for higher $N$, we see the number of clusters decrease in Fig. 3.2b by increasing $d_{\text{max}}$ for longer before it starts to flatten out. Correspondingly, the rate in Fig. 3.2a increases with $d_{\text{max}}$ for longer when $N$ is larger.

We now move to heterogeneous systems and evaluate the performance of our proposed NOMA-MEC algorithm, compared against OMA and NOMA-BB with slight modifications to account for the heterogeneous decoding capability constraints. We note that there are no direct user clustering schemes in the literature that considers individual user decoding capabilities for us to compare against. NOMA-BB is fairly
Figure 3.2: Performance of NOMA-BB for a homogeneous system where each users' SIC decoding capability constraint $d = d_{\text{max}}$, effectively making $d_{\text{max}}$ the maximum number of users that can be placed in a cluster. (a) Overall spectral efficiency compared to OMA (b) Number of clusters, $K$, required to serve the $N$ users.
Figure 3.3: Performance of NOMA-MEC for a heterogeneous system where each user has its own decoding capability constraints. NOMA-MEC is compared against OMA and NOMA-BB-Het, a modified version of NOMA-BB to account for the heterogeneous user decoding capabilities.

typical of most NOMA clustering schemes in the literature that do not have individual restrictions on each user’s SIC decoding position, and so offers good insights for us to compare our proposed NOMA-MEC against. However, to run NOMA-BB in a heterogeneous system, we cannot set $m = d_{\text{max}}$ like we could for a homogeneous system, since each user has its own decoding capability constraint and so not all clusters will result in feasible decoding capability constraint and so not all clusters will result in feasible decoding order combinations, even if the cluster size is capped at $d_{\text{max}}$. To make NOMA-BB work for a heterogeneous system, we need to separate out all users with $d < m$, for any $m \in [1, d_{\text{max}})$, and then divide the remaining users into $\lceil \frac{n}{m} \rceil$ clusters. This would ensure that the arbitrary user ordering done by the NOMA-BB scheme does not violate any user’s SIC decoding capability constraint. A larger $m$ means we can form larger clusters but will have to exclude more users with the extreme case of $m = d_{\text{max}}$ equivalent to OMA and so we exclude it, while a smaller $m$ means we will form smaller clusters, but exclude less users. We term this modified version of NOMA-BB for heterogeneous systems as NOMA-BB-Het and run it with all possible values of $m \in [1, d_{\text{max}}), d_{\text{max}} = 5$, for the simulations in Fig. 3.3.
Figure 3.4: NOMA-MEC run in a heterogeneous system with users having their SIC decoding capability, $d$, randomly distributed in $[0, d_{\text{max}}]$, is compared against NOMA-BB and NOMA-MEC run for a homogeneous system where all users have $d = d_{\text{max}}$ ($d_{\text{max}} = 5$).

which we discuss next.

Analyzing the performance of NOMA-MEC from Fig. 3.3, we see that despite the restrictions put in place by the heterogeneous user decoding capabilities, we still see a significant performance gain over OMA. It also outperformed NOMA-BB-Het for all values of $m$, because NOMA-BB-Het and other such clustering algorithms from the literature do not consider restrictions on each individual user’s capabilities while clustering. In Fig. 3.4, the NOMA-MEC heterogeneous scheme running in a heterogeneous system with all users having a random value of $d$ in the range $[0, d_{\text{max}}]$, is compared against a hypothetical homogeneous system where all users have $d = d_{\text{max}}$.

For the hypothetical homogeneous system, the original NOMA-BB and NOMA-MEC that assumes homogeneous user decoding capability with $m = d_{\text{max}}$ is run. We note that the NOMA-MEC algorithm can easily be run for a homogeneous system, with all users having $d = d_{\text{max}}$. We see that the NOMA-MEC algorithm for the heterogeneous deployment closely shadows, but always trails, the NOMA-MEC run for the homogeneous deployment. The flexibility of the proposed NOMA-MEC algorithm is highlighted by this observation since it says that even though each user is posing its own decoding restrictions of $d \leq d_{\text{max}}$, we are still able to achieve close to the
Figure 3.5: Analyzing NOMA-MEC in terms of the number of clusters, $K$, outputted by the algorithm, as the number of heterogeneous users in the system increases for different values of $b$.

Performance we could if there was a simple maximum users per cluster constraint of $d = d_{\text{max}}$. The hypothetical homogeneous deployment is still better because in the NOMA-MEC for the homogeneous deployment, all $I_1$ cluster combinations examined is step-1 of NOMA-MEC are valid and entered into $C_v$ for the MEC algorithm to choose from. However, the NOMA-MEC for heterogeneous systems strips a large chunk of these $I_1$ combinations away due to not satisfying the user decoding capability constraints and so gives fewer pairing options in $C_v$ for the greedy MEC algorithm to work with when trying to minimize $K$. Looking at just the homogeneous curves in Fig. 3.4, NOMA-MEC (Hom.) with $b \in [2, 4]$ outperforms NOMA-BB. This is expected as the NOMA-MEC algorithm is more advanced, allowing users to pick multiple candidate beams for clustering, giving more clustering opportunities.

Additionally, analyzing the trends of NOMA-MEC in both Fig. 3.3 and Fig. 3.4, we see the rate increase at first as $b$ increases, but then starts to drop-off as we increase $b$ further. This is a consequence of the trade-off between a larger search space to reduce $K$ and the beamforming gain from allowing users to be served on their stronger beams, as discussed in Section 3.3. A larger choice of $b$ implies a larger candidate cluster list $C_v$, in the NOMA-MEC algorithm, allowing the MEC...
part of the algorithm to find solutions with a lower number of clusters, $K$. This is illustrated in Fig. 3.5, where for any number of users in the system, the number of clusters required to serve the $N$ users, i.e., $K$, decreases as $b$ increases. However, as $b$ increases beyond the $b_{th}$, users are adding beams to their user-beam-set, $B_u$, that they are less aligned with in terms of the cosine similarity metric. In other words, $\forall$ beam-$b$ in $B_u$, $b > b_{th}$, NOMA-MEC can potentially schedule the user in a cluster served by beam-$b$, even though the user is out of the coverage area of beam-$b$. As seen in Fig. 3.3 and Fig. 3.4, at first as $b$ goes from one to two, the extra clustering opportunities allow us to reduce $K$ as well as not incur too much of a penalty in terms of the beamforming gain. However, as $b$ increases further, the penalty from sacrificing the beamforming gain outweighs the further cluster reduction we are able to achieve and hence we see the spectral efficiency start to drop off after that. The exact value of $b$ at which this reversal occurs depends on the number of candidate beams, $B$ and the width of these candidate beams, which is a consequence of the number of transmit antennas, $M$. Next, we discuss the impact of $B$ and $M$ on the choice of $b$ from a performance perspective.

The simulations in Fig. 3.6 were conducted to understand the impact of parameters $B$ and $M$ respectively on the performance of NOMA-MEC. In particular, we are focused on the trend where we first see a performance improvement as $b$ increases, but then a decline as $b$ is further increased. In Fig. 3.6a, we vary $M$ while keeping $B$ fixed, $B = 20$. As the number of transmit antennas at the BS, $M$, increases, the BS is able to form more narrow beams. For a fixed value of $B$, as $M$ increases, the amount of overlap between the candidate beams decreases. This means that a user is located in the coverage area of a smaller number of beams, i.e., $b$ needs to be smaller to realize the beamforming gain. Conversely, as $M$ decreases for a fixed $B$, the overlap between the beams increases, as we have the same number of wider beams. This allows for a user to be located in the coverage area of more number of beams, i.e., a larger $b$ can be chosen. The same analysis can be done for a fixed $M$ and varying $B$ in Fig. 3.6b. In this case, the beam width is fixed since $M$ is fixed. However, as $B$ increases, the overlap increases as we have more candidate beams covering the same coverage area, $\bar{\phi}$. As a result, we see the performance gain from increasing $b$ for longer in Fig. 3.6b for larger values of $B$. We see that with every increase of $B = 10$, the drop-off starts to occur $b = 1$ later. Hence, irrespective of the values of $B$ and $M$, the trend is the same - we first see a gain from increasing $b$, but then the performance starts to drop
Figure 3.6: Analyzing the impact of $b$ on the performance of NOMA-MEC as system parameters $B$ and $M$ vary.

off once we start to lose the beamforming gain as $b$ is further increased. For a small value of $M$ and a large value of $B$, the value of $b$ before the drop-off starts to occur is larger than if we had a small $B$ or large $M$. Hence, from a performance perspective, $b$ needs to be selectively tuned as a function of $B$ and $M$. We discuss the impact of parameter $b$ on the complexity of the algorithm next.

As described in Section 3.3, the NOMA-MEC algorithm presented in Algorithm 1 consists of two steps. The first is the formation of candidate list $C_v$ by iterating
through each beam in the list of candidate beams, looking for all valid combinations of users who could be scheduled together in a cluster served by the beam. Each possible combination requires users to be ordered by their effective channel gain and then the decoding capability of each user has to be checked to see if the combination is valid or not. The second part of the algorithm involves taking the candidate list, \( C_v \), and running the greedy algorithm to solve the minimum exact cover problem. Simulation runs to present the number of iterations required at both steps of the NOMA-MEC algorithm, \( I_1 \) and \( I_2 \) respectively, are presented in Fig. 3.7a and Fig. 3.7b respectively.

The number of iterations in step-1 of NOMA-MEC, \( I_1 \), is determined by \( n_b, l \) and \( B \) as equation (3.16) shows. The term \( n_b \), which corresponds to the number of users that contain each beam-\( b \) in their user-beam-set, is influenced by the number of users in the system \( N \) and the number of beams per user set, \( b \). Since NOMA-MEC iterates through \( \binom{n_b}{l} \), \( l = \{2, \ldots, d_{\text{max}}\} \), combinations to check for valid cluster combinations based on the user decoding capability, the impact of \( l \) on \( I_1 \) is entirely determined by \( d_{\text{max}} \). The complexity of step-2 depends entirely on the size of set \( C_v \), i.e., the number of valid combinations found after step-1, i.e., \( I_2 \). If \( I_1 \) is large, \( I_2 \) is likely to be large too and so the same factors that influence \( I_1 \) in step-1 also affect the complexity of step-2. However, since a large number of combinations examined in step-1 are deemed invalid and not added to \( C_v \) in step-1 due to the SIC decoding capability restrictions, the size of \( C_v \) will still be small. This is illustrated in Fig. 3.7b, where we see a significant reduction in the candidate cluster list size, \( C_v \) compared to the number of cluster combinations examined in Fig. 3.7a. As discussed in Section 3.3, if NOMA-MEC were run on a homogeneous system, \( I_2 = I_1 \), leading quickly to a prohibitively high complexity for the greedy algorithm to solve the MEC problem. On the other end of the spectrum, if a large number of combinations examined in step-1 are deemed invalid due to the SIC decoding capability constraints, the size of \( C_v \) will still be small. This is likely in deployments where the majority of users are IoT users with \( d \in [0, 1] \) and there are only a handful of users with a larger value of \( d \), e.g., cellular users. In such a case, the complexity of step-1 will be high since \( d_{\text{max}} = \max(d) \) is still large. However, if most users have \( d \in [0, 1] \), then most examined combinations in step-1 will be deemed invalid, still leaving a manageable size of \( C_v \) for the greedy algorithm in step-2 to work with. However, in the simulations in Fig. 3.7, we only considered the case where the users’ decoding capabilities, \( d \), are randomly generated from \([0, d_{\text{max}}]\). In future works, we will consider more skewed distributions of \( d \) and explore how...
Figure 3.7: Impact of the number of beams per user set, $b$, on the complexity of NOMA-MEC in terms of (a) $I_1$ which represents the number of cluster combinations examined in step-1 of NOMA-MEC and (b) $I_2$, which represents the number of valid combinations in $C_v$ to be considered by the greedy algorithm to solve the MEC problem.

The complexity of step-1 of the NOMA-MEC algorithm can be reduced by exploiting advanced knowledge of the distribution of the users’ decoding capabilities.

Finally, it is worth mentioning that the parameter $b$ in the NOMA-MEC algorithm can be set by considering the performance-complexity tradeoff as follows. We have
seen that increasing $b$ improves performance up to $b = b_{th}$, but then the performance declines as $b$ is increased any further. Depending on system parameters $B$ and $M$, we can find the value of $b_{th}$ at which the performance gains from increasing $b$ peaks. After that, the complexity aspect can be considered. If the complexity is acceptable at $b_{th}$, that would be a logical choice for $b$. However, for systems with a large $N$ or $d_{\text{max}}$, setting $b = b_{th}$ could result in a prohibitively high algorithm complexity as seen in Fig. 3.7. In such settings, $b$ can be reduced to bring down the complexity as seen from Fig. 3.7, at the expense of performance.

3.5 Conclusion

In this chapter, we proposed a joint user clustering and user ordering scheme, namely, NOMA-MEC, for an ABF mmWave-NOMA system that can serve a set of users that each have their own SIC decoding capability constraints. By using the reported SIC decoding capability constraint from each user to set the maximum position in the SIC decoding order for clusters that are always decoded in the order of their effective channel gains, we framed the problem as a minimum exact cover optimization problem. Despite each user posing individual conditions on how many other users’ signals it can decode in the SIC decoding order, simulation results demonstrated that the proposed algorithm still offers significant spectral efficiency gains over OMA as well as over other NOMA clustering algorithms that do not have the flexibility to accommodate for such user decoding requirements. We provided a detailed analysis of the performance-complexity trade-off from the setting of parameters related to the NOMA-MEC algorithm as well as system-level parameters. Finally, for a homogeneous system where all users have the same decoding capability requirements, we showed that this boils down to a simpler condition of restricting the number of users per cluster. We proposed a simpler NOMA-BB algorithm for the homogeneous system and also evaluated its performance through simulations.
Chapter 4

Neural Network Aided User Clustering in mmWave-NOMA Systems with User Decoding Capability Constraints

4.1 Introduction

The problem with applying optimization techniques, even low-complexity heuristics such as the NOMA-BB and NOMA-MEC used in Chapter 3 in mmWave-NOMA systems, in live networks for user clustering is that they require a very large number of computation steps to make a clustering decision. If these clustering decisions are based on the instantaneous channel of hundreds of users, it becomes prohibitively complex to implement in practical systems on a millisecond granularity as required by beyond 5G (B5G) systems. To address this issue, this chapter builds on the work of Chapter 3 and proposes a computationally efficient two-stage machine learning based approach using neural networks to solve the cluster assignment problem in a millimeter wave-non orthogonal multiple access (mmWave-NOMA) system where each user’s individual successive interference cancellation (SIC) decoding capabilities are taken into consideration. This work implements a version of the proposal we presented in the patent [131].

The artificial neural network (ANN) is applied in real time to assign users to clusters taking each user’s instantaneous channel state information (CSI) and SIC decoding capabilities as inputs. The algorithm is trained offline on cloud resources, i.e., not using the base station (BS) compute resources. This training is done using
Table 4.1: Comparison of this work with existing literature

<table>
<thead>
<tr>
<th>Papers Category</th>
<th>References</th>
<th>System Model</th>
<th>Clustering Problem Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIC decoding paper</td>
<td>[132]</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Normal clustering papers</td>
<td>[118]</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>Unsupervised clustering papers</td>
<td>[12, 123, 124, 133]</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>Supervised clustering papers</td>
<td>[134, 135]</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>This work</td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

a dataset obtained by offline computation of input parameters using the NOMA-BB and NOMA-MEC optimization algorithms. As a result, we term the proposed algorithms in this chapter as ANN-NOMA-MEC and ANN-NOMA-BB, respectively.

4.1.1 Related Work

To address the real-time complexity issue with user clustering schemes that need to be run on a millisecond granularity, machine learning has often been proposed an enabler. The work in [138] classified user clustering problems in mmWave-NOMA systems into joint resource aware user clustering techniques such as the ones applying the cosine similarity metric as described in Chapter 3 and a second class of algorithms called learning assisted user clustering techniques to bring down the complexity. In [138], the complexity of several user clustering schemes in the mmWave-NOMA literature is analyzed and it is shown that a significant run-time complexity is inherited by all schemes to make clustering decisions on a millisecond granularity, especially as the network size grows. However, the machine learning techniques applying K-means like clustering algorithms bring down the complexity compared to traditional optimization schemes.

There are two main machine learning based themes of work that are relevant to the discussion in this chapter. The first is the theme of work in [12, 123, 124, 133], where the user clustering and ordering problem in mmWave-NOMA systems is solved using an unsupervised clustering ML approach. These works exploit the high correlation amongst users’ channels and the fact that mmWave propagation is dominated by the LoS path to effectively employ K-means clustering. In [134], an advancement on k-means for the cluster formation problem is proposed. Since the effects of multipath propagation are limited in mmWave spectrum, the user clustering in mmWave-NOMA systems comes down to finding spatially correlated users with the available CSI at the BS. A concrete clustering metric based on channel correlation among users is proposed.
This is precisely what unsupervised clustering algorithms are capable of achieving without any labeled training data. On a similar line, in [140], a multi-label classification problem is framed to solve the user clustering problem. One other theme of work in general NOMA systems is that from Kumaresan et al. in [135–137]. The authors in [137] propose a ANN-based user clustering framework that learns from a training data set obtained through simulation settings and a brute force search approach of all possibilities [135]. The dataset is used to train the neural network that is then applied to the clustering decisions in the network. In [136], the same authors extended the machine learning scheme to instead use the extreme machine learning (ELM) method to solve the clustering problem. Finally, user clustering solutions in other NOMA based system models such as one incorporating device-to-device communication proposed in [141] also run into similar problems of exponential run-time complexity to solve the clustering problem in a live network.

One important limitation with these schemes is that there is no flexibility incorporated to account for the SIC decoding capability limitations of each individual user. The work in the previous chapter, published in [132], proposed a user clustering and ordering scheme for a mmWave-NOMA system that takes into consideration the SIC decoding capability of each individual user in the system. To this end, two heuristics were proposed:

1. A NOMA-minimum exact cover (NOMA-MEC) algorithm for heterogeneous systems.

In this chapter, we use these algorithms to build up the labelled training dataset for the proposed neural network to learn from, as discussed next.

4.1.2 Motivation and Contributions

In order to actually deploy mmWave-NOMA systems, practical considerations need to be factored in. One example is the availability of instantaneous CSI of large number of users, and the complexity involved in making clustering decisions with this large amount of CSI information. In order to mitigate the availability of CSI, location aided [142] and vision aided [143] clustering techniques have been proposed in the literature as well as techniques to combat imperfect CSI availability [144]. However,
in this work we will assume the full CSI of each user is available to the BS. Even so, a user clustering and ordering problem (UCOP) in mmWave-NOMA systems that relies on the instantaneous channel information of hundreds of users, needs to find the right balance between performance and complexity while considering practical limitations such as the SIC decoding capability of the user and the availability of CSI in order to be usable in practical deployments. In terms of performance, the UCOP scheme needs to generate a clustering result that maximizes or close to maximizes the system throughput while satisfying each users minimum QOS constraint. However, to lead to a feasible solution, the UCOP needs to factor in the SIC decoding capabilities of the users in the system. For a homogeneous system, that amounts to just limiting the number of users per cluster while for heterogeneous systems, it means accounting for each users SIC decoding capabilities. The NOMA-MEC and NOMA-BB schemes proposed in [132] addressed these requirements but the complexity of the algorithms was quite large to be run at a millisecond granularity. As shown in [132], depending on the parameter settings, up to 60,000 computation steps need to be executed to make a clustering decision every millisecond. Due to the latency sensitive nature of this computation, it cannot be offloaded to the cloud or an external entity as the round-trip delay would be too large. With low-cost and small cell BSs being increasingly studied for B5G systems, there is a need to manage the computational resource needs available at the cell site. To this end, we propose a novel machine learning based approach using neural networks to factor in the SIC decoding capabilities of the users and
build a low-complexity scheme that produces clustering results on par with NOMA-MEC and NOMA-BB for heterogeneous and homogeneous systems, respectively. We distinguish our work from the other machine learning user clustering schemes in the NOMA literature by factoring in the SIC decoding capabilities of the users into the machine learning scheme. For example, compared to the unsupervised clustering schemes in [12, 123, 124], we use a supervised learning scheme as it is a better fit to the clustering problem once SIC decoding capabilities are considered. Compared to the supervised learning schemes presented in [135–137], we distinguish ourselves again by studying a neural network that is capable of forming clusters while factoring in the SIC decoding capabilities of the users which these other works do not take into consideration. Table 4.1 illustrates how the system model and clustering problem formulation of this work is distinguished from existing work in the literature.

The proposed neural network is trained offline on labelled data samples generated using the NOMA-MEC and NOMA-BB algorithms. Concretely, for different possible inputs of user positions, channels and SIC decoding capabilities, the NOMA-MEC or NOMA-BB algorithm is run and a labelled data set is generated. This labelled dataset is used to train the neural network. As a result, we term the two proposed algorithms in this chapter as ANN-NOMA-MEC and ANN-NOMA-BB, for heterogeneous and homogeneous systems, respectively. This training of the neural network can be done offline from the BS, possibly on cloud compute resources. The trained neural network is then applied in the live network with the input comprising of the users channels and SIC decoding capabilities. This two-stage flow is illustrated in Fig. 4.1.

Applying this trained neural network model directly on the users inputs allows for it to be run at the millisecond granularity that is required to make optimal clustering decisions as users channels change or users enter or leave the system. Users SIC decoding capabilities can also change over time as a function of the battery level of the users. Hence, this model continues to make a clustering decision considering a whole fresh set of input every millisecond or so, but in directly applying a trained model to do so, this is a far more reasonable cost to take at the BS. Our contributions in this chapter can thus be summarized as follows:

- We propose two ANNs for user clustering factoring in SIC decoding capability constraints in mmWave-NOMA systems, namely ANN-NOMA-MEC and ANN-NOMA-BB, for heterogeneous and homogeneous systems respectively, that are trained offline using labelled datasets generated from running the NOMA-MEC
and NOMA-BB algorithm on randomly generated inputs of users channels and SIC capabilities.

- Simulation results from running the trained neural network with test data shows that it performs comparably to the heuristics it was trained from. This validates the proposed approach of training networks to make clustering decisions offline and then applying the trained model directly in live networks to make clustering decisions on a milisecond granularity.

4.2 System Model and Problem Formulation

The system model used in this section is the same as the one used in Chapter 3. For completeness here, we outline again the key aspects of the objective cluster minimization sub-problem to the overall rate maximization problem and the relevant details of the NOMA-MEC algorithm from [132] that the neural network uses to create the data sets that it learns from.

There are $N$ single-antenna users, each with a minimum QoS constraint that need to be served. Each of these users have their own individual end user signal processing capabilities related to the SIC decoding procedure in NOMA that needs to be modelled. These $N$ users are served by a single-cell BS having $M$ antennas and operating in the mmWave spectrum; where the mmWave channel between the BS and user-$u$ can be modelled in a similar way to [120, 132] and is given as

$$h_u = a(\theta_u)\frac{\alpha_u}{\sqrt{L(1 + r_u^\eta)}}, \quad (4.1)$$

where $a(\theta_u)$ represents the steering vector, $\alpha_u$ represents the complex channel gain for user-$u$, $L$ denotes the number of paths, $r_u$ denotes the distance between the BS and user-$u$ and $\eta$ denotes the path loss exponent. Further, like [132], analog beamforming (ABF) is used and only one beam can be transmitted at a time, which can be equated to forming one beam to serve one cluster of NOMA users per channel use, e.g., per time slice. Let the entire coverage region, $\bar{\theta}$, from $-\pi/2$ to $\pi/2$ be covered by a set of $B + 1$ candidate beams that the BS can create through ABF. Thus, a NOMA cluster of users will be served on an orthogonal channel using one of these $B + 1$ beams. Each
beam-\(b\) in this list of candidate beams has the precoding vector,

\[ \mathbf{w}_b = a(\bar{\theta}_b), \forall b \in [0, B], \]  

(4.2)

where the parameter \(\bar{\theta}_b\) is

\[ \bar{\theta}_b = -\pi/2 + (b \times \pi/B). \]  

(4.3)

Thus, there are a set of \(B + 1\) precoding vectors from (4.2) that cover the coverage region of \(\bar{\theta} = -\pi/2\) to \(\pi/2\), including some possible overlap between the beams, depending on the value of \(B\). Using the same notations as [132], let \(B_c\) represent this list of candidate beams, such that \(B_c = \{\text{Beam-0}, ..., \text{Beam-}\(B\)\}\), with their respective list of candidate precoding vectors being \(W_c = \{\mathbf{w}_0, ..., \mathbf{w}_B\}\).

There are \(N\) users in the system and they can be clustered into \(K\) NOMA clusters, \(K \leq N\). Each cluster is served by one of the \(B + 1\) precoding vectors in the candidate list in one orthogonal channel, e.g., time slice. Let \(C = \{C_1, ..., C_K\}\) represent the \(K\) clusters that are selected to collectively serve the \(N\) users in \(K\) time slices, where \(C_k\) refers to the \(N_{C_k}\) users selected to serve in the cluster with index-\(k\), \(N_{C_k} \leq N\). Let beam-\(b_k\) with a precoding vector \(\mathbf{w}_{b_k}\) represent the beam selected for cluster \(C_k\). In each cluster, \(C_k\), the BS applies superposition coding (SC) as follows:

\[ s_k = \sum_{u=1}^{N_{C_k}} \sqrt{p_u} s_{k,u}, \]  

(4.4)

where \(p_u\) represents the power allocated to user-\(u\). The received signal at user-\(u\) in cluster \(C_k\) is

\[ y_u = h_u^H \mathbf{w}_{b_k} s_k + \xi_u, \]

\[ = h_u^H \mathbf{w}_{b_k} \sqrt{p_u} s_{k,u} + h_u^H \mathbf{w}_{b_k} \sum_{u' \neq u, v=1}^{n_k} \sqrt{p_v} s_{k,v} + \xi_u. \]  

(4.5)

We let \(\pi_k(j)\) denote the user index for the \(j\)-th decoded user in the cluster \(C_k\) serving \(N_{C_k}\) users, \(j \leq N_{C_k}\). The \(j\)-th user needs to first decode the signals of users \(\{\pi_k(1), ..., \pi_k(j)\}\) before decoding its own signal in the SIC procedure. When decoding
user $\pi_k(j)$ at user $\pi_k(j')$, $j' > j$, the signal-to-interference-plus-noise ratio (SINR) can be represented as

$$
\Gamma_{\pi_k(j)}^{\pi_k(j')} = \frac{p_j |h_{(j')}^H w_k|^2}{|h_{(j')}^H w_k|^2 \sum_{v > j} N_{C_k} p_v + \sigma^2},
$$
(4.6)

where $\sigma^2$ is the noise power. If we let $R_k$ denote the rate achieved in NOMA cluster $C_k$, the effective sum rate of the system, $R_{\text{sum}}$ can be expressed as the sum of the rates from each cluster, $R_k$, given that each cluster is served by one channel. The effective sum rate can thus be represented as

$$
R_{\text{sum}} = \frac{\sum_{k=1}^{K} R_k}{K} = \frac{\sum_{k=1}^{K} \sum_{u \in C_k} \log_2 \left(1 + \Gamma_{\pi_k(u)}^{\pi_k(u)}\right)}{K},
$$
(4.7)

expressed in bits per second (bps) per channel-use.

The SIC decoding capability constraint of each user-$u$ is modelled as $d_u$ [132]. For example, $d_u = 4$ implies a user capable of decoding four other users’ signals. This impacts the user ordering done within a cluster. If a user-$u$ is placed in cluster-$k$ at position $j$, then the maximum value of $j$ is $d_u$. We let $d_{\text{max}} = \max(d_u), \forall u = [1, .., N]$, represent the maximum decoding capability among the $N$ users in the system. A heterogeneous system allows each user to have its own value of $d_u$, $d_u \leq d_{\text{max}}$; while a homogeneous system puts the additional constraint that all users have the same value of $d_u = d_{\text{max}}$.

The objective optimization problem that needs to be solved is the same rate maximization problem as in [132], which we will instead tackle with the neural network approach and can be given as

$$
\max_{\{C_k\}, \{w_k\}, \{\pi_k\}, \{p_u\}} R_{\text{sum}},
$$
(4.8a)

s.t. $R_u \geq \log_2 (1 + \Gamma_{\text{min}})$, $\forall u = 1, .., N$
(4.8b)

$$
d_{\pi_k(j)} \geq j - 1, j = 1, .., N_{C_k}, \forall k = 1, .., K,$$
(4.8c)

$$
\sum_{i=1}^{N_{C_k}} p_i \leq P, \forall k = 1, .., K,$$
(4.8d)
where $\Gamma_{\text{min}}$ denotes the minimum SINR with which each user needs to be served, i.e., $\Gamma_{\pi_k(u)} \geq \Gamma_{\text{min}}, \forall u = [1, .., N]$, (4.8b) represents the QoS constraint, (4.8c) represents the decoding capability constraint, and (4.8d) represents the power per channel constraint. As discussed in our survey paper in [132], the rate optimization problem in (4.8a) is a joint optimization problem that involves optimal power allocation along with optimal user-clustering, that leads to a non-convex and combinatorial non-polynomial (NP)-hard optimization problem for which an exhaustive search is needed to find the optimal solution. Hence, the most common approach as shown in [132] is to break down the problem into a user clustering followed by a power allocation problem.

In this work, the user clustering and ordering problem is what is relevant and can be given as

$$\min_{\{u_{i,k}\},\{b_k\},\{\pi_k\}} K, \quad (4.9a)$$

s.t. $\sum_{k=1}^{K} u_{i,k} = 1, \forall i = 1, .., N$, \quad (4.9b)

$$b_k \in B_u, \ u_{i,k} = 1, \forall k = 1, .., K, \quad (4.9c)$$

$$d_{\pi_k(j)} \geq j - 1, \ j = 1, .., N_{C_k}, \forall k = 1, .., K, \quad (4.9d)$$

$$u_{i,k} \in \{0,1\}, \quad (4.9e)$$

$$b_k \in B_c, \forall k = 1, .., K, \quad (4.9f)$$

where (4.9b) makes sure a user is placed in exactly one cluster, (4.9c) assures that the beam chosen to serve a cluster of NOMA users belongs to an list contained in the user-beam list of each of those users and (4.9d) is the key constraint that ensures the SIC decoding capability constraints of each user in the system is respected. It is worth noting that for homogeneous systems, constraint (4.9d) can be simplified down to

$$\sum_{i=1}^{N} u_{i,k} \leq d_{\text{max}}, \forall k = 1, .., K, \quad (4.10)$$

since all users have the same decoding capability and hence, only the overall number of users per cluster needs to be limited.

Even after breaking down the original non-convex and combinatorial problem
in (4.8a) into sub-problems, the optimal user clustering solution to the problem in (4.9a) requires an exhaustive search and may not be affordable for practical systems and networks with a large number of users [20]. The objective function defined in (4.9a) represents an NP-complete problem, a class of problems widely known for its complexity in wireless communications and other fields of study. In the literature, several optimization based techniques such as the monotonic optimization approach, combinatorial relaxation approach and matching theory approach have been used to solve the user clustering problem [132]. With our system model that additionally considers individual user decoding capability constraints for each user, NOMA-MEC and NOMA-BB were two heuristics we proposed to solve the user clustering problem in [132]. As shown in Chapter 3, these are low complexity heuristics that scale with the following parameters: 1) the number of beams each user picks in its beam set, \( b \), which is an algorithm specific parameter, 2) the number of candidate beams, \( B \), which is a system level design parameter that we can control and 3) the number of users, \( N \), that need to be served along with their respective decoding capabilities, \( d_u, u \in [1, N] \). If some of these parameters are sufficiently large, it was shown in Chapter 3 that upwards of 100,000 computations would need to be computed at the BS compute resources on a millisecond granularity as discussed earlier in the motivation for this chapter. On the other hand, a fully trained ANN can be directly applied at the network irrespective of how large or small the parameters \( b, B \) or \( N \) are. Of course, the larger these values, the more complex it is to train the neural network but this training can be done offline from the BS on much more powerful servers on a cloud or elsewhere.

NOMA-MEC solves the cluster minimization problem in (4.9a) for a heterogeneous system while NOMA-BB is a lower complexity algorithm for the simpler homogeneous system. While we refer the reader to [132] for the details on the NOMA-MEC and NOMA-BB algorithm, these algorithms are used to create the training data set for the ANN-NOMA-MEC and ANN-NOMA-BB algorithms that are proposed. In other words, by generating datasets using the NOMA-MEC heuristic in a wide variety of simulation settings and knowing its performance in terms of \( R_{sum} \), we establish the ground truth for the proposed artificial neural networks, i.e., the truth with which the neural networks are trained. We note that being heuristics, NOMA-MEC and NOMA-BB are approximations to the real solutions. However, the same neural network can just as easily be trained on a dataset created from the user clustering solution found
from a brute-force style approach as well.

4.3 Proposed Algorithm(s)

In this section, we outline the proposed ANN-NOMA-MEC and ANN-NOMA-BB neural networks in terms of the neural network architecture in Section 4.3.1 and the detailed training and testing steps in Section 4.3.2.

4.3.1 Neural Network Structure

The proposed neural network architecture to search and learn underlying patterns in the datasets containing the user clustering solutions generated from the NOMA-MEC and NOMA-BB schemes is illustrated in Fig. 4.2. The neural network architecture consists of three fully connected layers: the input layer, a hidden layer and the output layer. The first layer is the input layer where the number of neurons is defined by the number of users $N$. Each neuron in this layer receives a data sample $t$ containing a vector of features per user-$u$, $F_{u}^{(t)}$. The relevant features of each user in $F_{u}^{(t)}$ needed to train the neural network are defined by the user-beam set, $B_{u}^{(t)}$, containing the best $b$ beams for user-$u$; the users mmWave channel expressed through the physical departure angle between the BS and the user-$u$, $\theta_{u}^{(t)}$, and the distance between the
BS and the user-\(u\), \(r^{(t)}_{u}\); and finally the SIC decoding capability of user-\(u\), \(d^{(t)}_{u}\). Thus, the \(F^{(t)}_{u}\) user’s feature vector will be a column vector with the following structure:

\[
F^{(t)}_{u} = \left[ B^{(t)}_{u}, \theta^{(t)}_{u}, r^{(t)}_{u}, d^{(t)}_{u} \right]^T. \tag{4.11}
\]

Let \(n_f\) be the number of features contained in the feature vector. Finally, let \(F^{(t)}\) be the matrix of user features presented to the neural network at data sample \(t\) and is defined as

\[
F^{(t)} = \left[ F^{(t)}_{1}, F^{(t)}_{2}, \cdots, F^{(t)}_{N} \right]^T. \tag{4.12}
\]

We can infer from (4.12) that the input to the neural network will be a matrix, whose dimension will be defined by the length of the feature vector, \(n_f\), and the number of users, \(N\), i.e., \(F^{(t)} \in \mathbb{R}^{N \times n_f}\). It is worth noting that \(n_f\) can vary based on the number of beams per user set, \(b\), in \(B_u\). For example, if \(b = 2\) beams in \(B_u\), then \(n_f = 5\) after adding the other features \(\theta_u\), \(r_u\) and \(d_u\). Thus, each configuration of \(N\) and \(b\) will result in a different network size and, therefore, in a different neural network.

The second layer is the hidden layer consisting of \(H\) hidden neurons denoted by \(h \in 1, 2, \cdots, H\). The input feature matrix, \(F^{(t)}\), containing the relevant user information from all input layer neurons \((i \in 1, 2, \cdots, N)\) are connected to each neuron \(h\) present in the hidden layer. The selected hidden layer in our neural network architecture is the powerful long short-term memory (LSTM) layer, which makes each hidden neuron to be modeled as an LSTM unit. Fig. 4.3 illustrates in more detail the internal composition of hidden neurons and their connections with respect to the overall structure of the neural network as described later in this section. The output layer is the final layer and has \(N\) number of neurons denoted by \(o \in 1, 2, \cdots, N\). The outputs from all neurons in the hidden layer are connected to each neuron \(o\), which provides the estimated cluster assignment for the respective user-\(u\), \(\hat{y}_u\).

Our framework proposes a fully connected neural network structure. In the data sample \(t\), the feature vector of each user-\(u\), \(F^{(t)}_{u}\), is received by each input layer neuron. The fully connected layout between the input and hidden layers allows the entire input feature matrix, \(F^{(t)}\), to be processed by each hidden neuron. As shown in Fig. 4.3, hidden neurons are composed of a cell state denoted by \(c^{(t)}\) and four internal components defined by \(z^{(t)}, i^{(t)}, f^{(t)}\) and \(o^{(t)}\). These internal components manage the information flow by combining the input features, \(F^{(t)}\), weighted by the
Figure 4.3: Illustrating the fully connected structure of the hidden neurons in the proposed neural network.

weights $W_z^{(t)}$, $W_i^{(t)}$, $W_f^{(t)}$, and $W_o^{(t)}$, respectively; the hidden neuron outputs of the previous data sample $v_h^{(t-1)}$ weighted by $R_z^{(t)}$, $R_i^{(t)}$, $R_f^{(t)}$, and $R_o^{(t)}$, respectively; and the biases $b_z$, $b_i$, $b_f$, and $b_o$. Particularly, $f^{(t)}$ and $i^{(t)}$ add the cell state of the previous data sample $c^{(t-1)}$ weighted by $p_f^{(t)}$ and $p_i^{(t)}$, respectively; while $o^{(t)}$ adds the current cell state $c^{(t)}$ weighted by $p_o^{(t)}$. Then, the components $f^{(t)}$, $i^{(t)}$, and $o^{(t)}$ are regulated by an activation function defined by $\sigma$ while $z^{(t)}$ is regulated by the activation $\rho$. Finally, the current cell state $c^{(t)}$ is obtained from $c^{(t-1)}$, $f^{(t)}$, $i^{(t)}$ and $z^{(t)}$ while the output $v_h^{(t)}$ is obtained from $o^{(t)}$ and an regulated version of $c^{(t)}$ through the activation function $\omega$. Concretely, the idea behind the LSTM hidden layer is to maintain its state, $c^{(t)}$, over time and regulate the flow of information through nonlinear activation functions ($\sigma$, $\rho$, and $\omega$). Thus, learning in the hidden neurons is given by recursively connecting their cell states, $c^{(t-1)}$, and their outputs, $v_h^{(t-1)}$, with their inputs $F^{(t)}$. The above, gives rise to the STM and LTM concepts where: STM, short-term memory, refers to the learning at each data sample and is guided by the current output, $v_h^{(t)}$; while LTM refers to the learning over time, which is recorded and remembered in the cell state $c^{(t)}$. In this way, the hidden layer can remember relevant information and forget irrelevant information from user features, in favor of learning complex and unpredictable patterns to predict clustering solutions from the NOMA-MEC and NOMA-BB heuristics in the final layer.

The outputs of the hidden layer, $v_h^{(t)} = v_1^{(t)}, v_2^{(t)}, \ldots, v_H^{(t)}$, are then passed
through an activation function defined by $\varphi$, to obtain a regularized version, $h^{(t)} = h_1^{(t)}, h_2^{(t)}, \ldots, h_H^{(t)}$, that feeds the final layer. The full connection between the hidden and output layers allows each neuron in the final layer to process $h^{(t)}$ by a linear combination weighted by $W_y^{(t)}$, whose final result is also regulated by the $\varphi$ activation. In this way, the neural network obtains the estimated clustering assignment of each user-$u$, $\hat{y}_u$, as a result of a series of direct and indirect operations on the user features $F^{(t)}$ where all the weights involved in the neural network structure we have proposed intervene: $W_z^{(t)}, W_i^{(t)}, W_f^{(t)}, W_o^{(t)}, R_z^{(t)}, R_i^{(t)}, R_f^{(t)}, R_o^{(t)}, p_f^{(t)}, p_i^{(t)}, p_o^{(t)}$ and $W_y^{(t)}$. As we will see in the next section, during the training phase, our neural network learns to estimate clustering solutions from the NOMA-MEC and NOMA-BB heuristics based on an iterative adjustment of these weights.

4.3.2 Training and Testing

In this section we outline the training and testing phases for the proposed neural networks that applies to both ANN-NOMA-MEC and ANN-NOM-BB. Algorithm 3 summarizes the training phase in which our neural network will learn to cluster users based on the clustering solutions offered by the NOMA-MEC and NOMA-BB heuristics. Algorithm 4 summarises the testing phase by providing a method for evaluating the neural network as it is fed by user features to obtain the estimated clustering solution.

We begin with the training algorithm where we first need to describe the input defined by the sets $F_\tau$ and $Y_\tau$. Specifically, $F_\tau = \{F^{(1)}, F^{(2)}, \ldots, F^{(T)}\}$ is a set containing the feature matrix $F^{(t)}$ obtained from (4.12) for each data sample $t = 1, 2, \cdots, T$. On the other hand, $Y_\tau = \{y^{(1)}, y^{(2)}, \ldots, y^{(T)}\}$ is a set containing the labeled user clustering vector $y^{(t)}$ for each data sample $t$. Let $y_u$ be the actual cluster label of user-$u$, obtained from the clustering solution of the NOMA-MEC and NOMA-BB heuristics, the labeled vector following the structure $y^{(t)} = [y_1, y_2, \cdots, y_u, \cdots, y_N]$, can be seen as the target clustering formation that the neural network must learn to predict. In other words, each $y^{(t)}$ in $Y_\tau$ will be in charge of guiding the training of the neural network on each training data sample $t$ as we will describe later in this section. As we see, the cluster numbering of each user-$u$ in $y^{(t)}$ plays an important role and it should be labeled in a manner that can be easily learned by the neural
Algorithm 3: ANN Training

Input: $F_T, Y_T$
Parameters: $\alpha, n_{epochs}, H, T_{train}, T_{valid}$
Output: ANN, the Artificial Neural Network model

Step-1: Split $F_T, Y_T$ using $T_{train}, T_{valid}$
- Get training sets $F_{T_{train}}, Y_{T_{train}}$
- Get validation sets $F_{T_{valid}}, Y_{T_{valid}}$

Step-2: Initializations
- Initialize $W_y^{(0)}, W_r^{(0)}, R_r^{(0)}, p_i^{(0)}, p_f^{(0)}, p_o^{(0)}$
- Initialize $y^{(0)}, c(0)$
- Initialize bias $b^*$

Step-3: Learning of the network
for $e = 1$ to $n_{epochs}$ do
  for $t = 1$ to $T_{train}$ do
    Feed-forward computation:
    - Vectorize $F_{T_{train}}^{(t)}$
    - Compute $z(t), i(t), f(t), c(t), o(t)$
    - Compute $v_h(t), h(t)$
    - Compute $v_y(t), y(t)$
    - Compute the loss $L$ using $Y_{T_{train}}^{(t)}$
  Back-propagation and weight adjustment:
    - Compute $\delta_{vy}^{(t)}, \Delta_{yh}^{(t)}, \delta_{vh}^{(t)}$
    - Compute $\delta_{o}^{(t)}, \delta_{c}^{(t)}, \delta_{f}^{(t)}, \delta_{i}^{(t)}, \delta_{o}^{(t)}$
    - Compute $\delta_{W_y}^{(t)}, \delta_{W_r}^{(t)}, \delta_{R_r}^{(t)}, \delta_{p_i}^{(t)}, \delta_{p_f}^{(t)}, \delta_{p_o}^{(t)}$
    - Adjust $W_y^{(t)}, W_r^{(t)}, R_r^{(t)}, p_i^{(t)}, p_f^{(t)}, p_o^{(t)}$
  MSE computation using $Y_{T_{train}}^{(t)}$

Step-4: Validation
for $t = 1$ to $T_{valid}$ do
  Feed-forward computation:
    - Vectorize $F_{T_{valid}}^{(t)}$
    - Compute $z(t), i(t), f(t), c(t), o(t)$
    - Compute $v_h(t), h(t)$
    - Compute $v_y(t), y_{valid}^{(t)}$
  MSE computation using $Y_{T_{valid}}^{(t)}$

network. Therefore, let $C = \{C_1, C_2, \cdots, C_k, \cdots, C_K\}$ be the set of $K$ clusters and let $b_C = \{b_1, b_2, \cdots, b_k, \cdots, b_K\}$ be a set of $K$ beams where each beam $b_k$ represents
Algorithm 4: ANN Testing

Input: ANN, \( F_{\text{Test}} \), \( T_{\text{Test}} \)
Output: \( \hat{Y}_{\text{Test}} = \{ \hat{y}_{\text{Test}}^{(1)}, \ldots, \hat{y}_{\text{Test}}^{(T_{\text{Test}})} \} \), the predicted cluster formation of users for all \( T_{\text{Test}} \) data samples

Step-1: Initializations
- Initialize \( y^{(0)}, c^{(0)} \)

Step-2: Testing
for \( t = 1 \) to \( T_{\text{Test}} \) do
  Feed-forward computation:
  - Vectorize \( F_{\text{Test}}^{(t)} \)
  - Compute \( z^{(t)}, v^{(t)}, f^{(t)}, c^{(t)} \) and \( o^{(t)} \)
  - Compute \( v^{(t)}_h, h^{(t)} \)
  - Compute \( v^{(t)}_y, \hat{y}_{\text{Test}}^{(t)} \)
  Store the predicted \( \hat{y}_{\text{Test}}^{(t)} \):
  - \( \hat{Y}_{\text{Test}} \{ t \} = \hat{y}_{\text{Test}}^{(t)} \)

the most repeated beam among all user-beam sets \( B_u \) involved with users in \( C_k \), then, in \( y^{(t)} \) all users of a specific cluster \( C_k \) are labeled with a single number based the smallest beam in \( b_C \). More explicitly, the users of the cluster containing the smallest beam in \( b_C \) will be labeled as 1, the users of the cluster containing the second smallest beam will be labeled as 2, and so on.

The goal of Algorithm 3 is to train a neural network based on the structure captured in Figs. 4.2 and 4.3 so that it learns to cluster users on its own. To do this, we break down the training algorithm into four steps. In Step-1, we obtain the training and validation sets by splitting the input sets \( F_T \) and \( Y_T \). The training sets denoted by \( F_{\text{Train}} \) and \( Y_{\text{Train}} \) are used to train the neural network and are obtained by selecting \( T_{\text{Train}} \) number of data samples from \( F_T \) and \( Y_T \), respectively. The validation sets denoted by \( F_{\text{Valid}} \) and \( Y_{\text{Valid}} \) are used to assess the performance of the neural network after training and are obtained by selecting \( T_{\text{Valid}} \) number of data samples from \( F_T \) and \( Y_T \), respectively. It should be noted that the data samples for each set can be selected randomly but with no overlap between each set. In Step-2, we initialize all the training variables that require initialization at time step \( t = 0 \) with zeroes or random values. This includes all the weights that compose the network, the output and cell state of the hidden layer denoted by \( v^{(0)}_h \) and \( c^{(0)} \), respectively as well as the biases involved in the four components of the hidden layer.
In Step-3, we perform the learning of the network based on the user features and the labeled user clustering formation of all data samples \( t = 1, 2, \ldots, T_{\text{train}} \) contained in the training sets \( F_{T_{\text{train}}} \) and \( Y_{T_{\text{train}}} \), respectively. Let \( e \) be a training epoch in which the network is learning to estimate the desired user clustering formations in \( Y_{T_{\text{train}}} \) from the corresponding user features in \( F_{T_{\text{train}}} \), the training algorithm needs to go through several epochs until it ensures that the estimated clustering formation of the network is close enough to the desired ones. The learning at each epoch \( e \) is performed by iterating through the sets \( F_{T_{\text{train}}} \) and \( Y_{T_{\text{train}}} \) where each iterative step is defined by the data sample \( t \). Let \( F_{T_{\text{train}}}^{(t)} \) and \( Y_{T_{\text{train}}}^{(t)} \) be the user feature matrix and the labeled clustering vector of the data sample \( t \) from \( F_{T_{\text{train}}} \) and \( Y_{T_{\text{train}}} \), respectively. We then break down the learning in two stages. In the first stage the feed-forward computation of the neural network is performed starting from the feature matrix \( F_{T_{\text{train}}}^{(t)} \) to finish computing the network loss \( \mathcal{L} \) using the labeled clustering vector \( Y_{T_{\text{train}}}^{(t)} \). Then, in the second stage the back-propagation is carried out to adjust all the weights of the neural network where its learning is reflected.

The feed-forward computation begins by vectorizing the user feature matrix \( F_{T_{\text{train}}}^{(t)} \) to convert it into a column vector. The vectorization of the \( F_{T_{\text{train}}}^{(t)} \) matrix is obtained by stacking all its columns on top of one another. Let \( F_{v}^{(t)} \) be the vectorized version of \( F_{T_{\text{train}}}^{(t)} \) defined as,

\[
F_{v}^{(t)} = \text{vec} \left( F_{T_{\text{train}}}^{(t)} \right) = \left[ B_{1}^{(t)}, \ldots, B_{N}^{(t)}, \theta_{1}^{(t)}, \ldots, \theta_{N}^{(t)}, r_{1}^{(t)}, \ldots, r_{N}^{(t)}, d_{1}^{(t)}, \ldots, d_{N}^{(t)} \right]^T. \tag{4.13}
\]

We can infer from (4.13) that \( F_{v}^{(t)} \in \mathbb{R}^{N_{f} \times 1} \) is a column vector containing the features of all users and whose length \( N_{f} \) is defined by the number of elements in \( F_{T_{\text{train}}}^{(t)} \), i.e., \( N_{f} = N \cdot n_{f} \).

Next, the feed-forward uses the vectorized version \( F_{v}^{(t)} \) to compute the four internal components, \( \{ z, i, f, o \} \), along with the cell state of the hidden layer as outlined in
4.3, as follows [145]:

\[
\begin{align*}
z^{(t)} &= \rho \left( W_z^{(t)} F_v^{(t)} + R_z^{(t)} v_h^{(t-1)} + b_z \right), \\
i^{(t)} &= \sigma \left( W_i^{(t)} F_v^{(t)} + R_i^{(t)} v_h^{(t-1)} + p_i^{(t)} \odot c^{(t-1)} + b_i \right), \\
f^{(t)} &= \sigma \left( W_f^{(t)} F_v^{(t)} + R_f^{(t)} v_h^{(t-1)} + p_f^{(t)} \odot c^{(t-1)} + b_f \right), \\
c^{(t)} &= z^{(t)} \odot i^{(t)} + c^{(t-1)} \odot f^{(t)}, \\
o^{(t)} &= \sigma \left( W_o^{(t)} F_v^{(t)} + R_o^{(t)} v_h^{(t-1)} + p_o^{(t)} \odot c^{(t)} + b_o \right).
\end{align*}
\]

Let \(*\) refer to any of the internal components of the hidden layer \( \{z, i, f, o\} \), then \( W_*^{(t)} \in \mathbb{R}^{H \times N_t}, R_*^{(t)} \in \mathbb{R}^{H \times H}, b_* \in \mathbb{R}^{H \times 1}, *^{(t)} \in \mathbb{R}^{H \times 1} \). The operator \( \odot \) denotes the pointwise multiplication between two vectors that must necessarily have the same length. For example, in expression (4.17) the operation \( z^{(t)} \odot i^{(t)} \) is performed, for which the block input \( z^{(t)} \in \mathbb{R}^{H \times 1} \) and the input gate \( i^{(t)} \in \mathbb{R}^{H \times 1} \) are two column vectors of length \( H \). The result of this operation is another column vector of length \( H \) in which each element reflects the multiplication of the corresponding elements in \( z^{(t)} \) and \( i^{(t)} \). Thus, we can deduce that the cell state \( c^{(t)} \in \mathbb{R}^{H \times 1} \) in which each element is associated to each hidden neuron \( h = 1, \ldots, H \). It is worth highlighting the recurrence of the second term in the definition of the four internal components of the hidden layer, \( v_h^{(t-1)} \in \mathbb{R}^{H \times 1} \), that represents the output of the hidden layer in a previous iterative step. During the first iteration, \( t = 1 \), \( v_h^{(0)} \) which was initialized in Step-2 is applied.

Then, both the output of the hidden layer \( v_h^{(t)} \) and its regularised version \( h^{(t)} \) are computed by applying the following expressions,

\[
\begin{align*}
v_h^{(t)} &= \omega \left( c^{(t)} \right) \odot o^{(t)}, \\
h^{(t)} &= \varphi \left( v_h^{(t)} \right).
\end{align*}
\]

We can infer from (4.19) that \( v_h^{(t)} \in \mathbb{R}^{H \times 1} \) is a vector of length \( H \) in which each element represents the output of each hidden neuron \( h = 1, \ldots, H \). While \( h^{(t)} \) is simply the result of passing \( v_h^{(t)} \) through the activation function \( \varphi \). Similar to [137], the activation \( \varphi \) that we apply here is the ReLu function for which each element in \( v_h^{(t)} \) is regulated depending on its sign. Elements with positive sign are treated linearly, while elements with negative sign are set to be zero. In this way, \( h^{(t)} \in \mathbb{R}^{H \times 1} \) results
in a column vector of length $H$ where each regulated element is associated to each neuron $h = 1, 2, \cdots, H$.

Subsequently, both the output of the final layer $v_y(t)$ and the estimated user clustering formation $\hat{y}(t)$ are computed following the expressions,

$$v_y(t) = W_y h(t),$$

(4.21)

$$\hat{y}(t) = \varphi(v_y(t)).$$

(4.22)

In (4.21), $W_y(t) \in \mathbb{R}^{N \times H}$ represents a weight matrix containing all the weights associated with the output layer. Specifically, each row in $W_y(t)$ is associated with each output neuron $o = 1, 2, \cdots, N$. Since each of these neurons is fed with the vector $h(t)$, the matrix multiplication defined in (4.21) results in the vector $v_y(t) \in \mathbb{R}^{N \times 1}$ in which each element is related to a neuron $o$. Similar to (4.20), in (4.22), $v_y(t)$ is passed through the ReLu activation function to obtain the estimated user clustering formation $\hat{y}(t) \in \mathbb{R}^{N \times 1}$ as a vector of length $N$, in which each element represents the estimated cluster assignment of the user-$u$, $\hat{y}_u(t)$, for each $u = 1, 2, \cdots, N$.

Finally, feed-forward stage computes the network loss denoted by $L$ in which the estimated clustering formation, $\hat{y}(t)$, is compared with the labeled clustering formation $Y_{train}^{(t)}$ by the following expression,

$$L = \frac{1}{N} \sum_{u=1}^{N} (y_u - \hat{y}_u)^2.$$  

(4.23)

The back-propagation stage seeks to adjust the learnable parameters involved in the network based on the minimization of the network loss, $L$, obtained in the previous stage. As pointed out in [145], at each iterative step $t$, the back-propagation begins by applying the generalized delta rule to get the delta vectors $\delta_{v_y}^{(t)}$ and $\Delta_{v_h}^{(t)}$ that enable the update of weights for the output and hidden layers, respectively, and can be expressed as

$$\delta_{v_y}^{(t)} = \varphi'(v_y(t)) \odot e_y(t),$$

(4.24)

$$\Delta_{v_h}^{(t)} = \varphi'(v_h(t)) \odot e_h(t),$$

(4.25)

where $\varphi'$ represents the derivative of the ReLu activation function upon $v_y(t)$ and $v_h(t)$. On the other hand, $e_y(t)$ represents the derivative of the network loss while $e_h(t) = (W_y)^T \delta_{v_y}^{(t)}$. Next, we compute the accumulated gradient vector $\delta_{v_h}^{(t)}$, which can
be defined as the combination between $\Delta_{vh}^{(t)}$ and the recurrent dependencies according to the expression,

$$
\delta_{vh}^{(t)} = \Delta_{vh}^{(t)} + R_z^{(t-1)} \delta_{\hat{z}}^{(t+1)} + R_i^{(t-1)} \delta_i^{(t+1)} + R_f^{(t-1)} \delta_f^{(t+1)} + R_o^{(t-1)} \delta_o^{(t+1)},
$$

(4.26)

such that $R_s^{(t-1)}$ (where $*$ can be $z, i, f, o$) represent recurrent weights that were adjusted at the previous iterative step; while $\delta_{s}^{(t+1)}$ represent the gradient vectors associated with the four components of the hidden layer at $t + 1$. Subsequently, we compute the delta vectors $\delta_{o}^{(t)}$, $\delta_{c}^{(t)}$, $\delta_{f}^{(t)}$, $\delta_{i}^{(t)}$, and $\delta_{\hat{z}}^{(t)}$ according to,

$$
\delta_{o}^{(t)} = \delta_{vh}^{(t)} \odot \omega \left( c^{(t)} \right) \odot \sigma' \left( o^{(t)} \right),
$$

$$
\delta_{c}^{(t)} = \delta_{vh}^{(t)} \odot o^{(t)} \odot \omega' \left( c^{(t)} \right) + p_o \odot \delta_{o}^{(t)} + p_i \odot \delta_{i}^{(t+1)} + p_f \odot \delta_{f}^{(t+1)} + \delta_{c}^{(t+1)} \odot f^{(t+1)},
$$

$$
\delta_{f}^{(t)} = \delta_{c}^{(t)} \odot c^{(t-1)} \odot \sigma' \left( \hat{f}^{(t)} \right),
$$

$$
\delta_{i}^{(t)} = \delta_{c}^{(t)} \odot z^{(t)} \odot \sigma' \left( \hat{i}^{(t)} \right),
$$

$$
\delta_{\hat{z}}^{(t)} = \delta_{c}^{(t)} \odot \hat{i}^{(t)} \odot \rho' \left( \hat{z}^{(t)} \right),
$$

where $\hat{i}^{(t)}$, $\hat{f}^{(t)}$ and $\hat{o}^{(t)}$ are vectors defined from (4.15), (4.16) and (4.18), respectively, but without applying the activation function $\sigma$; $\hat{z}^{(t)}$ is a vector defined from (4.14) but without applying the activation function $\rho$; while $\rho'$, $\omega'$ and $\sigma'$ are the derivatives of the activation functions tanh and sigmoid, respectively.

Based on the computed delta vectors, the gradients $\delta_{W_y}^{(t)}$, $\delta_{W_z}^{(t)}$, $\delta_{R_z}^{(t)}$, $\delta_{p_i}^{(t)}$, $\delta_{p_f}^{(t)}$ and $\delta_{p_o}^{(t)}$ used to adjust the weights are calculated as follows:

$$
\delta_{W_y}^{(t)} = \delta_{v_y}^{(t)} \otimes h^{(t)}, \quad \delta_{p_i}^{(t)} = c^{(t)} \odot \delta_{i}^{(t+1)},
$$

$$
\delta_{W_z}^{(t)} = \delta_{*}^{(t)} \otimes F_v^{(t)}, \quad \delta_{p_f}^{(t)} = c^{(t)} \odot \delta_{f}^{(t+1)},
$$

$$
\delta_{R_z}^{(t)} = \delta_{*}^{(t+1)} \otimes v_h^{(t)}, \quad \delta_{p_o}^{(t)} = c^{(t)} \odot \delta_{o}^{(t)},
$$

with $*$ being any of the four internal components in the hidden layer $\{z, i, f, o\}$ and the operator $\otimes$ representing the outer product between two vectors. Unlike pointwise multiplication, the outer product $\otimes$ performs a matrix computation in which the second vector is transposed. For example, if we wish to compute the gradient $\delta_{W_y}^{(t)} = \delta_{v_y}^{(t)} \otimes h^{(t)}$, we first transposes $h^{(t)}$ to become $h^{(t)^T}$ and then performs the outer product. This is because the outer product is defined as $A \otimes B = [a_{ij}b_{jk}]$, where $A$ and $B$ are matrices.


\[ \delta_f^{(t)} \otimes F_v^{(t)} = \delta_f^{(t)} [F_v^{(t)}]^T, \] for which if \( \delta_f^{(t)} \in \mathbb{R}^{H \times 1} \) and \( [F_v^{(t)}]^T \in \mathbb{R}^{1 \times N_f} \) the result of this operation is the matrix \( \delta_f W_f^{(t)} \in \mathbb{R}^{H \times N_f} \).

Finally, similar to the work in [137], we apply the stochastic gradient descent (SGD) scheme in order to perform the adjustment of all the weights associated with our neural network. Mathematically, the adjustment of the weights can be achieved by following the expressions,

\[
\begin{align*}
W_y^{(t)} &= W_y^{(t-1)} + \alpha \delta_{W_y}^{(t)}, \\
W_*^{(t)} &= W_*^{(t-1)} + \alpha \delta_{W_*}^{(t)}, \\
R_*^{(t)} &= R_*^{(t-1)} + \alpha \delta_{R_*}^{(t)}, \\
p_i^{(t)} &= p_i^{(t-1)} + \alpha \delta_{p_i}^{(t)}, \\
p_f^{(t)} &= p_f^{(t-1)} + \alpha \delta_{p_f}^{(t)}, \\
p_o^{(t)} &= p_o^{(t-1)} + \alpha \delta_{p_o}^{(t)}, \\
R_*^{(t)} &= R_*^{(t-1)} + \alpha \delta_{R_*}^{(t)}, \\
p_i^{(t)} &= p_i^{(t-1)} + \alpha \delta_{p_i}^{(t)}, \\
p_f^{(t)} &= p_f^{(t-1)} + \alpha \delta_{p_f}^{(t)}, \\
p_o^{(t)} &= p_o^{(t-1)} + \alpha \delta_{p_o}^{(t)}, \\
\end{align*}
\]

where \( \alpha \) is a high-impact scalar parameter in the training of the network which defines the learning rate that regulates how much the weights are adjusted at each iterative step. At the end of step-3, the mean square error (MSE) metric is computed which measures the performance of the network at the end of each epoch and is defined as,

\[
\text{MSE} = \frac{1}{T_{\text{train}}} \sum_{t=1}^{T_{\text{train}}} \left( Y_{\text{train}}^{(t)} - \hat{Y}_{\text{train}}^{(t)} \right)^2. \tag{4.27}
\]

In Step-4 of Algorithm 3, the performance is assessed at the end of each epoch \( e \) using the validation sets, \( F_{\text{valid}}^{(t)} \) and \( Y_{\text{valid}}^{(t)} \), whose data samples have not been used during learning and, therefore, helps measure the generalization capability of the network. To do this, we iteratively go through \( F_{\text{valid}}^{(t)} \) and \( Y_{\text{valid}}^{(t)} \) running the feed-forward stage described in Step-3 but using \( F_{\text{valid}}^{(t)} \) at each iterative step in order to obtain the corresponding estimated clustering formation \( \hat{y}_{\text{valid}}^{(t)} \). Finally, we compute the MSE metric using \( y_{\text{valid}}^{(t)} \) and \( Y_{\text{valid}}^{(t)} \) to measure the validation performance. If the validation MSE is relatively high, then the learning needs to go through more epochs until the validation MSE reaches an acceptable performance.

The training algorithm is highly influenced by the parameters \( T_{\text{train}}, \alpha, H \) and \( n_{\text{epochs}} \). In particular, \( T_{\text{train}} \) defines the number of training data samples selected from \( F_T \) and \( Y_T \) that are used for network learning. If \( T_{\text{train}} \) is too small the learning of the network in Step-3 can be highly compromised because the weights would be adjusted for a small set of solutions from the NOMA-MEC and NOMA-BB heuristics, as the case may be. This would reduce the generalization capability of the network, resulting in a low validation MSE. The learning rate \( \alpha \) controls how much the network weights
are adjusted during Step-3 at each iterative step $t$. In other words, this parameter defines how fast the network learns. Considering that $\alpha$ can take values between 0 and 1, a desired value of $\alpha$ should be low enough for the network to converge to something useful, but high enough that it can be trained in a reasonable time. In this sense, a smaller value of $\alpha$ requires more training epochs since only small changes are made to the weights at each adjustment; while a large value of $\alpha$ results in rapid changes and requires fewer training epochs but can lead to a sub-optimal set of weights. On the other hand, the design parameter $H$ directly influences the complexity of the training algorithm. A large value of $H$ leads to an exaggerated number of neurons in the hidden layer, making the neural network structure very large and therefore very slow. However, if $H$ is too small, it reduces the ability of the hidden layer to learn to remember desired patterns as well as to forget undesired patterns. Finally, the parameter $n_{\text{epochs}}$ which defines the number of times the learning works through the complete $F_{\tau_{\text{train}}}$ and $Y_{\tau_{\text{train}}}$ sets, is a critical parameter as it ensures the convergence of the network to an acceptable performance. A desirable value of $n_{\text{epochs}}$ should be large enough so that the train MSE is low enough while the validation MSE is not affected. Therefore, the design of the neural network relies on an appropriate choice of these four parameters in order to find an optimal combination that leads to a high performing network that can be trained in an acceptable amount of time. In Section 6.5, we illustrate the tuning process of these four parameters for the proposed neural netowrk.

In the testing algorithm described in Algorithm 4, the input is defined by the artificial neural network ANN and the testing set $F_{\tau_{\text{test}}}$. The ANN is a network that has been trained through Algorithm 3 for which all weights involved have been optimally adjusted. The $F_{\tau_{\text{test}}}$ set containing $\tau_{\text{test}}$ feature matrices from $F_{\tau}$ is used to test the ANN in order to obtain the corresponding estimated clustering formations returned in $\hat{Y}_{\tau_{\text{test}}}$. To do this, we break down Algorithm 4 into two steps. In Step-1 we simply need to initialize the output and cell state of the hidden layer denoted by $y^{(0)}$ and $c^{(0)}$, respectively, to consider that these have initial values when $t = 0$. The initial values can be zeros. In Step-2, we perform ANN testing by iterating through $F_{\tau_{\text{test}}}$ where, at each iterative step $t = 1, 2, \ldots, \tau_{\text{test}}$, the feed-forward stage described in Algorithm 3 is performed based on the feature matrix $F_{\tau_{\text{test}}}^{(t)}$ from $F_{\tau_{\text{test}}}$ to obtain the estimated clustering formation $\hat{y}_{\tau_{\text{test}}}^{(t)}$. Then each $\hat{y}_{\tau_{\text{test}}}^{(t)}$ is orderly stored in the set $\hat{Y}_{\tau_{\text{test}}}$ to ensure that the output of the testing algorithm contains all the test solutions
4.4 Simulation Results and Discussion

In this section, we illustrate simulation results to show the process of fine-tuning the neural network training algorithm parameters and then compare the performance of the network against the heuristics from which it was trained. We work with a total dataset of $T = 12000$ samples obtained by running the NOMA-MEC and NOMA-BB heuristics and unless otherwise stated, the number of training samples used is 70% of the data set, i.e., $T_{\text{train}} = 8400$. Table 4.2 shows the simulation settings used in this section. For the simulations related to the tuning of neural network parameters, we present the ANN-NOMA-MEC case as it is the more complicated network given that it supports heterogeneous systems where each user has its own individual SIC decoding capability presented to the network to factor into the clustering algorithm. By going through different controlled settings of hidden neurons, $H$, learning rate, $\alpha$, training data samples, $T_{\text{train}}$ and number of epochs, $n_{\text{epochs}}$, we train different neural networks with the objective of choosing the combination of parameters that optimizes the validation performance of the ANNs. In particular, to measure the validation performance, we use both the MSE metric and the effective sum rate $R_{\text{sum}}$ [132] in the simulation runs. The MSE metric introduced in Section 3.3, is chosen because it is commonly used to measure the performance of an ANN model, while $R_{\text{sum}}$ provides us with a performance measure in terms of the user clustering formation predicted by the ANN. We also run simulations for different network sizes, i.e., using different configurations of number of users $N$ and number of best $b$ beams per user, to demonstrate the stability of the proposed neural network structure in ANN-NOMA-MEC and ANN-NOMA-BB, respectively. Finally, the effective sum rate for our optimally trained ANN-NOMA-MEC and ANN-NOMA-BB networks are benchmarked with the original NOMA-MEC and NOMA-BB heuristics, respectively.

We start by investigating the effect of the number of hidden neurons, $H$, on the performance of the proposed neural network as the number of users in the network grows. In Fig. 4.4, we run simulations for different choices of $H$ for an ANN-NOMA-MEC neural network trained on a dataset of 8400 samples obtained from NOMA-MEC heuristic with $b = 2$. As seen in Fig. 4.4, in general, an ANN-NOMA-MEC setting with $H = 1$, i.e., a single hidden neuron, leads to the highest MSE which
implies the lowest performance, because the clustering formations predicted by the network are very poor and are further away from the clustering formations solved by the NOMA-MEC heuristic. At $H = 10$, we get good performance at 50 users but the performance deteriorates as the number of users is further increased. In general, when $H = 50$ we observe the best MSE performance for all users. However, when we go to even higher number of hidden layer neurons, the neural network starts to overfit to the training data and the performance deteriorates when tested on the validation dataset. Therefore, we use $H = 50$ hidden neurons in the further simulations.

We now move to study the impact of the learning rate, $\alpha$, on the performance of ANN-NOMA-MEC. To do this, in Fig. 4.5, we compare both the MSE performance (left axis) and the effective sum rate $R_{\text{sum}}$ (right axis) measured by the training algorithm when using learning rates of 0.001 and 0.01, while keeping the other training parameters fixed with $H = 50$ and $b = 2$. It is worth noting that a lower MSE and a higher value of $R_{\text{sum}}$ are indicators of better performance. It is clear from the results in Fig. 4.5a and Fig. 4.5b, which are presented at $n_{\text{epochs}} = 1$ and $n_{\text{epochs}} = 10$, respectively, that at $\alpha = 0.01$, the training algorithm finds it easier to control the learning of the ANN-NOMA-MEC network, which leads to better MSE performances.

### Table 4.2: Simulation parameters

<table>
<thead>
<tr>
<th>Parameter name, notation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users, $N$</td>
<td>$[50, 100, 150, 200]$</td>
</tr>
<tr>
<td>Number of best beams, $b$</td>
<td>$[1, 2, 3, 4]$</td>
</tr>
<tr>
<td>Number of hidden neurons, $H$</td>
<td>$[1, 10, 50, 100, 150, 200]$</td>
</tr>
<tr>
<td>Learning Rate, $\alpha$</td>
<td>$[0.01, 0.001]$</td>
</tr>
<tr>
<td>Training samples, $T_{\text{train}}$</td>
<td>$[6000, 7200, 8400]$</td>
</tr>
<tr>
<td>Validation samples, $T_{\text{valid}}$</td>
<td>1800</td>
</tr>
<tr>
<td>Testing samples, $T_{\text{test}}$</td>
<td>1800</td>
</tr>
<tr>
<td>Number of epochs, $n_{\text{epochs}}$</td>
<td>20</td>
</tr>
</tbody>
</table>
Figure 4.4: Effect of the number of neurons in the hidden layer, $H$, on the performance of ANN-NOMA-MEC for different numbers of users, $N$, with $b = 2$.

that translate into better estimates of clustering formations and, therefore, this leads to higher effective sum rates. Hence, we use $\alpha = 0.01$ as the learning rate in the simulations that follow.

Through the simulations in Fig. 4.6 and Fig. 4.7, we illustrate the impact of the parameters $n_{epochs}$ and $T_{train}$ on the performance of ANN-NOMA-MEC during learning. In Fig. 4.6, we can observe that setting the training algorithm with $n_{epochs} = 1$ is inadequate for any value of $T_{train}$, which implies that the learning must go through more than one epoch for the ANN to better adjust its weights in order to achieve better estimates of clustering formations. The results between $n_{epochs} = 10$ and $n_{epochs} = 20$ are comparable for all values of $T_{train}$, which suggests that the network has reached its learning peak at around the tenth epoch. At $n_{epochs} = 10$, we see that with $T_{train} = 8400$, i.e., 70% of the total samples, we achieve the highest performance. We illustrate this further with the simulation results in Fig. 4.7, where we compare both the MSE performance (left axis) and the effective sum rate $R_{sum}$ (right axis) measured by the training algorithm when considering 6400, 7200 and 8400 training data samples, while fixing the other parameters according to $H = 50$, $\alpha = 0.01$ and $n_{epochs} = 10$. From Fig. 4.7, it is clear that $T_{train} = 6400$ is an insufficient number of samples for the network to train with for optimal performance. At $T_{train} = 8400$ training samples, we find the best network performances, with the lowest validation MSEs, leading to the highest $R_{sum}$. Therefore, $T_{train} = 8400$ training samples are
used in the next set of simulations.

In this way, the parameters of ANN-NOMA-MEC and ANN-NOMA-BB can be fine-tuned and for further simulations, we set the training algorithm to work with $H = 50$ hidden neurons, a learning rate $\alpha = 0.01$, $T_{\text{train}} = 8400$ training data samples and $n_{\text{epochs}} = 10$ training epochs. Using this combination of parameters, we
can train ANN-NOMA-MEC and ANN-NOMA-BB offline from a dataset created by the NOMA-MEC and NOMA-BB parameters, respectively, and then apply them to make clustering decisions in a live network. In Fig. 4.8, we illustrate the evolution of both MSE training and MSE validation through 20 epochs for an ANN-NOMA-MEC network trained in offline mode focused on a system of $N=200$ users and considering $b=2$ best beams. As the figure shows, there is an aggressive reduction of the training MSE through the first 7 epochs where the ANN progressively adjusts its clustering formation estimation to the desired solutions. In parallel, a similar behavior can be observed with the MSE validation also, but with a less aggressive reduction rate. From epoch $e=10$, both training MSE and validation MSE converge to an approximately constant performance measure that is kept until training ends. This confirms the hypothesis that running the ANNs offline up to epoch-10 allows the ANN-NOMA-MEC network to effectively learn to predict the clustering formations of the heterogeneous system at hand and, therefore, this ANN is suitable to be applied directly in live networks.

Finally, using the neural network parameters as described earlier, we run the ANN-NOMA-MEC for different settings of number of users, $N$ and choices of number of best beams per user set, $b$, and compare it against the clustering outputs provided by the NOMA-MEC heuristic. Later, we do the same for ANN-NOMA-BB and compare
it against NOMA-BB for homogeneous systems. In Fig. 4.9a and 4.9b, we show the effective sum rate $R_{sum}$ and the number of clusters $K$, respectively, obtained from the estimates of clustering formations returned by the ANN-NOMA-MEC testing algorithm, considering $T_{test} = 1800$ testing data samples. We plot that against the performance achieved by the clustering results from the baseline heuristic, NOMA-MEC. In general, based on Fig. 4.9a, we can see that the performance $R_{sum}$ is quite stable between the different configurations of $N$ and $b$, which suggests us that the networks have learned similar patterns from the NOMA-MEC heuristic. Although in the cases where $b = 2$ and $b = 3$ the spectral efficiency of the networks is visibly lower with respect to the original heuristic, the plot scale reflects that these values are very close. Therefore, we can infer that our ANN-NOMA-MEC models are able to attain the near-optimal $R_{sum}$ performance as compared to NOMA-MEC heuristic. Furthermore, as seen in Fig. 4.9b, the original behavior of the NOMA-MEC heuristic is such that for each number of users in the heterogeneous system, the number of clusters required to serve the $N$ users, i.e., $K$, decreases as $b$ increases. This same behavior is followed by the ANN-NOMA-MEC models but showing slightly larger values. In any case, through Figs. 4.9a and 4.9b, we can assert that the ANN-NOMA-MEC scheme we have proposed is shown to be effective enough to closely follow the behavior of the NOMA-MEC heuristic. Similarly, in Fig. 4.10, we evaluate the homogeneous case by comparing the performance of ANN-NOMA-BB with its

**Figure 4.7:** Analyzing the effect of the number of training data samples $T_{train}$ on MSE and $R_{sum}$ considering different numbers of users, $N$. ANN-NOMA-MEC with $b = 2$. 
Figure 4.8: Evolution of ANN-NOMA-MEC learning through 20 epochs with $N = 200$ and $b = 2$. Training in offline mode.

respective baseline heuristic, NOMA-BB. Unlike the heterogeneous case, the NOMA-BB heuristic is designed to always work with $b = 1$ while solving homogeneous systems where the SIC decoding capability $d_{\text{max}}$ is an input parameter. This allows us to evaluate ANN-NOMA-BB models through different configurations of $d_{\text{max}}$ as reflected in Fig. 4.10. Once again, we see that the performance of ANN-NOMA-BB closely follows the NOMA-BB heuristic it learned from both in terms of $R_{\text{sum}}$ and number of clusters predicted, $K$. Since NOMA-BB operates on a homogeneous system which is simpler in terms of the underlying patterns of the data, ANN-NOMA-BB is able to more closely match or even slightly exceed the $R_{\text{sum}}$ that NOMA-BB achieves; unlike ANN-NOMA-MEC which has to learn much more complex data patterns in heterogeneous systems and so we saw a slight decrease in $R_{\text{sum}}$ compared to its baseline heuristic of NOMA-MEC.

The results from Fig. 4.9 and Fig. 4.10 show that the neural networks that were trained offline for both homogeneous and the more complex heterogeneous systems can achieve performance on par with the baseline heuristics it was trained on when applied in live networks. While the training and fine tuning of the neural network parameters is a tedious process, this can be done offline and off the BS compute resources, e.g., on cloud resources. When the trained ANNs are applied directly for NOMA clustering in live networks by the BS compute resources, it is a low complexity algorithm. This compared to the baseline heuristics of NOMA-MEC and NOMA-BB that execute thousands of operations to find the optimal cluster assignment on a millisecond granularity. Hence, the proposed ANN-NOMA-MEC and ANN-NOMA-BB
Figure 4.9: Evaluation of different ANN-NOMA-MEC models trained offline with different configurations of $N$ and $b$, benchmarked with the NOMA-MEC heuristic. (a) Performance $R_{\text{sum}}$ and (b) Number of user clusters, $K$. 
Figure 4.10: Evaluation of different ANN-NOMA-BB models trained offline with different number of users, $N$, benchmarked with the NOMA-BB heuristic for different values of the SIC decoding capability, $d_{\text{max}}$. (a) Performance $R_{\text{sum}}$ and (b) Number of user clusters, $K$. 
allow for practical realizations of NOMA in mmWave systems that need to consider a mix of users with different SIC decoding capability constraints.

Finally, we show that the proposed approach generalizes well to datasets other than one prepared using the NOMA-MEC heuristic. To do this, we create a dataset for a heterogeneous system with $b = 2$ using a brute force search (BFS) of all possible clustering options as opposed to one obtained using the NOMA-MEC heuristic. We then train the same ANN on this dataset and term this as ANN-NOMA-BFS to indicate that it has been trained on the BFS dataset. The results are shown in Fig. 4.11 where it is firstly worth noting that BFS performs marginally better than the NOMA-MEC heuristic, as expected. The results in Fig. 4.11 also show that the ANN-NOMA-MEC performs nearly on par as the ANN-NOMA-BFS, indicating there was not much performance lost from using the dataset generated using the NOMA-MEC heuristic as opposed to a full brute force search.

4.5 Conclusion

In this chapter, we proposed a neural network aided machine learning approach to the user clustering and ordering problem in mmWave-NOMA systems that can factor in the individual SIC decoding capability of each user in the system. The proposed neural networks, called ANN-NOMA-MEC and ANN-NOMA-BB, are trained offline on datasets generated from running simulated settings using the NOMA-MEC and
By training the neural network offline using cloud computing resources, the heavy computational steps are executed away from the BS compute resources. Instead, in the live network, where clustering decisions have to be made at a millisecond granularity, the trained neural network can be directly applied to output a clustering result when provided each user’s mmWave channel and SIC decoding capability as input. Simulation results show the effectiveness of the ANN-NOMA-MEC and ANN-NOMA-BB schemes as the neural network trained on offline simulation data performs comparably with the NOMA-MEC and NOMA-BB heuristics that is applying computationally intensive algorithms to make every clustering decision in a live network.

This research can be extended to consider user clustering problems in digital beamforming systems where multiple clusters can be formed and served at the same time, leading to inter-cluster interference that needs to be mitigated. With a large and diverse enough dataset, it would be interesting to learn if artificial neural networks can also learn the underlying patterns involved in cancelling out the inter-cluster interference on top of the SIC decoding capability constraints as considered in this chapter. Since ANN-NOMA-MEC relies on instantaneous channel conditions, another future direction is to consider the impacts of imperfect CSI on the performance. Thus, one extension of this work is to consider non-CSI based feedback, e.g., location based feedback, in the training of the neural network for making clustering decisions.
Chapter 5

Vision-Assisted User Clustering for Robust mmWave-NOMA

When operated in the mmWave band, user channels get highly correlated which can be exploited in mmWave-NOMA systems to cluster a set of “correlated” users together. Identifying the set of users to cluster greatly affects the viability of NOMA systems. Typically, only channel state information (CSI) is used to make these clustering decisions. When any problem arises in accessing up-to-date and accurate CSI, user clustering will not properly function due to its hard-dependency on CSI, and obviously, this will negatively affect the robustness of the NOMA systems. To improve the robustness of the NOMA systems, we propose to utilize emerging trends such as location-aware and camera-equipped base stations (CBSs) which do not require any extra radio frequency resource consumption. Specifically, we explore three different dimensions of feedback that a CBS can benefit from to solve the user clustering problem, namely CSI-based feedback and non-CSI-based feedback, comprised of user equipment (UE) location and the CBS camera feed. We first investigate how the vision assistance of a CBS can be used in conjunction with other dimensions of feedback to make clustering decisions in various scenarios. Later, we provide a simple use case study to illustrate how to implement vision-assisted user clustering in mmWave-NOMA systems to improve robustness, in which a deep learning (DL) beam selection algorithm is trained on the images captured by the CBS to perform NOMA clustering. We demonstrate that user clustering without CSI can achieve comparable performance to accurate CSI-based solutions like the NOMA-BB scheme proposed in Chapter 3, and user clustering can continue to function without much performance loss even in the scenarios where CSI is severely outdated or not available at all. This
work implements a version of the proposal we presented in the patent [146].

5.1 Introduction

Non-orthogonal multiple access (NOMA) techniques offer a way to serve multiple users in the same orthogonal resource (e.g., time, frequency, orthogonal frequency division multiplexing resource block (RB), etc.) by separating the users in the power or code domains instead. In mmWave bands, users’ channels are strongly correlated due to the highly directional nature of mmWave transmission [119, 120]. The strong correlations among users’ channels in mmWave and higher bands make them ideal for the formation of user clusters that can be served by a single beam and separated in the power or code domain through NOMA.

As discussed in previous chapters, for mmWave-NOMA, the method of clustering users is an important aspect to achieving the desired level of network performance. Obtaining the global solution for this clustering problem, particularly for large-size networks, is a formidable task due to its combinatorial nature; but local solutions could be efficiently obtained, for instance, using optimization schemes (e.g., [120,132]) or machine learning (ML) approaches [12,123]. However, these approaches are all based on the strong assumption of accurate instantaneous channel state information (CSI) from the users. Firstly, since clustering is an aspect of user scheduling, executing clustering algorithms based on instantaneous CSI is problematic, as the CSI acquired might be stale by the time it is used for user clustering. Secondly, for mmWave-NOMA systems, CSI acquisition and user tracking, which are required to establish and support highly directional transmission links, can create a tremendous amount of overhead and latency [147]. Thirdly, in practice, the availability and the reliability of CSI at base stations cannot be always guaranteed due to the following reasons: 1) Non-ideal hardware behavior. 2) Poor performance of physical downlink control channel. 3) Poor performance of physical uplink shared channel. 4) Errors in reported CSI. 5) Errors in decoded CSI. 6) Long CSI reporting duration and having outdated CSI. 7) Frequency gap between uplink and downlink in frequency division duplex mode. All these make it important to find other dimensions of user feedback that the BS can exploit for NOMA clustering decisions, such as user location information or pictures from a scene captured by a camera-equipped BS (CBS).

Beyond 5G (B5G) systems can access the location information of users [148] and
exploit the directional nature of mmWave transmission, location aided beamforming (BF) strategies have been developed, e.g., [149]. Extending this idea to NOMA systems, in [150], a location-aided NOMA clustering strategy was developed that exploits the user location to assign a user to pre-defined cluster angles. CBSs, on the other hand, can be used to capture red-green-blue (RGB) images of users at a scene and utilize the vast potential of deep learning (DL) algorithms on these images for performing wireless communication tasks [151–153]. In [151], a synthetic data generation framework for RGB images and the associated user channels for mobile users was developed. In [153], for instance, Alrabeiah et al. applied convolutional neural networks (CNN) to residual networks (ResNets) [154] to solve a beam prediction problem. An 18-layer residual network (ResNet-18) was adopted and customized to fit the beam prediction problem.

In this chapter, we motivate the use of CBSs to enhance the performance and robustness of mmWave-NOMA systems, which does not need to consume any extra radio frequency (RF) resources. The camera feed of CBSs can be used for user clustering exclusively when CSI is hard to obtain, or the feed can be used in conjunction with CSI not only to improve the accuracy and the quality of the scheduling decisions, but also to reduce the amount of overhead and power consumption in the system. We highlight the different dimensions of user equipment (UE) feedback that a CBS can exploit for NOMA clustering, namely CSI-based feedback from the user and non-CSI-based feedback, comprised of the images captured by the CBS and UE location. Through a simple case study that performs NOMA clustering exclusively based on CBS images and user location feedback, we show that applying deep learning techniques to the images captured by the CBS can be exploited to achieve a spectral efficiency performance comparable to where NOMA clustering decisions are taken using the full accurate CSI of users. The results of our investigation highlight that in practical NOMA systems, either the CSI of users or visual feed of cameras or some combination of the two can be used interchangeably for NOMA clustering, depending on what feedback is available to the CBS in different situations.
5.2 Why Camera-equipped BS in mmWave-NOMA Clustering?

5.2.1 Dimension Space for Possible Clustering Approaches in mmWave-NOMA Systems

Figure 5.1: Illustrating the dimensions of feedback that a CBS can exploit for NOMA clustering decisions, categorized into CSI-based feedback and non-CSI-based feedback, comprised of UE location and BS camera feed images.

In this section, we illustrate the NOMA clustering problem statement for a CBS and the different dimensions of UE feedback that a CBS can use to solve this problem. Consider a multi-antenna, mmWave-NOMA enabled BS as shown in Fig. 5.1. Using the spatial dimension, users can be separated through beamforming. As we can see in Fig. 5.1, NOMA allows multiple users to be served in one beam. The goal of user clustering schemes in mmWave-NOMA systems is to identify sets of correlated users, so that they can be grouped into a NOMA cluster. Within a cluster, the users are separated via the power domain, called power-domain NOMA [113]. Clusters are separated from each other in the spatial domain, using beamforming techniques. Clustering, therefore, represents a user scheduling problem of identifying the set of
users to be served in each time slot and of identifying the candidate beams, such that a minimum quality-of-service is guaranteed to each user.

To solve this NOMA clustering problem, we can see that Fig. 5.1 shows the different dimensions of user feedback that a CBS can exploit. This feedback that the CBS relies on for NOMA clustering can be classified under two categories: CSI-based feedback and non-CSI-based feedback. The CSI-based feedback (i.e., the channel information dimension shown in the figure) incurs wireless channel overhead. The non-CSI-based feedback is comprised of two entities, namely user location feedback and pictures from the camera feed of the CBS. The latter is the main focus of this article. Any combination of these three dimensions of feedback may be available at different times, for different users, and in different situations. We note that while obtaining BS images and user location has an associated cost, it is not a cost on the wireless channel resources itself, the most precious commodity in a wireless system.

With ever-decreasing camera prices and location information already a common feature of 5G systems [148], it is not unrealistic to expect that CBSs will be able to exploit this information in the near future. CBSs can make clustering decisions with information from anyone dimension or combination of dimensions, depending on what is available.
5.2.2 Role of the Vision Dimension in Clustering

To make clustering decisions, current mmWave-NOMA clustering schemes are based on the CSI feedback from the users, typically using the cosine similarity or Euclidean distance metrics to determine the correlation between users [12]. A CBS allows for a new dimension of UE feedback that can be used for UE clustering decisions in NOMA systems. The visual dimension can be used in conjunction with or as a replacement for CSI, which is traditionally used for NOMA clustering. In other words, the visual information provided by an external unit can be used by the baseband on the CBS to make a decision about how to form the NOMA clusters using the following possible permutations:

- In case there is no access to the visual information, the CBS will rely on the channel information.
- In case there is no access to the channel information, the CBS will rely on the visual information.
- In case there is access to both channel and visual information, the CBS can combine information from both sources, e.g., through weighted summation, to make a better decision. Or, the CBS can resort to visual information as a fall-back solution, only when the reliability of channel information is low. To detect whether there is a reliability issue with channel information, the number of consecutive Hybrid Automatic Repeat Request (HARQ) Non-Acknowledgements (NACKs) and the number of consecutive Discontinuous Transmissions (DTXs) might be used.

Obtaining CSI is often challenging and problematic in terms of its availability and reliability [147]. This is particularly true for NOMA systems that are designed to serve a large number of users. Even when CSI is available, there is a large cost involved in obtaining the channel information of all users; something that conventional NOMA clustering schemes rely on to work. In particular, obtaining CSI for clustering in this way is wasteful for users who cannot be scheduled on the basis of current CSI, and a newly updated CSI is needed when they are actually scheduled. Additionally, using instantaneous CSI for UE scheduling decisions in future transmission slots presents issues related to the quality of CSI being used. Hence, even when CSI is available, there is an incentive to rely on CSI-free NOMA clustering approaches.
In mmWave systems that have dominant line-of-sight (LoS) paths, approaches such as the location-aware NOMA clustering scheme proposed in [150] can be used. However, a NOMA clustering approach based exclusively on user location feedback is limited to LoS paths and simple channel settings, such that the best beam for a user can be determined exclusively on the basis of user location and no other feature of the channel or surroundings. However, with CBSs, we can feed captured images to powerful DL image processing techniques. These DL techniques, using neural networks, can learn advanced features of the channel and make good beam predictions or cluster formation decisions. These neural networks can also be fed the location of the users as an additional input, particularly for identifying users in images captured by the CBS. Fig. 5.2 illustrates an example with a classroom deployment where a CBS captures the image of all users along with user location feedback. A deep learning algorithm parses the images from the classroom, using the user location feedback to identify the users. From a codebook of beams, the best beam for each user can be predicted; which can in turn be used for the clustering of users that all select the same best beam. This concept is illustrated through a case study in the next section.

5.3 A Case Study: Robust User Clustering in mmWave-NOMA via Vision-Based Deep Learning

For this case study, we build on the basic concepts of the NOMA-BB algorithm, presented in Chapter 3. Of course, NOMA-BB is an algorithm that works with perfect CSI available at the BS, but the goal of this case study is to derive an algorithm similar to NOMA-BB that can operate CSI-free, instead using the non-CSI dimensions of user feedback such as user locations and the camera feed of the CBS. For completeness sake, we briefly describe the key aspects of NOMA-BB again here that are relevant to this case study.

NOMA-BB is a simple CSI-based mmWave-NOMA clustering scheme. We focus on a single-path mmWave channel model in this case study. From a codebook of beams, such as the one shown in Fig. 5.3, the idea behind NOMA-BB is for the BS to identify the best beam for each user from the set of candidate beams. Using
Figure 5.3: Illustrating the NOMA-BB algorithm, where from a codebook of beams each user's best beam is determined and users that share the same best beam are clustered together.

the available CSI of each user, the BS uses the cosine similarity metric between the user channels and the fixed beamforming directions to determine the best beam for each user, $b_u^*$. NOMA-BB then clusters users that all have the same best beam, hence the name NOMA-BB. Additionally, there is a constraint on the maximum number of users, $n_{\text{max}}$, not to overload certain clusters with too many users, since in practice, each digital unit has a limit on the number of parallel transmissions per time slot. If the number of users with the same BB is greater than $n_{\text{max}}$, NOMA-BB simply splits them into more than one cluster, all served by the same beam but in different time slots. A similar but complementary constraint is considered in [132] where each user had their own individual successive interference cancellation (SIC) decoding capability, since in practice, the NOMA users cannot perform SIC to decode an infinite number of other users’ signals due to the limited computational and energy and memory resources of the UEs. The goal of the case study is to see if NOMA clustering can be done using only the visual feedback from the CBS and
how it performs compared to approaches such as accurate CSI-based NOMA-BB. The key point is to show that CSI is not essential for determining the best beam of each user, \( b_u^* \). If this \( b_u^* \) can be determined through the other non-CSI-based UE feedback dimensions discussed in Section 5.2.1, the rest of NOMA-BB can be used for NOMA clustering.

For this case study, we expand on the deep learning (DL) based beam prediction scheme proposed in [153], where a ResNet-18 neural network pre-trained on the popular ImageNet2012 dataset is customized for the purpose of beam prediction. Typical DL neural networks for image classification are skilled at classifying images into the appropriate class with sufficient training examples of images that belong to the different classes. In [153], it is shown that with a pre-defined BF codebook like in our problem, learning beam prediction from the RGB images degenerates to an image classification task where the goal of the system is to identify to which sector a user belongs. In other words, since the set of candidate beam vectors divides the scene (spatial dimensions) into multiple sectors, and single-user images are used to train the neural network, the image classification DL algorithm identifies the sectors to which a user belongs. Thus, the algorithm finds the users best-beam direction using the image captured by the CBS only. However, the images used to train the ResNet-18 neural network are single-user images (i.e., they contain only one user per image).

This concept can be extended to NOMA clustering problems where a CBS will capture images with hundreds of users that it needs to serve. As described above, an algorithm like NOMA-BB can be adapted to make the user clustering decisions based on the best-beam prediction from a neural network. Fig. 5.2 shows what this deployment would look like in a classroom, where a CBS captures images of all users. Unlike in [153] that assumes single-user images, in our case and in reality, we will have multiple users in an image that are to be served by the CBS, particularly in NOMA use-cases where the goal is to serve a large number of users. However, the location information can be used to break down these multi-user images into single-user images. We just need to learn the best beam from each user’s perspective, and this can be done with training examples from any user in the scene. This is because in a single-path LOS scenario with no obstacles, and where all users have equal channel gain, the best beam learned for user A at location \((X,Y)\) implies that the best beam for user B at location \((X,Y)\) is also the same. In more complex channel models, we
Algorithm 5: Proposed DL-based NOMA clustering algorithm using the images captured by the CBS.

Stage 1 (Training):
1. CBS collects CSI from users and images from the scene.
2. CBS uses CSI to generate training data for the ResNet-18 beam prediction algorithm.
3. NOMA clustering is done using CSI at this stage.

Stage 2 (Execution):
1. CBS uses UE’s location input and the trained ResNet-18 model to pick the best beam, \( b_u^* \), for each user \( u \).
2. NOMA-BB:
   \[
   \text{for } (beam-b : B_c) \text{ do}
   \]
   A. Group all \( n \) users that picked \( b_u^* = b \) into a cluster, to be served by beam-\( b \).
   B. If (\( n = 0 \)), do not form a cluster to be served by beam-\( b \).
   C. If (\( n > n_{\text{max}} \)), split the \( n \) users into \( \lceil \frac{n}{n_{\text{max}}} \rceil \) clusters, all served by beam-\( b \).
   \[
   \text{end}
   \]

Stage 3 (Validation):
1. Collect CSI of random users and validate against best beam, \( b_u^* \), predicted by ResNet-18 model.
2. If error threshold reached, go back to Stage 1.

Note: Stage 2 and Stage 3 run in parallel.

We divide the entire algorithm into three distinct stages, as shown in Algorithm 5. The first stage involves the training phase where the CBS learns how to make best beam predictions for all users in an image. To achieve this, the CBS collects CSI of all users to determine each user’s best beam using the cosine similarity metric like in [120], and that is provided as a training sample to the DL algorithm along with a picture of the scene. During this phase, since the CBS is not yet trained to make predictions using the DL algorithm, the CSI is used to make NOMA clustering decisions to provide system continuity. Once the DL algorithm is sufficiently trained, the CBS moves to Stages 2 and 3 jointly. In Stage 2, the CBS stops collecting the users’ CSI and instead starts using the beam predictions made by the trained DL algorithm to feed to NOMA-BB for the final NOMA clustering decisions. In parallel to this, Stage 3 is run where the CSI of a few users is collected in order to validate against the best beam prediction from the DL model. If a defined error threshold is
Figure 5.4: Simulation results highlighting the close performance of the proposed DL approach using camera and location feed (non-CSI feedback) with the clustering schemes that use CSI feedback. Reverted to Stage 1 if the environment has sufficiently changed from the originally trained model. For instance, in LoS dominated scenarios like classrooms, coffee shops, etc., where there can be many users to serve (user devices
like today, but also internet-of-things (IoT) devices in the future), the environment
does not change that often and so we will not have to keep reverting from Stage 3 to
Stage 1.

The performance of this proposed scheme is shown in Fig. 5.4, where we can see
that the DL camera-based clustering scheme is able to achieve comparable perform-
ance in spectral efficiency to the CSI-based scheme. We can see that the perform-
ance of the camera feed approach improves as more training data is used in Stage
1 to train the ResNet-18 based DL model. The difference in performance between
the CSI and camera feed schemes shown here is entirely due to the beam prediction
aspects. Errors in predicting the best beam for some users mean they are placed in
clusters where they do not receive the maximum possible signal-to-noise and interfe-
rence ratio (SINR). It is worth mentioning that the user ordering and power allocation
parts of the NOMA scheme depend on the CSI of the user, as it is most optimal
to order the users for SIC decoding in the order of their channel gains. In a truly
CSI-independent scheme, we would have to do this arbitrarily, which would cause a
loss of performance as well.

5.4 Challenges and Future Directions

In this section, we discuss potential challenges and future research directions for
implementing DL-based NOMA clustering algorithms using the images captured by
CBSs.

**Camera coverage and cost**: The number of cameras used, the placement of
these cameras, and the quality of the pictures captured by the CBS are all interesting
questions that will have an impact on the NOMA performance. Of course, better
quality images equal more cost, so exploring such trade-offs, particularly for low-cost
small cell BSs is an important aspect for practical deployments.

**Frequency of camera updates and mobility aspect**: A cluster of users has
to be determined by the baseband unit on the order of hundreds of milliseconds. The
current camera feed might not provide updates that often and then have clustering
decisions be made accordingly. Hence, the control interface between providing new
camera updates, making beam predictions and accordingly NOMA clustering deci-
sions is an important challenge to address. On the other hand, in this chapter, it
is assumed that user locations do not change faster than every few seconds, since in
practice, how often the user location (i.e., beam-index information) changes will have a direct impact on the load of the control interface and the design of the interface.

**Imprecise location information:** Another challenge is addressing imprecise location information. Since the system relies on the location information to identify users on the scene, other object detection techniques such as [155] can be used for user identification to complement user location feedback.

**Complexity of channel models and amount of data:** In multi-path settings, learning the best beam from a set of candidate beams could involve more advanced features, and thus also, require more training data. For example, multi-path poses a challenge to the neural network to learn the requisite advanced features of the channel. Additionally, users themselves can be obstacles and alter the multi-path setting and so the best beam of the users.

**Privacy and security:** An important concern with the use of cameras for clustering is the user privacy and security concerns that come with it. This is true for any vision-assisted scheme where the BS gets access to camera images that identify individual users. With the proposed scheme, given that the camera feed is used only for clustering decisions, regulations can be built-in to ensure the BS does not store any history related to the user.

For cases where the pictures might be stored for some time for offline training of the model, the user context and identity of the users can be removed; so nothing can be traced back.

**Anticipatory networking with visual information:** To reduce operational and cost inefficiencies of the next-generation wireless networks, knowing the future user distribution in both spatial and temporal domains, i.e., forecasting the future state of the network, at various time scales is critical [156]. Visual information collected by various cameras on the BSs can be leveraged to predict user distribution with high precision, and to optimize not only the network functions such as handover, but also, the network performance through techniques such as cell dimensioning, cell switch-off, and load balancing.

### 5.5 Conclusion

In this chapter, the potential of CBSs in mmWave-NOMA systems was explored to improve the robustness of user clustering. We showed three different dimensions of
user feedback that a CBS can exploit for user clustering, depending on the situation; CSI-based feedback or non-CSI-based feedback, including the UE location and camera feed of the CBS. Exploiting the advances in deep learning, the NOMA clustering problem can be addressed using the images captured by CBSs, without consuming extra RF resources. Through a case study, we showed that such an approach achieves comparable performance to a CSI-based approach, and user clustering can continue to function without much performance loss even in the scenarios where CSI is severely outdated or not available at all. Lastly, open challenges and future research directions in this area were discussed.
Chapter 6

PAPR in Uplink SCMA-OFDM Systems: The Impact of Modulation Schemes and Sub-carrier Allocation

Sparse code multiple access (SCMA) is a code-domain NOMA uplink solution that overloads resource elements (RE’s) with more than one user. Given the success of orthogonal frequency division multiplexing (OFDM) systems, SCMA will likely be deployed as a multiple access scheme over OFDM, called an SCMA-OFDM system. One of the major challenges with OFDM systems is the high peak-to-average power ratio (PAPR) problem, which is typically studied through the PAPR statistics for a system with a large number of independently modulated sub-carriers (SCs). In the context of SCMA systems, the PAPR problem has been studied before through the SCMA codebook design for certain narrowband scenarios, applicable more for low-rate users. However, we show that for high-rate users in wideband systems, it is more meaningful to study the PAPR statistics. In this chapter, we highlight some novel aspects to the PAPR statistics for SCMA-OFDM systems that is different from the vast body of existing PAPR literature in the context of traditional OFDM systems. The main difference lies in the fact that the SCs are not independently modulated in SCMA-OFDM systems. Instead, the SCMA codebook uses multi-dimensional constellations, leading to a statistical dependency between the data carrying SCs. Further, the SCMA codebook dictates that an UL user can only transmit on a subset of the available SCs. We highlight the joint effect of the two major factors that influence the PAPR statistics - the phase bias in the multi-dimensional constellation design along with the resource allocation strategy. The choice of modulation scheme and SC allocation strategy are static configuration options, thus allowing for PAPR reduction...
opportunities in SCMA-OFDM systems through the setting of static configuration parameters. Compared to the class of PAPR reduction techniques in the OFDM literature that rely on multiple signalling and probabilistic techniques, these gains come with no computational overhead. In this chapter, we also examine these PAPR reduction techniques and their applicability to SCMA-OFDM systems.

6.1 Introduction

Non-orthogonal multiple access (NOMA) solutions are being actively studied to address the massive connectivity requirements for 5G and beyond 5G (B5G) communication systems [157]. The sparse code multiple access (SCMA), proposed in [158], is one such NOMA scheme that has received a lot of attention particularly for the uplink (UL) direction [159]. SCMA will likely be used as a multiple access scheme over orthogonal frequency division multiplexing (OFDM), which is referred to as an SCMA-OFDM system [160,161]. In SCMA-OFDM systems, the orthogonal OFDM sub-carriers (SCs) are the resource elements (RE’s) over which the SCMA codewords are spread. Traditional OFDM systems that independently modulate the individual SCs are known to create large power peaks compared to the average power, resulting in the well known peak-to-average power ratio (PAPR) problem [162]. A high PAPR means that the power amplifier needs to operate in an inefficient region to avoid power leakage, which in turn affects the battery-life of the transmitting UL end-user device. Hence, studying the PAPR problem in an OFDM system that uses a NOMA scheme like SCMA is an important problem for 5G and beyond communication systems.

In the context of traditional OFDM systems that independently modulate a large number of SCs, the PAPR attained is a random quantity, since it depends on the sequence of complex-valued constellation points transmitted in the OFDM SCs along with the symbol rate. The PAPR can then be analysed in terms of its maximum theoretically attainable value based on the constellation scheme used to modulate the individual SCs. Alternatively, the PAPR can be studied in terms of its statistics using the complementary cumulative distribution function (CCDF), often referred to as the PAPR statistics [163]. It has been shown that when the number of SCs is sufficiently large, the maximum theoretically attainable PAPR value occurs with negligible probability and the PAPR statistics offer more meaningful insights [163–166]. Since traditional OFDM systems typically involve independently modulated
SCs, the PAPR statistics have been characterized with this assumption in several studies [164,166,167]. However, using SCMA as a multiple access scheme over OFDM means that the individual SCs are not independently modulated; thus motivating the need to characterize the PAPR analysis specifically for SCMA-OFDM systems.

In the SCMA construct, each user maps its incoming bits to a multi-dimensional modulation symbol coded over multiple RE’s, which is termed as a codeword. Each modulation symbol has its own codeword and together they form a codebook that is ideally unique to a user. In this way, the users are separated by their unique codebooks. The SCMA codebook design problem involves the multi-dimensional constellation design [159] and several such constellations have been proposed in the literature [168–172]. This SCMA codebook design provides an additional degree of freedom to the PAPR problem in SCMA based systems [173] and corresponding multi-dimensional constellation schemes that minimize the PAPR that can theoretically be attained have been proposed in [174–177]. However, we show in this chapter that like with traditional OFDM systems, this maximum theoretically attainable PAPR based on the codebook design is a meaningful metric only for low-rate users or in narrow-band systems. For high rate users, i.e., when a larger number of modulation symbols are transmitted in the same OFDM symbol duration, it is more meaningful to study the PAPR statistics. While the SCMA paradigm is often discussed for massive connectivity deployment involving low-rate IoT devices [173], SCMA can just as easily be used for traditional wireless devices and other high-rate users in 5G and B5G networks [160,161,178]. The characteristics of the SCMA codebook that affect the PAPR statistics are different from those studied to date in the SCMA literature to the best of our knowledge, and is the focus of this chapter.

Since SCMA employs multi-dimensional constellations, each modulation symbol is transmitted over multiple individual SCs. Thus, the transmitted OFDM SCs are not independently modulated like they are in traditional OFDM systems. The multi-dimensional constellation design dictates what is transmitted in the individual SCs. The PAPR statistics in SCMA-OFDM systems will thus reflect this dependency between the modulated SCs. For instance, if the SCMA codeword consists of constellation points all of the same phase and the SCs over which they are transmitted are contiguous, it will have a detrimental effect on the PAPR statistics. Alternatively, an SCMA scheme where the codewords contain constellation points of opposite phases and transmitted over contiguous SCs is likely to have a positive effect on the PAPR.
statistics. However, the SCs carrying SCMA codewords do not necessarily have to be contiguous SCs. In other words, the SCs that carry the codewords can be located anywhere in the frequency spectrum. Hence, the joint impact of the SCMA modulation scheme along with the OFDM SC placement must be considered when studying the PAPR statistics for high-rate SCMA-OFDM users.

In this chapter, we highlight these novel aspects to the PAPR analysis for SCMA-OFDM systems that is different from the vast body of existing PAPR literature in the context of traditional OFDM systems. We highlight two major factors that influence the PAPR statistics - the phase bias in the multi-dimensional constellation design along with the resource allocation strategy\(^1\). By considering such a cross-layer systematization perspective, we motivate the fact that significant PAPR reduction can be achieved through the setting of static configuration parameters. Such gains could come without any additional overhead to the system that is typically introduced by most multiple signalling based PAPR reduction techniques that work with the assumption of independently modulated SCs, i.e., no a-priori knowledge of any statistical dependencies in the transmitted signal \([162]\).

Further, in the context of traditional OFDM systems, numerous techniques that improve the PAPR statistics have been proposed, as captured by the surveys in \([162,179,180]\). Broadly, there are two categories of such PAPR reduction techniques: signal distortion based techniques and multiple signalling and probabilistic based techniques. Signal distortion techniques like clipping distort the transmitted signal by not transmitting any power peaks above a certain threshold \([181]\). Alternatively, PAPR reduction techniques based on multiple signalling generate a set of candidate signals every OFDM symbol and transmit the signal with the least PAPR. The set of candidate signals are generated through operations like phase changes \([182]\) or interleaving \([183]\) on the original data set. There is a significant complexity overhead in generating these extra candidate signals as well as some throughput loss since sidelink information about the operations performed on the data set needs to be transmitted to the receiver \([162]\).

In \([160]\) and \([161]\), the class of signal distortion techniques is studied in the context of high-rate SCMA-OFDM systems. Specifically, these two papers investigate the challenge of allowing the SCMA receiver to cope with the distortions introduced by signal clipping at the transmitter. However, the class of multiple signalling and

\(^1\)The placement of the SCs that carry an SCMA codeword is basically a resource allocation problem. In this chapter, we use the terms resource allocation and SC placement interchangeably.
probabilistic techniques has not been thoroughly examined in the context of SCMA-OFDM systems to the best of the authors knowledge. As we discuss in this chapter, some of the techniques that involve constellation shaping [184,185] cannot easily lend itself to SCMA systems because it affects the SCMA constellation design. However, other multiple signalling techniques such as selective mapping (SLM) [182], partial transmit sequences (PTS) [186] and interleaving (IL) [183] can be tailored to meet the constraints of an SCMA-OFDM system. In this chapter, we discuss what adaptations are needed to these well known PAPR reduction techniques to make them work in SCMA-OFDM systems. Moreover, in traditional OFDM systems where each SC is independently modulated, until the SCs to be transmitted in an OFDM symbol are known, there is no way to know which SC allocation strategy results in the least PAPR. Hence, a certain number of permutations are tried dynamically every OFDM symbol and the one with the least PAPR is transmitted. However, with SCMA-OFDM systems, we can exploit the statistics known in advance through the novel aspects we present in this chapter to reduce or even eliminate the complexity and sidelink information overhead typically incurred by these PAPR reduction techniques. Further, it is worth mentioning that these PAPR reduction techniques have also been recently investigated in other non-SCMA based NOMA systems [187–190], but these are beyond the scope of this chapter.

The contributions of this chapter can then be summarized as:

- We highlight the two main factors that impact the PAPR statistics in SCMA-OFDM systems as a result of the dependency between data carrying SCs - the phase bias in the SCMA constellation design and the accompanying resource allocation strategy.

- We show that such a resource allocation based PAPR analysis allows for PAPR gains through the setting of static configuration parameters that does not incur any computational overhead. Such gains are not possible in traditional OFDM systems that individually modulate the SCs.

- Finally, we analyse the class of PAPR reduction techniques based on multiple signalling and probabilistic techniques in the context of SCMA-OFDM systems. We then compare the static PAPR gains from the resource allocation based strategies with the gains from these well known PAPR reduction techniques and offer some insights into how they can be used together to improve the
PAPR while minimizing complexity and throughput loss.

6.2 System Model: Uplink SCMA-OFDM Transmission

Consider an SCMA-OFDM system where the total bandwidth is comprised of \( Z \) OFDM SCs. The \( Z \) SCs are uniformly divided into SCMA blocks of size \( N \). Therefore, a total of \( N_B = Z/N \) SCMA blocks can be configured in the system. Within one SCMA block, \( K \) users share the \( N \) SCs such that \( N < K \). Due to the sparse overloading of SCMA, each user is assigned to only \( d_v \ll N \) SCs. If the SCMA block is fully loaded, each user is assigned a unique combination of \( d_v \) SCs and

\[
K = \binom{N}{d_v}.
\]  

(6.1)

In an \( M \)-point signal constellation, each \( N_M = \log_2 M \) bits for each user is sent over the \( d_v \) SCs. A user-to-SC binary allocation matrix \( S \) of dimensions \( N \times K \) dictates which \( d_v \) SCs are assigned to which users. Every row in \( S \) represents a SC, while every column represents a user. For example, for \( N = 4 \), \( d_v = 2 \) and \( K = 6 \), a sample user-to-SC allocation matrix is

\[
S = \begin{bmatrix}
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 & 0
\end{bmatrix}.
\]  

(6.2)

As the first and third positions in the first column of \( S \) in (6.2) are non-zero, we can say the first user has allocation “1010” in one SCMA block. In other words, user 1 is assigned to the first and the third SCs. Similarly, the second user is assigned the second and fourth SC in the SCMA block, i.e., it has allocation “0101” and so on. Since this is a fully loaded system, all unique combinations of \( d_v = 2 \) SCs are covered in the \( K \) columns of \( S \).

We consider a modulation symbol to be the representation of \( \log_2 M \) bits, e.g., when \( M = 4 \), the modulation symbols are ‘00’, ‘01’, ‘10’ and ‘11’. Each user can send
Figure 6.1: The transmitter for User-1 in an SCMA-OFDM system where 6 users share 4 SCs, with 2 non-zero SCs allocated to each user.

$L (L \leq N_B)$ modulation symbols over $L$ different SCMA blocks in the same OFDM symbol duration. For simplicity, we assume each user has the same user allocation in each of these $L$ blocks. For instance, for the user-to-SC allocation matrix in (6.2), user 1 is assigned to the allocation “1010” over all the $L$ blocks.

In an SCMA-OFDM system with $K$ users, each symbol of the $k^{th}$ user is mapped to a $d_v$-dimensional complex constellation $\tilde{x}_k = (\tilde{x}_{1,k}, ..., \tilde{x}_{d_v,k})^T$, that is selected from the columns of a $d_v \times M$ matrix called $X_k$. As in [159], for uplink transmission, we assume that the constellation scheme used is the same for all $K$ users. Hence, when describing the transmitter for one user in an SCMA-OFDM system, the index $k$ can be dropped and the $d_v$-dimensional SCMA constellation scheme for the user can be represented by $X$. Each column represents what the user transmits for the $m = \{1, .., M\}$ symbol, i.e., $X = (x_1, ..., x_M)$ and $x_m = (x_{1,m}, ..., x_{d_v,m})^T$.

In an SCMA-OFDM system, the transmitter maps $L$ sets of $N_M$ bits to $L$ modulation symbols, based on $X$. These $L$ modulation symbols will be carried over the $L \times d_v$ SCs assigned to it and the user is required to leave the other SCs in the system as null SCs, i.e., the user does not transmit anything on these null SCs. Let $Y_i$ denote the complex constellation point transmitted in SC$_i$ in the system. $Y_i$ will either be null if SC$_i$ is not assigned to the user, or else $Y_i$ will contain $x_j \in \tilde{x}_m$, $j \in \{1, .., d_v\}$, for the symbol $m$ transmitted in SC$_i$. An inverse fast Fourier transform (IFFT) based implementation of an OFDM transmitter uses the input on these SCs to generate the...
discrete time domain samples of the signal, $y[n]$, as follows:

$$y[n] = \frac{1}{\sqrt{Z}} \sum_{i=0}^{Z-1} Y_i e^{j2\pi in/Z}.$$  \hfill (6.3)

Fig. 6.1 illustrates an example of the transmitter of an SCMA-OFDM system. In the top-left part of the figure, one SCMA block with $K = 6$, $N = 4$, $d_v = 2$ with the user-to-SC allocation matrix $S$ from (6.2) is depicted. The users are labelled from $U_1$ through to $U_6$ and the SCs from SC$_1$ through to SC$_4$. The coloured boxes indicate that the corresponding SC is assigned to the user from $S$. The IFFT based implementation of the transmitter for $U_1$ is then illustrated. Since, from (6.2), $U_1$ has allocation “1010”, it uses the first and third SC in every SCMA block it transmits on. In this example, $U_1$ transmits one modulation symbol in each of the available SCMA blocks, i.e., $L = 5$. This corresponds to the user using 5 blocks with 2 SCs in each block, which constitutes a total of 10 SCs.

The $M$-point SCMA modulation scheme determines how $N_M$ bits of user data are mapped to the $d_v$ SCs allocated to a user. The known multi-dimensional constellation schemes in the literature are outlined in detail in the survey in [159]. In Fig. 6.2, we show multi-dimensional SCMA constellation schemes when $M = 4$ and $d_v = 2$. The first two constellations are named 4-LDS and 4-Bao respectively, which follows the same naming convention as the authors in [159] for consistency. The third, named 4-OPP, is a new constellation we will introduce in Section 6.3.2.

### 6.3 PAPR in Traditional OFDM Systems vs. PAPR in SCMA-OFDM Systems

In this section, we present the key differences between studying the PAPR in traditional OFDM systems that independently modulate the SCs vs. in an SCMA-OFDM system. As we discussed in the introduction in Section 6.1, the PAPR attained during an OFDM signal transmission is a random quantity that can be analysed in terms of the maximum theoretically attainable PAPR in a given OFDM symbol duration or through a statistical characterization of the PAPR, called the PAPR statistics. In traditional OFDM systems, the maximum theoretically attainable PAPR can be determined through the knowledge of the constellation scheme used to independently
Figure 6.2: The 4-point SCMA constellations used in this study.
modulate each SC, e.g., \( M \)-QAM. In SCMA-OFDM systems, the theoretically attainable PAPR can be determined in a similar way but must account for the multi-dimensional constellation scheme in use. However, in the vast body of PAPR literature for OFDM systems, it has been shown that when the number of SCs is sufficiently large, i.e., high-rate users, the maximum theoretically attainable PAPR value occurs with next to negligible probability and the PAPR statistics offer more meaningful insights [163–166]. The same is true for an SCMA-OFDM transmitter that transmits over a large number of SCs in the same OFDM symbol duration, i.e., high-rate users. However, the discussion on PAPR statistics for SCMA-OFDM systems is different because of the statistical dependency between SCs. In what follows, we discuss how the combination of the multi-dimensional modulation scheme along with the accompanying resource allocation strategy impacts the PAPR statistics in such high-rate SCMA-OFDM systems.

### 6.3.1 PAPR in OFDM systems

An OFDM signal, \( y(t) \), that is generated from individually modulated subcarriers transmitted in the same OFDM symbol duration can be represented as

\[
y(t) = \sum_{k=0}^{Z-1} a_k \exp(j2\pi (f_c + k\Delta f)t) \\
= \exp(j2\pi f_ct) \sum_{k=0}^{Z-1} a_k \exp(j2\pi kt/T_s),
\]

where \( a_k \) is the complex-valued constellation point transmitted in SC \( k \), \( f_c \) is the centre frequency of SC \( k \) and \( Z \) is the total number of SCs in the system. If \( y(t) \) is sampled at a rate of \( Z/T_s \), i.e., every sample is taken at multiples of \( T_s/Z \), the discrete time version for the baseband part of \( y(t) \) from (6.4) can be expressed as

\[
y[n] = \sum_{k=0}^{Z-1} a_k \exp(j2\pi kn/Z).
\]

When the transmitted OFDM signal is generated from independently modulated SCs, the non-constant envelope creates large instantaneous peaks in the signal. These power peaks occur when the individual signals align in phase, which can be much
Figure 6.3: Illustration of the power peaks produced when summing sinusoids of evenly spaced frequencies as is the case in an OFDM signal.

larger than the average power of the transmitted signal and results in a high PAPR. This is illustrated in Fig. 6.3, where four sinusoids with equal subcarrier spacing are added together and the resultant signal has large power peaks. While PAPR applies to the continuous time transmitted signal, studies have shown that if oversampled with a sufficiently high ratio \( [191] \), PAPR can be accurately calculated from the discrete time samples \( y[n] \) from (6.5) as follows:

\[
PAPR \text{ (dB)} = 10 \log_{10} \left( \frac{\max(|y[n]|^2)}{E(|y[n]|^2)} \right). \tag{6.6} \]

The PAPR attained is a random quantity, since it depends on the sequence of complex-valued constellation points transmitted in the \( Z \) SCs along with the symbol rate. If \( a_k \) is selected from a QAM constellation of \( M \) points of equal magnitude, e.g., 4-QAM, then the maximum theoretically attainable PAPR is \( Z \) \([165]\). This is because when all the SCs add coherently, the instantaneous power is \( Z^2 \), while the average power to transmit \( Z \) SCs of unit energy signals each is \( Z \). However, it has been shown that when \( Z \) is sufficiently large, this theoretically attainable PAPR value occurs with negligible probability and the PAPR statistics offer more meaningful insights \([163–166]\). For example, with \( Z = 32 \), 4-ary modulation and an OFDM symbol duration of 100 \( \mu s \), the authors in \([164]\) showed that the theoretically attainable PAPR occurs every 3.7 million years. For an OFDM system with \( Z \) SCs, \( M^Z \) unique symbol sequences and thus \( M^Z \) unique OFDM waveforms per block can be generated \([163]\). Some of these waveforms will have low PAPR and some will have a
higher PAPR value. Since traditional OFDM systems typically involve independently modulated SCs, the PAPR statistics have been characterized with this assumption in several studies [164, 166, 167]. Since the Z SCs are individually modulated, and if Z is large, the central limit theorem dictates that the real and imaginary parts of the transmitted OFDM signal can be modelled by Gaussian random processes. As a result, the overall envelope of the transmitted signal follows a Rayleigh distribution [166].

### 6.3.2 PAPR in SCMA-OFDM Systems

When \( L \) modulation symbols, carried over \( L \) SCMA blocks, are transmitted in the same OFDM symbol duration, the maximum theoretically attainable PAPR of the transmitted signal can be computed from (6.6). Let \( X_{m\text{,max}} \) denote the maximum possible instantaneous peak to transmit symbol \( m \). Thus, \( X_{m\text{,max}} \) is the sum of the amplitudes of all \( d_v \) dimensions of \( x_m \). Let \( X_{\text{max}} \) denote the maximum possible instantaneous peak from the constellation scheme, i.e., the peak which occurs when the modulation symbol that contains the maximum peak is transmitted. The maximum attainable peak from transmitting \( L \) symbols is achieved when the symbol corresponding to \( X_{\text{max}} \) is transmitted on all SCMA blocks and each of the SCs line up in phase. Also, let \( P_{X,m} \) represent the power required to transmit \( x_m \) and \( P_{X,\text{avg}} \) represent the average power of transmitting a symbol from the constellation. The maximum possible PAPR value, calculated per OFDM symbol duration, is computed as follows:

\[
PAPR (\text{dB}) = 10 \log_{10} \left( \frac{|L \times X_{\text{max}}|^2}{L \times P_{X,\text{avg}}} \right), \tag{6.7}
\]

where

\[
X_{m\text{,max}} = \sum_{i=1}^{d_v} x_{i,m}, \forall \ m \in \{1..M\}
\]

\[
X_{\text{max}} = \max(X_{m\text{,max}})
\]

\[
P_{X,\text{avg}} = \frac{\sum_{m=1}^{M} \sum_{i=1}^{d_v} x_{i,m}^2}{M}.
\]

If \( L = 1 \), we can consider that as the constellation PAPR. It is clear that this constellation PAPR value is determined entirely from the design of the SCMA multidimensional constellation. For example, for the constellation schemes shown in Fig.
6.2, $X_{\max} = 1.4$, $P_{X,\text{avg}} = 1$ and so the constellation PAPR is 3.04 dB. In this context, codebook designs that minimize this theoretically attainable PAPR have been proposed [174–177]. For example, in [174], a low-PAPR codebook that minimizes the number of projections, i.e., non-zero dimensions in the constellation scheme, is proposed. If a zero dB PAPR constellation design is required, i.e., 0 dB constellation PAPR when $L = 1$, then each modulation symbol should be coded with the same amplitude on only one of the $d_v$ SCs. The constellation design approach for this is outlined at the end of this chapter as an Appendix.

We described in Section 6.3.1 that for traditional OFDM systems, when the number of modulated SCs is large, the theoretically attainable PAPR occurs with negligible probability and the PAPR statistics are more meaningful to study. An SCMA-OFDM user transmits $L$ modulation symbols per OFDM symbol duration over $L \times d_v$ SCs. Hence, when $L$ is large, the PAPR statistics become more meaningful to study in an SCMA-OFDM system. Since $L$ corresponds to the number of modulation symbols transmitted per OFDM symbol duration, we can equate a large value of $L$ with high-rate users. Hence, the PAPR statistics should be investigated for such high-rate users. The SCMA codebook influences the PAPR statistics in ways that are different from the PAPR perspective studied in SCMA systems for low-rate users in [174–177]. Further, the characterization of the PAPR statistics is different from traditional OFDM systems for the reasons we discuss next.

The PAPR statistics in SCMA-OFDM systems are different from traditional OFDM systems because of two main factors. Firstly, the presence of null SCs in the codebooks means that each user transmits on only a small fraction of the total available SCs. Secondly, the data carrying SCs are not independently modulated. One modulation symbol dictates what is transmitted in $d_v$ SCs. Hence, these $d_v$ SCs are dependent in the statistical sense. As we discussed in the system model in Section 6.2, it is the multi-dimensional constellation used in the SCMA codebook that determines what the user transmits on each of the $d_v$ SCs. This creates a statistical dependency between these $d_v$ SCs, since they are collectively determined by the choice of one modulation symbol. This dependency affects the PAPR statistics through the level of phase bias in the constellation design. To illustrate this concept of phase bias in an SCMA multi-dimensional constellation, we use two of the known multi-dimensional constellations from the SCMA literature when $M = 4$ and $d_v = 2$, named 4-LDS and 4-Bao respectively [159]. To these, we introduce another SCMA
constellation, namely 4-OPP. These three constellations are depicted in Fig. 6.2 and all have the same constellation points in each dimension, but are combined differently to form the codewords for the respective modulation symbols. The 4-LDS scheme repeats the same constellation point over all the dimensions in which they are coded. This means that all the SCs carrying an LDS modulation symbol are guaranteed to have the same phase. In 4-Bao, two symbols, “00” and “11”, are coded with the same point in both SCs (i.e., same phase), while the other two symbols, “01” and “10”, are coded with constellation points of exactly opposite phase. This means if the symbols “00” and “11” are transmitted, a phase bias of having two SCs with the same phase guaranteed will occur. On the other hand, when “01” and “10” are transmitted, the opposite bias is introduced. Given each of the four symbols are equally likely to be transmitted, we can expect these two opposing bias effects to statistically cancel each other out. 4-OPP on the other hand has the opposite effect of LDS and introduces a guaranteed 180° phase difference between the two constellation points for each and every symbol.

The impact that such a phase bias in the constellation scheme has on the PAPR statistics comes from the a-priori statistical dependency it introduces between transmitted SCs in the same OFDM symbol. This is a feature that is very different from traditional OFDM systems where the SCs are independently modulated. However, whether the phase bias has a positive or negative effect on the PAPR statistics compared to independently modulating the SCs is influenced by whether these dependent SCs are of similar or very different centre frequencies in the spectrum. For example, when 4-LDS with the phase bias of having constellation points all of the same phase is transmitted over \( d_v \) statistically dependent SCs that are near contiguous in the spectrum, it will have a detrimental effect on the PAPR statistics. Alternatively, when 4-OPP that contains a phase bias of having constellation points of exactly opposite phases is transmitted over contiguous SCs, it is likely to have a positive effect on the PAPR statistics. These are just a simple consequence of the way the signals in OFDM SCs add up to form the equivalent OFDM signal, as illustrated earlier in Fig. 6.3. However, the SCs carrying SCMA codewords do not necessarily have to be contiguous SCs. It depends on which SCs are allocated to the SCMA blocks assigned to the user. Hence, we refer to this as the SC allocation or resource allocation strategy. In what follows, we first describe different resource allocation strategies that could impact the PAPR statistics in Section 6.4 and then analyze the joint impact of the phase bias in
the constellation design and the accompanying SC allocation strategy on the PAPR statistics in Section 6.5.

### 6.4 Resource Allocation Schemes

As we discussed in Section 6.3.2, the placement of the dependent SCs in an SCMA-OFDM system influences the PAPR statistics. The placement of the dependent SCs is determined by how the SCs in the spectrum are assigned to carry the SCMA blocks, i.e., the resource allocation strategy. While any number of such resource allocation strategies can be considered, we focus on two resource allocations which represent the two extremes in terms of frequency separation, i.e., the spacing between the dependent SCs. In the illustration shown in the system model in Fig. 6.1, every $N$ contiguous SCs is grouped into an SCMA block. We term this as the *regular allocation*. However, if the grouping of every $N$ consecutive SCs to form an SCMA block is considered as a virtual view of the system SCs, it can be mapped in any way to the physical OFDM SCs. In the second resource allocation strategy, we separate the individual SCs that make up an SCMA block by as much as possible. Since separating the SCs that carry the SCMA codeword equates to providing frequency diversity between the $d_v$ SCs, we call this the *diversity allocation*. An illustration of this mapping for a system with 16 SCs is shown in Fig. 6.4. The virtual allocation, depicted as va in the figure, is the same as the regular allocation. The physical allocation, represented as pa in the figure, represents the diversity allocation. The algorithm for constructing this diversity allocation is discussed next.
6.4.1 Constructing the Diversity Allocation Scheme

Attaining frequency diversity translates to providing as much separation as possible between the non-zero dimensions over which an SCMA modulation symbol is transmitted. Since every pair of SCs in an SCMA block belong to some user’s allocation in user-to-SC allocation matrix $S$ as shown in (6.2), we seek some guaranteed minimum level of separation between every pair of SCs that belong to an SCMA block.

Let the OFDM SCs in the system be indexed as $\{f_1, f_2, ..., f_Z\}$, with the SC spacing between any two SCs denoted by $\Delta f$. If $\{f_1, ..., f_N\}$ constitutes the first SCMA block, $\{f_{N+1}, ..., f_{2N+1}\}$ constitutes the second SCMA block and so on, it is termed as the “regular allocation”. On the other hand, in the diversity SC allocation scheme, we distribute $N$ SCs to each SCMA block such that there is a minimum number of SCs that separate any pair of SCs in an SCMA block. In order to define this diversity allocation scheme, we will treat the regular allocation as the virtual allocation of SCs to SCMA blocks and define a mapping from the virtual allocation to the physical allocation of SCs in the system. We thus define two $Z$-dimensional vectors, $\mathbf{va}$ and $\mathbf{pa}$, to represent the virtual and physical allocation of SCs in the system, respectively. The goal of this diversity scheme is to provide a mapping from $\mathbf{va} \rightarrow \mathbf{pa}$ such that in the $\mathbf{pa}$ vector, every pair of SCs in an SCMA block is separated by at least a certain number of SCs. An example of this mapping was illustrated in Fig. 6.4 for a system with $Z = 16$ and $N = 4$.

We first seek to concretely determine this minimum level of SC separation that can be attained. Let $\mu$ represent the minimum number of SCs that separate any pair of SCs in an SCMA block. The number of SCMA blocks is $N_B = \frac{Z}{N}$. If the first SC in every block from $\mathbf{va}$ is mapped contiguously to the first available index in $\mathbf{pa}$, the last block starts at index $N_B$. For the example in Fig. 6.4, indexes $\{1, 5, 9, 13\}$ represent the first index of each block in the virtual allocation, which are placed contiguously at the start of the physical allocation. Clearly, the first index of the last block, i.e., index 13 from $\mathbf{va}$, gets placed at index $N_B = 4$ in $\mathbf{pa}$. The last available SC index in the system is $Z$, so $\mu_{\text{max}}$ for the remaining $(N - 1)$ SCs in each SCMA block will be such that they are equally spread apart and can be determined as

$$\frac{Z}{N} + (N - 1)\mu_{\text{max}} = Z. \quad (6.8)$$

Solving (6.8), we get $\mu_{\text{max}} = \frac{Z}{N} = N_B$, which is the maximum amount of SC
spacing we can guarantee to any pair of SCs in every SCMA block in the system. Algorithm 6 then describes the mapping from \( \text{va} \rightarrow \text{pa} \) such that in the \( \text{pa} \) vector, every pair of SCs in an SCMA block is separated by at least \( \mu_{\text{max}} \) SCs.

The algorithm iterates through each of the SCs in \( \text{va} \) but operates on an SCMA block by SCMA block basis. When it detects the start of a new SCMA block in \( \text{va} \), it takes the first SC in the block and assigns it to the smallest available index in \( \text{pa} \). Every subsequent SC in the block from \( \text{va} \) is then placed \( \mu_{\text{max}} \) SCs apart. This is done for all the \( N_B \) blocks in the system. For example, if \( Z = 128 \) and \( K = 4 \), then \( \mu_{\text{max}} = 32 \). The algorithm starts from \( \text{va}_1 \) which is assigned to \( \text{pa}_1 \). The SC in \( \text{va}_2 \) will then be placed 32 SCs apart at \( \text{pa}_{33} \). Similarly, allocations \( \text{va}_3 \rightarrow \text{pa}_{65} \) and \( \text{va}_4 \rightarrow \text{pa}_{97} \) are made. Now, \( \text{va}_5 \) represents the start of a new SCMA block and is hence assigned to \( \text{pa}_2 \), the smallest available index since \( \text{pa}_1 \) is used. Again, the remaining SCs in this block are placed 32 SCs apart starting from index 2 and the process repeats for the remaining blocks. The final SCMA block will have its first SC placed at \( \text{pa}_{32} \) and final SC at the last available index at \( \text{pa}_{128} \).

Returning to the example in Fig. 6.4, it also depicts how different users get a different level of SC separation depending on their user allocation, from the user-to-SC matrix in (6.2). A user with allocation “1100” has a spacing of four SCs between the coded dimensions while the user with allocation “1001” gets a much larger separation of twelve SCs. However, all users are assured a separation of at least \( \mu_{\text{max}} \) between their coded dimensions.

### 6.5 Simulation Results

We describe the joint impact of the SCMA constellation design and SC allocation on the PAPR statistics of high-rate users with the help of the MATLAB simulations presented in Fig. 6.5. All experiments were run for a large number of OFDM symbols, in the order of \( 10^4 \). The total number of SCs in the system, \( Z = 128 \), are divided into \( N_B = 32 \) SCMA blocks of \( N = 4 \) SCs each with \( L = 32 \). Also included in the results in Fig. 6.5 is a simulation run for randomly generated independent 4-QAM constellation points transmitted only on the data carrying SCs assigned to the user under test from the user-to-SC allocation matrix \( \text{S} \) in (6.2). Note that this is different from 4-LDS, because the SCs are being individually modulated. We illustrate the regular scheme for the user with allocation “0011” from matrix \( \text{S} \), while the diversity scheme for the
Algorithm 6: Proposed diversity-based SC mapping

**Input**: \( va \rightarrow \) a size \( Z \) vector representing the virtual allocation where every \( N \) consecutive entities represent an SCMA block

**Output**: \( pa \rightarrow \) a size \( Z \) vector representing the physical allocation of SCs, where at least \( Z/N \) SCs separate the entities of an SCMA block.

initialize \( pa \leftarrow \infty \) (all elements);

for \( i \leftarrow 1 \) to \( Z \) do

  if \( i \mod N = 1 \) then

    for \( j \leftarrow 1 \) to \( (Z/N) \) do

      if \( pa(j) \neq \infty \) then

        \( pa(j) \leftarrow va(i) \);

        break;

    else

      \( n \leftarrow i \mod N \)

      if \( n = 0 \) then

        \( n \leftarrow N \)

      \( pa(j + n \times (Z/N)) \leftarrow va(i) \)

users with allocations “1001” and “0011” for the reasons we outline next.

Since each user is allocated a different set of \( d_v \) SCs per SCMA block to transmit on (defined from matrix \( S \)), the PAPR statistics of each user will not necessarily be the same. With the regular scheme, in our illustration, an SCMA block is comprised of \( N = 4 \) contiguous SCs. Hence, any combination of \( d_v = 2 \) SCs comprises SCs of similar centre frequencies. Thus, any one of the possible user allocations, e.g., “0011”, is representative of the PAPR performance for all users. On the other hand, with the diversity scheme, different user allocations experience different levels of frequency separation. For instance, referring to the example in Fig. 6.4, the user with allocation “1001” has the first and fourth SC in the virtual allocation which are separated the furthest, while the user with “1001” has the third and fourth SC in the virtual allocation which is separated the least. Thus, the user with allocation “1001” represents the maximum SC separation scenario while “0011” corresponds to the user having the minimum SC separation.

We see from the results in Fig. 6.5 that with the regular allocation scheme, 4-OPP outperforms 4-LDS. With 4-OPP, we are placing \( d_v = 2 \) constellation points of
Figure 6.5: Comparing 4-LDS, 4-Bao and 4-OPP with different SC allocation schemes. For illustration, the curve when independent 4-QAM symbols (non-SCMA symbols) are transmitted in the same data carrying SCs is included.
equal magnitude but opposite phase in two near-contiguous SCs. While with LDS, we have the guaranteed placement of two points with the same amplitude and same phase in near contiguous SCs, i.e., SCs of similar centre frequencies. The sum of two sinusoids of similar frequencies will line up for higher peaks if they start at the same phase. However, with the diversity scheme, we see the trend shifts. With the “0011” allocation that provides the minimum frequency separation, we start to see 4-LDS and 4-OPP behave similarly, while with the maximum separation “1001” user allocation, the results are the exact opposite of the regular scheme. Increasing the SC separation between the $d_w$ SCs, means that we are adding sinusoids of increasingly different frequencies to generate the OFDM signal. As the frequency separation becomes large enough, the simulations show the modulation scheme biased to have both dimensions start at the same phase, i.e., 4-LDS, generates better PAPR statistics.

The complete change in the order from Fig. 6.5a to Fig. 6.5c highlights why it is important to study the PAPR statistics as the joint effect of the SCMA modulation scheme and the corresponding SC allocation strategy. Hence, when comparing the PAPR performance of different SCMA modulation schemes from the literature [159], it is not sufficient to conclude that one scheme outperforms the other. The modulation schemes have to be analyzed in conjunction with the associated SC allocation strategy to fully understand their impact on the PAPR statistics. Since the modulation scheme is a physical layer design parameter while the SC allocation comes from the layer-2 resource allocation strategy, the PAPR problem for high-rate users in SCMA-OFDM systems should be studied as a cross-layer systematization problem. As we showed in Section 6.3.2, this is in contrast to the low-rate users with a small value of $L$ where it is sufficient to analyse the PAPR purely from the layer-1 perspective of the SCMA multi-dimensional constellation design.

Further, as seen in Fig. 6.5, with each SC allocation, 4-Bao performs similar to just placing random 4-QAM points in the data carrying SCs. This is because it contains an equal mix of same and opposite phase bias among its constellation points, so the PAPR statistics reflect that it is no different from independently modulating the data carrying SCs. However, the SCMA codebook still plays an important role in the PAPR statistics for this 4-Bao scheme, even though there is no phase bias in the constellation. That is because only a subset of the SCs are being modulated with data carrying complex constellation points and the SC allocation strategy determines which are the data carrying SCs and which are the null SCs. As we can see, there
is a nearly 3 dB performance difference between 4-Bao with regular scheme and the diversity scheme for “1001” allocation in Fig. 6.5a and Fig. 6.5c, respectively. From the existing PAPR literature on OFDM systems, it is known that swapping the location of data carrying and reserved null SCs can lead to significant PAPR reduction [192,193]. It is the same observation we make here, except that the null SCs are determined by the SCMA codebook and the associated SC placement strategy. This highlights the fact that we can attain a better PAPR performance for any SCMA constellation through the SC placement strategy.

It is clear from the results in Fig. 6.5 that we should adopt a SC allocation strategy that shuffles around the SCs if we have a modulation scheme with the phase bias of having constellation points of the same phase, while we should use a contiguous SC allocation strategy for a scheme that has a phase bias of having constellation points of opposite phase. For the 4-LDS and 4-Bao schemes, which are schemes from the existing SCMA literature, we summarize the observations for these constellations in Fig. 6.6. We see that for a desired SCMA scheme, an appropriate SC allocation strategy based on the phase bias in the chosen SCMA scheme can be selected to reduce the PAPR or even vice-versa, i.e., for a desired SC allocation strategy, an SCMA scheme with favourable phase bias characteristics to reduce the PAPR can be selected. These results are important because the PAPR gains achieved between different configurations are a result of static configuration parameters, i.e., they are configured one-time on setup and come with no additional computational overhead. This is made possible by the SCMA codebook that introduces dependency between the transmitted SCs in the system. These statistical dependencies can be exploited to achieve PAPR reduction in a static manner, not possible in traditional OFDM systems that individually modulate the SCs. We discuss these opportunities for PAPR reduction in detail next in Section 6.6.

6.6 Exploiting Statistical Dependency in PAPR Reduction Schemes Based on Multiple Signalling

In this section, we investigate how the novel aspects to the analysis of PAPR statistics in SCMA-OFDM systems, discussed in Section 6.3.2, impact the class of PAPR
Figure 6.6: Comparing the PAPR statistics with different choices of SCMA modulation schemes and SC allocation strategies. These options are static configuration parameters that impact the PAPR statistics in SCMA-OFDM systems.

reduction techniques based on multiple signalling and probabilistic techniques [162]. The general idea with these PAPR reduction techniques is to generate a set of candidate signals every OFDM symbol and transmit the signal with the least PAPR. These techniques are information lossless, since they do not distort the transmitted signal. However, they come with the complexity overhead of generating the set of candidate signals every OFDM symbol as opposed to just one signal. They also incur a throughput loss due to the need to transmit sidelink information, not ideal for overloaded NOMA systems. When these PAPR reduction techniques are used in traditional OFDM systems, since each SC is independently modulated, there is no advance knowledge of any statistical dependencies between the SCs to exploit. Hence, the set of candidate signals can only be generated after the information sequence in that OFDM symbol is known. However, with SCMA-OFDM systems, the statistical dependency between the transmitted SCs can be exploited for PAPR reduction in conjunction with these well established techniques. We show that for a given level of PAPR reduction, the overhead incurred by these multiple signalling techniques can be reduced or even eliminated in some scenarios.

The PAPR reduction techniques described in [162] under the class of multiple signalling and probabilistic techniques all assume that the SCs in the system are independently modulated with QAM symbols. With SCMA-OFDM systems, some of these techniques can be applied with some modifications to satisfy the SCMA constraints while some techniques cannot be easily extended to the SCMA-OFDM
Figure 6.7: Block diagrams highlighting the PAPR reduction techniques of SLM and IL that are described in the context of SCMA-OFDM systems.

paradigm. For example, techniques that involve constellation shaping [184,185] or tone injection cannot easily lend itself to SCMA systems because it affects the SCMA constellation design. SCMA constellations are designed with a number of criteria [159] that will be affected by the constellation shaping and is beyond the scope of the discussion here. Similarly, techniques that involve using null SCs such as tone reservation [194] or the dynamic swapping of data and null SCs [192,193] is difficult to extend to the SCMA-OFDM paradigm. This is because the null SCs are an integral part of every SCMA block and cannot be rearranged randomly for PAPR reduction purposes. However, other multiple signalling techniques such as selective mapping (SLM) [182], partial transmit sequences (PTS) [186] and interleaving (IL) [183] can be tailored to meet the constraints of an SCMA-OFDM system. The block diagrams for these three techniques are depicted in Fig. 6.7 and in the discussion that follows, we focus on how they can be adapted to SCMA-based systems.

In the SLM technique used in traditional OFDM systems, a set of candidate OFDM symbols are generated that represent exactly the same information. The signal with the least PAPR is then transmitted. The set of candidate signals is generated by multiplying the original data carried in the SCs for that symbol with $R$ different sets of phase factors

\[ b_m = \left[ b^0_m, b^1_m, \ldots, b^{Z-1}_m \right], \quad 0 \leq m \leq R - 1, \]

\[ b^n_m = e^{j \theta^n_m}, \quad 0 \leq n \leq Z - 1. \]  

(6.9)

After the inverse discrete Fourier transform (IDFT) block, this multiplication generates $R$ sequences in time domain and the one with the least PAPR is transmitted.
Sidelink information about the phase factor is sent to the receiver to indicate which set of phase sequences were used, so that the receiver can undo the multiplication and regenerate the original data. The side link information is \( \log_2 R \) bits long, since we only have to identify which sequence was used. The set of possible sequences are known to both the transmitter and receiver. Additionally, there is significant complexity introduced by the extra IDFT operations every symbol that scales linearly with \( R \) [162].

SLM can be applied to SCMA-OFDM systems because the typical SCMA constellation design process used in the literature allows for random user-specific rotations to be performed without affecting the error rate performance in the uplink [159]. While it is a sub-optimal approach to SCMA constellation design to find the mother constellation and user-specific rotations separately, it is by far the most widely used approach in the literature [159]. Further, as shown in [159], in the uplink, the user-specific rotations designed as part of the constellation design process lose meaning due to the fact that different users experience different fading channels. As a result, the user-specific rotations in the UL SCMA systems can be designed for PAPR reduction purposes instead. These random user specific rotations translate to a random phase being multiplied to the \( d_v \) dimensions of the SCMA codeword. However, the \( d_v \) dimensions of the SCMA codeword cannot each be multiplied by their own phase factor, as doing so would destruct the SCMA. Hence, this additional constraint needs to be placed when generating the set of phase factors that make up the phase sequences. In other words, the set of candidate phase sequences should be generated such that each set of \( d_v \) SCs is assigned a phase factor, rather than each SC being assigned its own phase factor. The set of phase sequences generated for a Z-SC OFDM system in (6.9) needs to then be modified to only generate a list of \( L \) phase factors. These phase factors are multiplied by the original data sequence to generate a set of \( R \) different OFDM symbols, \( y_m, \forall m = 0, \ldots, R - 1 \), and the signal with the least PAPR is transmitted as follows:

\[
y_{\hat{m}} = m \text{ PAPR}(y_m), 0 \leq m \leq R - 1.
\]

In Fig. 6.8, we run MATLAB simulations for the 4-Bao scheme with both the regular and diversity-based SC allocation schemes. We run with \( R = 2 \) and \( R = 4 \), which corresponds to one and two bits of additional sidelink information respectively. With
Figure 6.8: Comparing the PAPR reduction achieved with SLM in an SCMA-OFDM system with 4-Bao constellation and different SC allocation strategies and different values of $R$ for the SLM reduction.

$M = 4$, that corresponds to one SCMA block of transmission reserved for sidelink information. This means there is a throughput loss from $L$ to $L - 1$ modulation symbols per OFDM symbol duration. Additionally, there is a computational complexity overhead that is higher when $R = 4$ compared to when $R = 2$. We can see that for this 4-Bao SCMA constellation, the PAPR reduction achieved with $R = 2$ for the diversity scheme is the same as that achieved with no PAPR reduction using the regular allocation. Similarly, the PAPR reduction achieved with $R = 4$ in the diversity scheme is achieved with $R = 2$ using the regular scheme. The key takeaway message here is that the statistical dependency introduced by the SCMA codebook between certain SCs transmitted in an OFDM symbol can be exploited to achieve PAPR reduction gains through the setting of static configuration parameters such as the SCMA constellation scheme, SC allocation strategy, SCMA block dimensions like $N$ and $d_v$ etc. Such gains are not possible in traditional OFDM systems where the SCs are independently modulated and so there is no advance knowledge of the statistics to exploit.

Another multiple signalling technique called partial transmit sequences (PTS) follows a similar idea to SLM. In PTS, the $Z$ SCs are divided into disjoint sub-blocks and the IDFT of each block is taken. Different phase factors are multiplied to these IDFT outputs, i.e., to the time-domain data, and once again the OFDM signal with the least PAPR is transmitted. When applied to SCMA-OFDM systems, as long as every $d_v$ SCs that make up an SCMA block are contained in the same sub-block,
none of the SCMA related constraints are violated. Hence, a logical split for these disjoint sub-blocks would be along the SCMA blocks. The computational overhead involved to generate the candidate signals scales with the number of sub-blocks and is larger than that incurred with SLM [162]. Like with SLM, there is also the sidelink information required which results in a throughput penalty. Simulation results for PTS are not shown here as they are very similar to the observations from SLM, where the static configuration gains from the choice of SC allocation strategy compares with the PAPR reduction gains from PTS when $R$ is small.

**Figure 6.9:** Comparing the PAPR reduction achieved with interleaving in an SCMA-OFDM system with 4-Bao constellation and different SC allocation strategies.

Interleaving is another probabilistic technique for PAPR reduction commonly used in OFDM systems. In interleaving, the idea is once again to create a set of data block candidates and select the block with the least PAPR to transmit. Compared to SLM, in this method, an interleaver block is used instead of phase sequences. Interleaver is a device which reorders the entries of a block of length $Z$ in a specific order. Similar to SLM and PTS, there is overhead in computing the IFFT of the different interleaved sequences and also the receiver needs sidelink information to de-interleave the received data block. When applied to SCMA-OFDM systems, we can only interleave the $L$ modulation symbols with each other. In other words, there are $L$ modulation symbols to be transmitted in an OFDM symbol, and they can be transmitted on any of the $L$ SCMA blocks. However, the $d_v$ SCs within a block that contains the codeword for a modulation symbol must remain within the same SCMA block and the interleaver cannot reorder these SCs. With these constraints, even after applying an interleaver
block, the PAPR statistics are still subject to the joint effect of the modulation scheme phase bias and SC placement that was described in Section 6.3.2. This is illustrated by the simulation results in Fig. 6.9 where the gains attained by interleaving in some configurations are small and do not compare with the PAPR reduction gains from interleaving in traditional OFDM literature [183].

6.7 Appendix: Zero dB PAPR Constellation Design

A zero dB constellation PAPR means zero dB PAPR for the transmission of one modulation symbol, i.e., $L = 1$. Such a constellation scheme requires each modulation symbol to be coded on only one of the $d_v$ SCs. The $N$ dimensional codewords already allocate $N - d_v$ null dimensions, but the zero dB PAPR constraint means that out of the remaining $d_v$ dimensions, only one can be non-zero. Further, the magnitude of each of these symbol points should also be constant. For an $M$-point constellation, this translates to having $N_d = M/d_v$ points per dimension evenly spaced around a circle. To normalize the energy of the constellation, this needs to be the unit circle. Therefore, we have an $N_d$-PSK constellation shape in the $d_v$ dimensions. The $M$ symbols are then assigned such that each symbol gets exactly one non-zero constellation point in any one dimension. However, which symbol is assigned to which constellation point is not important from a PAPR perspective. Algorithm 7 illustrates the construction of this constellation, that we call $M$-0PAPR because of the zero dB constellation PAPR property. The 4-point and 16-point version, namely 4-0PAPR and 16-0PAPR, are shown in Fig. 6.10a and 6.10b, respectively.

---

**Algorithm 7: $M$-0PAPR SCMA scheme**

<table>
<thead>
<tr>
<th>$M$</th>
<th>Number of points in constellation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_v$</td>
<td>Number of non-zero dimensions from $S$</td>
</tr>
</tbody>
</table>

**Step-1** Each RE is allocated $M/d_v$ points evenly spaced around the unit circle;

**Step-2** Each of the $M$ points is labelled uniformly so that it is non-zero in only one of the dimensions;

**Output:** A $d_v$-dimensional, $M$-point constellation having 0 dB constellation PAPR since each symbol is non-zero in only one dimension.
Figure 6.10: Proposed M-0PAPR scheme that has zero dB constellation PAPR.
Chapter 7

Summary and Future Research Directions

This chapter contains a brief summary of the key findings and contributions for the thesis in Section 7.1 followed by some insights into extending the work from this thesis in future research in Section 7.2.

7.1 Summary

This thesis addressed several practical challenges to realizing NOMA (PD-NOMA and CD-NOMA) systems in practical deployments. To start with, in PD-NOMA systems, we considered the impact on the user clustering and ordering schemes in ABF mmWave-NOMA systems where each user is modelled with their own SIC decoding capabilities. This is an important practical consideration as 5G systems are envisioned to support a wide variety of users that each have vastly different processing capabilities. As a result, NOMA ordering schemes cannot simply assume all users can be placed arbitrarily in a NOMA cluster and each user can be expected to process the same amount of SIC decoding computations as other users. This aspect was considered from three different perspectives in this thesis as outlined next.

Firstly, in Chapter 3, we proposed low-complexity heuristics to solve the joint user clustering and user ordering scheme in ABF mmWave-NOMA system that can serve a set of users that each have their own SIC decoding capability constraints. By using the reported SIC decoding capability constraint from each user to set the maximum position in the SIC decoding order for clusters that are always decoded in the order of their effective channel gains, we framed the problem as a minimum exact cover optimization problem. Also, for a homogeneous system where all users have the same decoding capability requirements, we showed that this boils down to
a simpler condition of restricting the number of users per cluster and proposed a simpler NOMA-BB algorithm for such homogeneous systems.

Secondly, in Chapter 4, we motivated the need for solutions other than optimization techniques for PD-NOMA clustering decisions considering end-user SIC decoding capabilities in live networks due to the large number of computation steps that need to run very often. Even low complexity heuristics like NOMA-MEC and NOMA-BB are run based on the instantaneous channel conditions of the users and so involve a large number of computation steps that need to be run each time a single users channel changes, i.e., almost on a millisecond granularity. Thus, in Chapter 4, the thesis proposed a computationally efficient two-stage machine learning based approach using neural networks to solve the same user clustering and ordering problem. The proposed neural networks, called ANN-NOMA-MEC and ANN-NOMA-BB, are trained offline on datasets generated from running simulated settings using the NOMA-MEC and NOMA-BB heuristics. By training the neural network offline, the heavy computational steps are executed away from the BS compute resources in this approach. We showed through simulation results that ANN-NOMA-MEC and ANN-NOMA-BB performed comparably with the NOMA-MEC and NOMA-BB heuristics.

Thirdly, the thesis addressed the challenges associated with the availability and reliability of the CSI used to operate these user clustering schemes. In Chapter 5, we motivated three different dimensions of feedback that a network can benefit from to solve the user clustering problem, categorized as CSI-based feedback and non-CSI-based feedback, where the non-CSI-based feedback is comprised of UE location and a camera feed obtained from a camera equipped BS (CBS). We showed how the vision assistance of a CBS can be used in conjunction with other dimensions of feedback to make clustering decisions in various scenarios. Exploiting the advances in deep learning, the NOMA clustering problem was addressed using the images captured by CBSs, without consuming extra RF resources. We showed that such an approach achieves comparable performance to a CSI-based approach, and user clustering can continue to function without much performance loss even in the scenarios where CSI is severely outdated or not available at all.

Finally, the thesis addressed the PAPR problem in CD-NOMA systems employing SCMA for uplink (UL) communications; an important practical consideration to the deployment of UL NOMA systems in the future. We showed that the PAPR statistics for high-rate users in SCMA-OFDM systems are influenced by two main factors, the
SCMA modulation scheme and the placement of the SC’s that carry the SCMA codewords. The PAPR statistics of different SCMA modulation schemes were exploited along with their accompanying resource allocation strategies to achieve PAPR reduction gains in novel ways that would not apply to conventional OFDM systems. Concretely, simulation results showed that the joint impact of the phase bias in the SCMA modulation scheme along with the placement of the SC’s can lead to a PAPR performance difference of up to four dB.

7.2 Future Research Directions

In this section, we highlight the different ways in which the research presented in this thesis can be extended.

7.2.1 User Clustering in PD-NOMA Considering SIC Decoding Capabilities in Networks Using Conventional BSs without Camera

Some possible future research directions related to the user clustering and ordering schemes presented in Chapters 3 and 4 are:

- To model the battery life of the user as part of the SIC decoding capability. It was assumed in the channel model in these sections that the SIC decoding capability of a user is fixed and never changes. However, even higher end devices might want to advertise a lower SIC processing capability when their battery life is low; while accepting a higher amount of SIC decoding work when their battery life is higher.

- To consider user clustering problems in digital beamforming systems where multiple clusters can be formed and served at the same time, leading to inter-cluster interference that needs to be mitigated. Both heuristics and machine learning based approaches can be considered.

  - For ML based approaches, it would be interesting to learn if artificial neural networks can also learn the underlying patterns involved in cancelling out the inter-cluster interference on top of the SIC decoding capability constraints as considered in this thesis.
Jointly considering the NOMA user clustering along with the SC placement strategy in an OFDM system employing digital beamforming while still considering the individual SIC decoding capability of the users is also worth exploring further.

- While the individual SIC decoding capability of each user was modelled, it was assumed that each user requires the same minimum rate, i.e., quality of service (QOS). In reality, users with lower processing capabilities likely need a lower data rate than high-end users. Modelling this aspect into the problem is an area worthy of future research.

- To investigate a reinforcement learning approach to solve the same user clustering and assignment problem incorporating user decoding capabilities and compare it against the ANN based approach proposed in Chapter 4.

### 7.2.2 User Clustering in PD-NOMA Considering SIC Decoding Capabilities in Networks Using CBSs Exploiting non-CSI Feedback

Several challenges around using CBSs were highlighted in Chapter 5 that form the base of possible future research work in this area as follows:

- To investigate the approach in multi-path settings where learning the best beam from a set of candidate beams could involve more advanced features, since users themselves can be obstacles and alter the multi-path setting and so the best beam of the users. This potentially requires modelling deeper neural networks and needs more training data.

- To leverage the visual information captured by the CBS to predict user distribution in advance and use it to predict optimal clusters of users for NOMA. This can help overcome challenges related to frequency of camera updates and address high mobility users in the system.

- To use object detection techniques such as [155] for user identification to complement user location feedback in settings where the location feedback is not precise.
• To weigh the tradeoff between better camera coverage to aid with the NOMA clustering decisions with the additional cost incurred to have this extra camera coverage.

7.2.3 PAPR Analysis and Reduction in SCMA-OFDM Systems

Some possible future research directions related to the PAPR reduction problem presented in Chapter 6 are:

• To capture the statistical dependencies between the transmitted sub-carriers (SCs) in an SCMA-OFDM system via a mathematical model. For example, a metric to capture the level of phase bias in an SCMA-OFDM constellation can be derived. In this way, the SCMA constellation design process can aim to maximize this metric, in order to be exploited later with the appropriate SC allocation scheme for PAPR reduction.

• To further study the impact of SCMA configuration parameters like $d_v$ and $N$ and how they can be exploited by PAPR reduction techniques. One would expect the level of statistical dependency to grow as $d_v$ increases. PAPR reduction techniques like SLM and PTS can also be enhanced to exploit these statistical dependencies. For small values of $L$, there are a limited number of possible sequences and so all possible combinations can be tried beforehand to find the favourable sequences. An interleaver algorithm can then be developed to quickly match a favourable sequence from a PAPR perspective with minimum computational overhead.

• Cross-layer systematization problems around the user and SC allocation schemes that include PAPR considerations is an interesting extension of the work presented in this thesis. For example, in [159, 195, 196], frequency diversity gains in terms of error rate performance were demonstrated. This could be coupled with selecting a scheme like LDS that has the phase bias for accompanying PAPR gains for high-rate users.

• SC allocation schemes can also factor in the PAPR tolerance among the users in the system. For instance, the thesis showed that with the frequency diversity
scheme, some users are offered more SC separation than others, offers better PAPR reduction to some users compared to others. Users with the least PAPR tolerance can be assigned the most favourable SC separation and so on; an important aspect in considering end-user capabilities in UL SCMA systems.
List of References


[38] Y. Sun, J. Zhou, Q. Cao, and S. Li, “Precoder design in statistical CSI aided non-orthogonal multiple access,” *IEEE Access*, vol. 6, pp. 16484–16492, 2018.


[130] K. Higuchi and A. Benjebbour, “Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access,” *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, Mar. 2015.


