Operating multi-user transmission for 5G and beyond cellular systems

by

Abdalla Hussein

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Ekram Hossain
Professor, Dept. of Electrical and Computer Engineering,
University of Manitoba

Supervisors: Catherine Rosenberg
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Patrick Mitran
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal Members: Ravi Mazumdar
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Mahesh Tripunitara
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal-External Member: Lukasz Golab
Professor, Dept. of Management Sciences,
University of Waterloo

Author's Declaration

Abstract

Every decade, a new generation of cellular networks is released to keep up with the ever-growing demand for data and use cases. Traditionally, cellular networks rely on partitioning radio resources into a set of physical resource blocks (PRBs). Each PRB is used by the base-station to transmit exclusively to one user, which is referred to as single-user transmission. Recently, multi-user transmission has been introduced to enable the base-station to simultaneously serve multiple users using the same PRB. While multi-user transmission can be much more efficient than its single-user counterpart, it is significantly more challenging to operate. Thus, in this thesis we study the operation, i.e., the Radio Resource Management (RRM), for two popular multi-user transmission technologies; namely, 1) NOMA (Non-Orthogonal Multiple Access) and 2) Multi-User Multiple-Input Multiple-Output (MU-MIMO).

For NOMA RRM, we study a multi-cell, multi-carrier downlink system. First, we formulate and solve a centralized proportional fair scheduling genie problem that jointly performs user selection, power allocation and power distribution, and MCS (Modulation and Coding Scheme) selection. While such a centralized schedule is practically infeasible, it upper bounds the achievable performance. Then, we propose a simple static coordinated power allocation scheme across all cells for NOMA using a simple power map that is easily calibrated offline. We find that using a simple static coordinated power allocation scheme improves performance by 80% compared to equal power allocation. Finally, we focus on online network operation and study practical schedulers that perform user-selection, power distribution, and MCS selection. We propose a family of practical scheduling algorithms, each of them exhibiting a different trade-off between complexity (i.e., run-time) and performance. The one we selected sacrifices a maximum of 10% performance while reducing the computation time by a factor of 45 with respect to the optimal user scheduler.

For MU-MIMO RRM, we focus on the study of the downlink of an OFDMA massive MU-MIMO single cell assuming ZFT (Zero Forcing Transmission) precoding. An offline study is initiated with the goal of finding the best achievable performance by jointly optimizing user-selection, power distribution and MCS selection. The best performance is analyzed by using both BRB (Branch-Reduce-and-Bound) global optimization technique for upper-bounding the achievable performance and a set of different greedy searches for lower bounding the achievable performance to find good feasible solutions. The results suggest that a specific search strategy referred to as greedy-down-all-the-way (GDAW) with full-drop (FD) is quasi-optimal. Afterwards, we design a simple practical scheduler that achieves 97% of the performance to GDAW with FD and has comparable runtime to that of the state of the art benchmark that selects all users, performs ZFT precoding followed by power distribution using water-filling. The proposed scheme performs a simple round robin grouping to select users, followed by ZFT precoding and joint power distribution and MCS selection via a novel greedy algorithm with a possible additional iteration to take zero-rate users into account. Our solution outperforms the benchmark by 281%.

Acknowledgements

I would like to express my gratitude to Prof. Catherine Rosenberg and Prof. Patrick Mitran for their ongoing support in both my academic and personal lives at Waterloo. I would not have been able to finish my thesis without their help and supervision. I am glad and proud to be under their supervision.

I want to thank the members of my PhD advising committee for taking the time to study my thesis and providing helpful feedback that helped enhance its quality. Also, I would like to especially thank Prof. Mazumdar for his advice and feedback throughout my years of study in Waterloo. Moreover, I would like to than professor Tripunitara for helping me on my first day teaching.

I would also like to thank my father and mother who never stopped loving and supporting me. I feel lucky every day for having such lovely parents like you. Words are never enough to tell how much I love you.

Of course, I am grateful to my siblings, who are my closest friends. I am grateful for your continued love, encouragement and support that helped me through out my journey. Our wonderful memories growing up has helped during the hardships I faced during my study and through the times of the pandemic.

I am grateful to my friends here in Waterloo for their support since the first day I arrived to Canada. Without their help, it would be extremely difficult for me to adjust to the new life here and to get through the pandemic.

Finally, I would like to thank all my relatives and my friends from Egypt and Canada for their support during my PhD studies. All of you have helped make me a better person, a stronger and a more positive person. You have given me strength and strength to pursue a PhD and I thank you for all the support and love that you give me every single day. Thank you my dear ones.

I also want to say thank you to my lab mates for all the fun we had together and the useful discussion we had across the years. These moments are dear memories that I greatly cherish and I hope you enjoyed my company as much as I did yours.

Dedication

This is dedicated to my family and loved ones.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

4G Fourth Generation 1

5G Fifth Generation 1, 2

BRB Branch-Reduce-and-Bound iv, 76, 77, 79, 82, 102, 103

DC Difference of Convex 18, 19

DL Downlink 1, 13, 19–21

eMBB Enhanced Mobile Broadband 1

ICIC Inter-Cell Interference Coordination 13

ILP Integer Linear Program 75

IoT Internet of Things 1, 12

KKT Karush–Kuhn–Tucker 19

LTE Long Term Evolution 1, 7

MCS Modulation and Coding Scheme iv, 3, 5, 13, 14, 19, 20, 22, 30, 108

MIMO Multiple Input - Multiple Output 2, 4, 7, 8, 61–63

MINLP Mixed Integer Non-Linear Program 24

MU-MIMO Multi-User MIMO 7–10, 53, 63, 68, 106

MUT Multi-User Transmission 2, 4, 61, 106, 108

NOMA Non-Orthogonal Multiple Access iv, xii, 2, 4–6, 10, 12–15, 18–22, 24, 29–33, 40, 41, 106

NR New Radio xiii, 1, 7, 59

OFDMA Orthogonal Frequency Division Multiple Access 1–3, 14, 19, 58, 71

OMA Orthogonal Multiple Access xii, 12–14, 18–20, 22, 29, 30, 32, 41

PRB Physical Resource Block 1–6, 12–15, 18, 20–24, 31, 33, 63, 68, 106

QoS Quality of Service 18

RF Radio Frequency 1

RRM Radio Resource Management 2, 4, 8, 10, 53

SIC Successive Interference Cancellation 2, 5, 6, 12–14, 17, 21, 22, 24, 30, 33, 40

SINR Signal-to-Noise and Interference Ratio 3, 5, 6, 14, 20, 22, 33, 104, 106, 107

SISO Single Input - Single Output 58, 106

SU-MIMO Single-User MIMO 7, 8, 63

SUT Single User Transmission 1, 106

UE User Equipment 2

UEs User Equipment 13

UL Uplink 1

ZFT Zero Forcing Transmission iv, 8, 9, 53–55, 68, 107

Chapter 1

Introduction

5G (Fifth Generation) and beyond cellular networks are designed to provide unprecedented network capacity. This is due to the need for ever higher data rates for eMBB (Enhanced Mobile Broadband) applications as well as the proliferation of connected devices brought by the IoT (Internet of Things). These networks must therefore provide higher system-level spectral efficiency than that provided by their 4G (Fourth Generation) counterparts.

A base station typically has access to a variety of communication resources, including RF bandwidth, transmission power, and antennas. The key to improving spectral efficiency lies in smart network operation, efficient resource allocation, and the use of advanced physical layer technologies.

The underlying physical layer structure has a significant impact on network operation. Up to 4G, the cellular network depended on SUT (Single User Transmission) at the physical layer. In SUT, network resources are divided into orthogonal units and each unit is used to serve a single user at a time. As an example, 4G LTE and 5G NR are based on OFDMA (Orthogonal Frequency Division Multiple Access) where the available RF band is divided into sub-channels and time is divided into slots. In the DL (downlink)[1] case, a single sub-channel for one time-slot is referred to as PRB (Physical Resource Block). In SUT systems, a PRB is used by the base-station to exclusively serve (i.e., transmit to) a single user.

Although SUT-based networks are simple to operate, they suffer from a major shortcoming. Information theory shows that capacity achieving transmission schemes should

[1] The same applies for communication in the UL (Uplink). However, we focus solely on DL OFDMA cellular networks in this thesis.

rely on MUT (Multi-User Transmission)[2]. In multi-user transmission based networks, a single PRB is used to simultaneously transmit to several users at the same time. This generates additional inter-user interference. However, when the network is operated properly, multi-user transmission can significantly outperform single-user transmission. Thus, MUT is one of the promising technologies for 5G and beyond cellular systems.

For implementing downlink multi-user transmission, there are two common types of physical layer technologies: 1) NOMA (Non-Orthogonal Multiple Access) and, 2) MIMO (Multiple Input - Multiple Output) transmission. In NOMA, MUT is implemented by superimposing the transmitted messages, typically in the power domain or the code-domain. At the receiver side, SIC (Successive Interference Cancellation) is used to remove unwanted weaker signals [5]. NOMA does not require multiple antennas at either the base-station or the UE; however, it needs complex SIC receivers.

On the other hand, MIMO allows the base station to generate multiple beams for the same PRB, each of which carries a signal dedicated to a particular user. Furthermore, massive MIMO takes MIMO to the next level by increasing the number of antennas dramatically which drastically increases the overall data rate. As a result, massive MIMO is an attractive approach.

Although information theory gives us valuable tools for establishing upper bounds on the total system throughput, this is usually done assuming impractical channel coding schemes. In reality however, a finite set of practical Modulation and Coding Schemes (MCS) is used. Moreover, information theory does not help us design practical RRM (Radio Resource Management) solutions.

This chapter is structured as follows. The RRM problem is first discussed in the context of traditional Single Input Single Output (SISO) downlink OFDMA systems and is followed by an overview of the multi-user transmission specific RRM characteristics. Following this, we go over the key background as well as the RRM problem for both NOMA and MIMO. Finally, we outline the main research questions and the thesis structure. Note that there is a more detailed background section for NOMA in Chapter 2 and for MU-MIMO in Chapter 3.

1.1 SISO RRM Problem

The RRM (Radio Resource Management) problem can be formulated as a joint optimization problem where the base-station jointly optimizes a set of decisions to maximize a

[2] This is true for both single antenna and multi-antenna systems [2, 3, 4].

certain fairness criteria, e.g. proportional fairness. In SISO OFDMA RRM, the base-station station has to allocate power to each PRB. This is referred to as *power allocation*. Then, it has to select a user to be served in each PRB. This is referred to as *user selection*. Finally, an estimate of the SINR (Signal-to-Noise and Interference Ratio) for the selected user is computed given the channel state information for that user, and, based on that SINR, a MCS (Modulation and Coding Scheme) is selected for transmitting to the selected user. Although MCS selection seems like a trivial step, it is almost always done in the literature with an approximated rate function (e.g., Shannon) as opposed to with the exact piecewise constant rate function corresponding to a small number of MCSs which has a significant impact on the results as will be discussed extensively in this thesis.

While assigning different radio resources, the base-station seeks to optimise total throughput while preserving fairness, i.e., it tries to ensure that no one is denied data service for a prolonged period. Typically, the base-station's goal is to enforce proportional fairness [6], which is defined as finding the set of rates in which no user can unilaterally increase their own rate without reducing another user's rate by at least the same percentage. This means that a user in good radio-conditions will get a high data rate and a cell-edge user will not be deprived. Although there are other proposed fairness criteria, operators typically prefer proportional fairness and thus, for the remainder of this thesis we consider proportional fairness in rates as our objective.

In cellular networks, the SINR estimate is a critical metric used in downlink scheduling to assess the quality of the communication link between the base-station and the UE. It is defined as the ratio of the received signal power to the summation of the interference and noise power at the receiver. The base-station uses the SINR estimate to determine the best MCS to efficiently transmit to each UE. To calculate the SINR estimate accurately, it is important to have accurate channel state information. Channel estimation in cellular systems is typically pilot-based. In pilot-based channel estimation, the base-station periodically transmits known pilot data symbols for channel estimation. The UE receives these pilot symbols and uses them to estimate the channel characteristics. The channel estimation process is performed periodically as the channel accuracy is critical for the cellular system performance. In this thesis we focus on analyzing MUT performance assuming perfect channel state information and leave the study of the system performance under imperfect channel estimation to future work.

1.2 RRM for MUT networks

With multi-user transmission, the RRM problem becomes much more complicated. In order to ease the discussion of RRM in MUT-based systems, we start by defining the main RRM procedures, that should be jointly optimized, then we go into the specifics for each MUT technology.

Definition 1.2.1 (Power allocation)**.** The process by which the base-station determines the total power to be allocated to a given PRB.

Power allocation is a process that is performed for both single-user transmission and multi-user transmission-based cellular networks. The defacto approach is to allocate power equally across the available sub-channels. Although this approach is simple and good for single-cell networks, it is sub-optimal in multi-cell network as will be shown in Chapter 2 for multi-cell NOMA downlink systems [7]. Power allocation optimization gain for single-user transmission systems was studied in [8].

For a given power allocation, the RRM problem can be approximated as a sequence of smaller per PRB joint optimization problems. Each per PRB problem includes the following procedures[3] and possibly others specific to the MUT technology (e.g., precoding for MIMO):

Definition 1.2.2 (User selection)**.** The process by which the base-station selects the set of users to be allocated a certain PRB.

One of the main differences between single-user transmission and multi-user transmission based cellular networks lies in the user-selection process. In single-user transmission, the process is simpler since the base-station has to only select one user to be scheduled in each PRB. On the other hand, in multi-user transmission a subset of users has to be scheduled to simultaneously transmit on the same PRB. Furthermore, depending on the multi-user transmission technology, i.e., NOMA or MIMO, there could be constraints on the selected user-set.

Definition 1.2.3 (Power distribution)**.** Another major difference between SUT and MUT is the need to distribute the power allocated to a PRB among the selected users in that PRB. This is what we call power distribution.

[3] As channel conditions remain almost the same across a range of PRBs, denoted as a coherence block, the base-station could reuse the same decisions for a subset of PRBs within that range. Although this approach reduces computational complexity, it also reduces the overall performance and thus in this thesis we only consider per PRB decisions.

Power distribution is performed to distribute the available power among the selected users. While computing the power distribution can be performed simply by equally distributing power among scheduled users, this yields poor performance. Instead, power distribution is performed through an optimization problem or an algorithm derived from it and that is when the piecewise rate function is replaced by a smooth approximation which is easier to handle in that context. A typical approximation is the Shannon capacity formula. As shown in the following chapters, power distribution is essential for enforcing fairness among the scheduled users as well as maximizing network performance.

Definition 1.2.4 (MCS selection)**.** The process by which the base-station selects the MCS to be used for communicating with each scheduled user on a given PRB based on the estimated SINR.

As previously mentioned, the RRM problem is equivalent to a joint optimization problem where the base-station optimizes the different RRM procedures to maximize the target fairness criteria. The MCS selection procedure is nearly always overlooked in the literature, despite the fact that it is crucial. Assuming that the relation between the received rate and the experienced SINR (Signal-to-Noise and Interference Ratio) is dictated by Shannon's rate function not only produces optimistic results, but it can also lead to inaccurate conclusions. Although, this assumption is typically performed to simplify the underlying optimization problem, the actual relationship between the SINR a user experiences and the rate it receives due to the finite number of MCS is defined by a piece-wise constant rate function. This is opposed to the smooth increasing $\log(1 + SINR)$ Shannon rate function and provides a crucial distinction especially at low SINR since a practical system cannot transmit successfully to a user that sees an SINR lower than that required for decoding the lowest rate MCS.

1.3 NOMA Background and RRM

NOMA (Non-Orthogonal Multiple Access) is one of the most recent advances to be considered for 5G cellular systems [9]. NOMA allows for multi-user transmission using a single antenna at the BS and at the UE, i.e., it allows multiple users to be simultaneously allocated to the same PRB. This is implemented by using superposition coding at the base-station and SIC at the receiver.

In NOMA, a single-antenna base station can transmit to multiple users in the same band and at the same time, with each user assigned a certain power. In other words, the base-station superimposes different messages in the power domain creating a power division multiple-access scheme. Alternatively, NOMA could superimpose messages using non-orthogonal codes which is referred to as code-domain NOMA.

SIC is used at the receiver to remove some of the unwanted signals at the receiver. For a UE to be able to remove an interfering signal, it needs to have at least the same SINR as the intended receiver. This is described by saying that a user A is at least as strong as user B and thus user A can use SIC to remove the interference caused by the message transmitted to user B.

For NOMA DL OFDMA cellular operation, RRM is composed of 4 steps; namely, 1) power allocation, 2) user-selection, 3) power distribution and 4) MCS selection. As mentioned in Section 1.2, power allocation is defined similarly for NOMA as well as traditional OFDMA systems. Although equal power allocation is the defacto approach, simple power allocation optimization can bring significant gains as shown in [8] for SISO systems as well as in Chapter 2 for NOMA (and published in [7]).

User-selection for NOMA is challenging because not only does the base-station have more options, but there are also various restrictions on user-set selection imposed by hardware limitations. Because NOMA requires the usage of SIC receivers, each selected UE must notify the base-station as to the maximum number of SIC operations it can perform. This is because it does not make sense to select users to be jointly scheduled on the same PRB if they cannot receive the sent messages due to the lack of hardware capability.

Additionally, with NOMA-N, the base-station selects an *ordered* set of N users $(i_1, ..., i_N)$ on a per PRB basis. The users are selected such that user i_n can use SIC to remove the interference caused by $i_{n+1}, ..., i_N$ and then user i_n decodes its message under the interference caused by the transmissions to users $i_1, ..., i_{n-1}$. This constrains which users can be selected together since not all users selections can satisfy this condition.

Power distribution for NOMA is tricky and typically confused with power allocation in the literature. Since with NOMA, the UE uses SIC to cancel the effect of some signals, precise power distribution is required to guarantee that the message is successfully received despite the presence of residual interference from signals not removed with SIC.

Finally, depending on the decoding order as well as the power allocated to each PRB and the power distributed to each selected user, the base-station selects a MCS to be used for transmission. This final step is critical and has a significant impact on performance as well as conclusions as will be studied extensively in Chapter 2.

1.4 MIMO Background and RRM

A MIMO system, as first proposed by Winters [10], uses multiple antenna elements for transmission and reception to boost wireless connection robustness and improve spectral efficiency capacity. The development of practical MIMO techniques over the last 20 years has been one of the reasons for the success of 4G/LTE (Long Term Evolution) and 5G/NR (New Radio).

In a MIMO system, the antenna elements can be used for *diversity* or *spatial multiplexing*. In diversity MIMO, the transmitter sends multiple copies of the message using its antennas, and the receiver decodes the message using all of its antennas. This method increases the reliability of a wireless connection by providing redundancy in the form of numerous copies of the same message, which reduces the likelihood of a failure. On the other hand, spatial multiplexing allows for a direct increase in link capacity by transmitting multiple data streams at the same time.

To exchange multiple data streams at the same time, the base-station employs linear *precoding*. Precoding is an extension of beamforming. In traditional single-stream beamforming, each antenna element is linked to an RF processing chain that adjusts the gain and phase of the sent signal to maximise the received power. Precoding allows the base-station to transmit multiple data streams simultaneously.

The precoding varies depending on the MIMO operation mode. MIMO can be utilised in either a SU-MIMO (Single-User MIMO) or a MU-MIMO (Multi-User MIMO) operation mode as shown in Fig. 1.1. In SU-MIMO operation mode (also known as point-to-point MIMO) the base-station employs all of its antennas to serve one *selected* multi-antenna receiver per PRB. In this case, precoding results in multiple (the number of streams is upper bounded by the number of receive antennas) data streams emitted from the transmit antennas to serve one user. In SU-MIMO OFDMA systems, only one user is selected per PRB but it is served with potentially multiple streams.

SU-MIMO is widely adopted in today's wireless networks due to its simplicity as the optimal precoder is a linear precoder [11]. However, the number of simultaneously transmitted streams is upper-bounded by the minimum of the number of antennas at the base-station and the UE. Because UEs are often constrained in both battery and space, the number of antennas that could be installed in a UE is small, limiting the benefits brought by MIMO. Furthermore, as the UE antennas are physically close, the probability that the receive channels at the UE antennas are correlated increases, and thus the channel matrix has a low rank, reducing the number of spatial streams that can be received simultaneously. To overcome these limitations, MU-MIMO has been proposed.

In MU-MIMO operation mode, the base-station uses its antennas to send multiple data streams, each intended to a different end user. Unlike with SU-MIMO, the MIMO link is between the base-station's antennas and all the antennas of the selected users. For simplicity, we will assume single-antenna receivers throughout this thesis. Precoding in MU-MIMO is used to generate multiple-streams, each intended for a different user chosen to maximize some metric of each user's throughput to balance fairness, performance and inter-user interference.

Precoding in 4G/LTE was codebook based, i.e., the precoders were chosen from a set of fixed precoders known as the codebook. Furthermore, the maximum number of streams that could be transmitted was limited by the codebook. This simple approach eased the adoption of MIMO technology into cellular networks. However, the codebook approach is sub-optimal.

Massive MIMO refers to MU-MIMO systems where the number of antennas is significantly large, typically at least 64 antennas. The relative gap between simple linear precoding techniques, such as ZFT (Zero Forcing Transmission) and optimal precoding vanishes as the number of antennas increase [12, 13]. Furthermore, due to the fact that ZFT forces inter-user interference to be zero, the design of RRM algorithms is simplified. This is due to the fact that ZFT cancels inter-user interference and transforms the channel into an equivalent system of parallel SISO channels thus decoupling the precoding and the power distribution procedures. As a result, we will concentrate on ZFT (Zero Forcing Transmission) precoding in the massive MIMO regime in this thesis since it strikes a favourable compromise between complexity and performance making it an attractive option for practical deployment.

The RRM problem for SU-MIMO DL OFDMA systems is a straightforward extension to the SISO DL OFDMA RRM problem. The key difference is that the rate that one user may get depends on the precoding used as well as the number of transmitted streams. Furthermore, assuming the UE has N antennas and the base-station has M antennas, the wireless channel between the base-station and the UE is defined by a complex channel matrix $H \in \mathbb{C}^{N \times M}$ rather than a scalar in the SISO case. Singular Value Decomposition (SVD) of the channel matrix determines the best precoding for SU-MIMO as well as per stream power assignment as discussed in [11]. MU-MIMO is much more complex and MU-MIMO RRM is the focus of Chapter 3.

Due to the lack of cooperation amongst end users when decoding in the downlink, designing practical schemes for MU-MIMO systems is significantly more challenging than developing practical schemes for SU-MIMO systems. In SU-MIMO, all streams belong to the same UE and thus a single decoder can jointly decode all streams. This is not possible

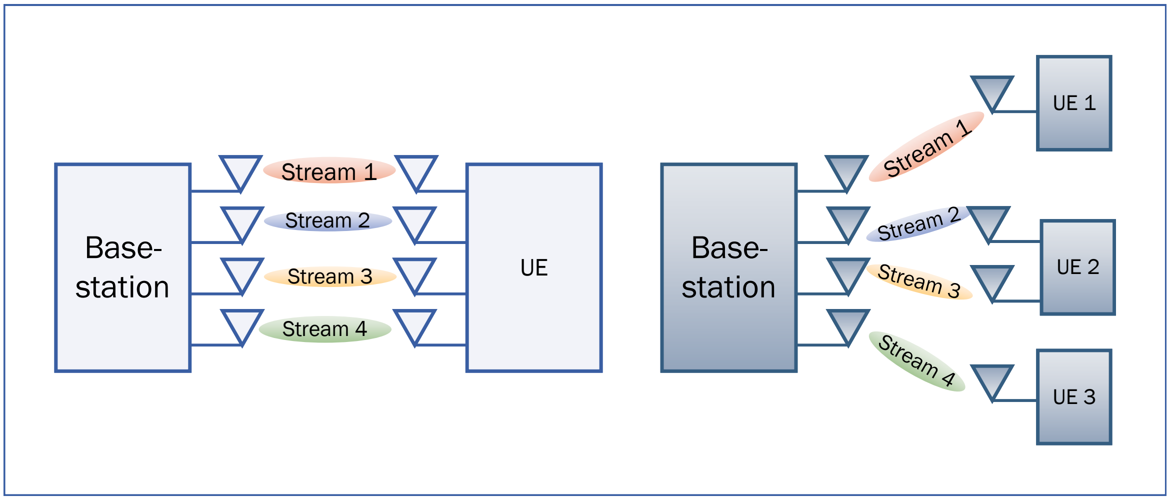

Figure 1.1: An illustration of SU-MIMO (left) and MU-MIMO (right).

in MU-MIMO since a UE cannot make use of the received copies at other UEs.

The RRM problem for MU-MIMO DL OFDMA systems is the problem of jointly optimizing; 1) Power allocation, 2) User-selection, 3) Precoding, 4) Power distribution, and 5) MCS Selection. The definitions of the power allocation and MCS selection steps are similar to the definitions established for single-user transmission techniques.

In general, the maximum number of streams that can be sent by a base-station is limited by the number of antennas installed at the base-station. With massive MIMO, the base-station has access to a large number of antennas which means that we could potentially select all users. However, as we will show in Chapter 3, although this strategy is very simple, it performs poorly.

In MU-MIMO, once a user-set is selected, the base-station needs to compute a *precoder* for it. The selected user-set as well as the precoding strategy and the power distribution impact the received signal strength not only for the user but also for other selected users. In general precoding and power distribution are coupled problems, however, in special cases such as with ZFT precoding, the problems can be decoupled.

The complexity of power distribution is a function of the used precoding. Depending on the precoding technique, the SINR experienced by the UE varies. For example, with ZFT, the precoder is designed to ensure that the intra-cell interference experienced by all selected users is nullified and power distribution can be decoupled.

1.5 Research Questions

In this thesis, we study the problem of operating, in a proportional fair manner, the downlink of an OFDMA network performing multi-user transmission. We study two types of multi-user transmission techniques, i.e., NOMA as well as MU-MIMO.

For both techniques we ask two main research questions: 1) what is the best we can do in terms of performance? and, 2) what can we do in real-time? The first question aims at assessing the best achievable performance under no runtime constraints to understand the achievable performance that we can aspire to in real network deployments. We address this question by writing the optimal RRM problem and then casting it into a problem with a favourable structure that can be utilized by specialized algorithms to efficiently find quasi-optimal solutions.

For the second question, we develop practical real-time solutions, i.e., algorithms that can produce a decision within a specific maximum amount of runtime, that can be used for network operation. To derive these solutions, we decouple the problem into several sequential processes corresponding to the different procedures defined in Section 1.2. Following this, we derive simple methods for performing each sequential step. Finally, we compare the performance achieved with the proposed online methods to those obtained during the initial offline assessment study via optimization algorithms.

1.6 Thesis outline

This thesis is structured as follows. In Chapter 2, we study in detail the RRM problem for multi-cell DL OFDMA systems performing multi-transmission using NOMA. We start by reviewing the related literature to position the main contributions with respect to the previous work on the subject. Then, we perform an offline study, i.e., we formulate the optimization problem to compute the optimal performance and present a method for providing a quasi-optimal offline solution. Moreover, we present our proposed solutions for real-time network operations that perform close to the quasi-optimal solutions derived in the offline solution. This chapter is based on the results published in [7].

In Chapter 3, we switch our focus to studying the problem of operating the downlink of a MU-MIMO single- cell system. In this chapter, we assume a single-cell DL OFDMA MU-MIMO system serving a set of single antenna UEs. Following a review of the related work, we present the system model as well as the RRM problem definition. Then we present the proposed methods for searching for good feasible solutions in an offline scenario. Finally, we present a real-time solution that performs similar to the best offline solutions while having reasonable computational complexity.

Finally, in Chapter 4, we summarize the main results and insights on NOMA and MU-MIMO and we describe possible directions for future work.

Note that the contributions are listed in Chapter 2 for NOMA and in Chapter 3 for MU-MIMO.

Chapter 2

Hybrid NOMA in Multi-Cell Networks: From a Centralized Analysis to Practical Schemes

In this chapter, we investigate the performance of a hybrid[1] NOMA (Non-Orthogonal Multiple Access) multi-cell downlink system. The results in this Chapter have been published in [7].

2.1 Motivation

NOMA (Non-Orthogonal Multiple Access) is one of the most recent advances to be considered for 5G cellular systems[9]. In contrast to traditional OMA (Orthogonal Multiple Access) systems, NOMA allows multiple users to be simultaneously allocated to the same PRB (Physical Resource Block). This is achieved by superimposing the transmitted messages, typically in the power or code domain. At the receiver side, SIC (Successive Interference Cancellation) is used to eliminate unwanted weaker signals in a certain order (referred to as decoding order in the following).

With the rapid increase in the number of active devices [14] and growth in popularity of IoT (Internet of Things) applications, the efficient use of radio spectrum is becoming increasingly critical. NOMA improves spectral efficiency by sharing the same resource among

[1] we call it hybrid because user equipment can have different SIC (Successive Interference Cancellation) capabilities.

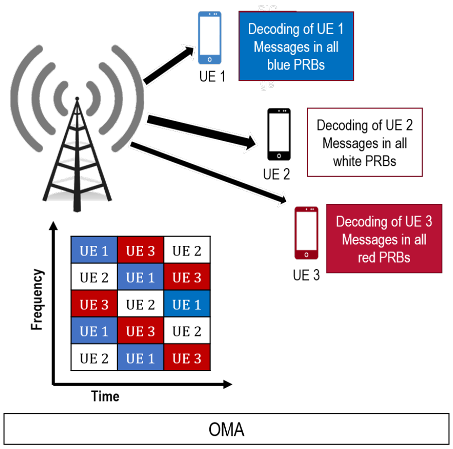
UE 1
Decoding of UE 1 Messages in all blue PRBs
UE 2
Decoding of UE 2 Messages in all white PRBs
UE 3
Decoding of UE 3 Messages in all red PRBs
Frequency
Time
UE 1 UE 3 UE 2
UE 2 UE 1 UE 3
UE 3 UE 2 UE 1
UE 1 UE 3 UE 2
UE 2 UE 1 UE 3
OMA

Figure 2.1: DL signalling process illustration for OMA.

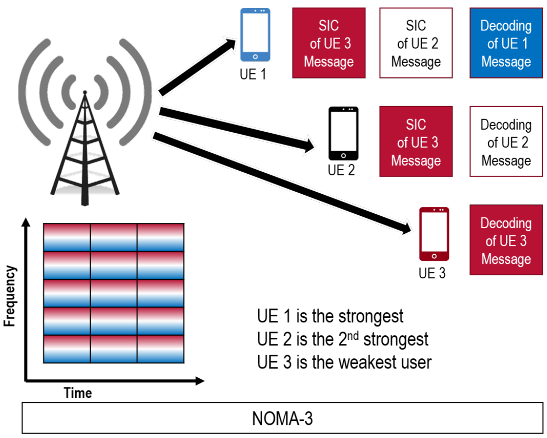

Figure 2.2: DL signalling process illustration for NOMA-3.

several users at the expense of some additional interference. ICIC (Inter-Cell Interference Coordination) schemes can therefore be even more important than for OMA.

As opposed to OMA where a PRB is used exclusively by a single user, NOMA-N is a technology that allows a base-station to simultaneously transmit to N users using the same PRB (Physical Resource Block) and the same antenna element. The base-station selects an **ordered** set of users $(i_1, ..., i_N)$ and transmit to them so that, at the receiver side, selected user i_n uses Successive Interference Cancellation (SIC) techniques to remove the interference induced by users $i_{n+1}, ..., i_N$ and then decodes its message under the interference created by users $i_1, ..., i_{n-1}$. For successful decoding, the receiver has to succeed in the SIC procedure as well as in decoding its own message. The main differences between NOMA and OMA are illustrated in Fig. 2.1 and Fig. 2.2, respectively.

This chapter investigates the scheduling of a hybrid NOMA-N multi-cell DL (Downlink) system where N denotes the maximum number of users that can be multiplexed in a given PRB (we call it hybrid to emphasize that UEs (User Equipment) have heterogeneous SIC capabilities). There is no clear consensus in the literature as to what falls under the term scheduling for a NOMA-N system. In this work, scheduling is defined as the joint process of i) allocating power to each PRB in each cell (i.e., the power allocation problem), ii) selecting for each PRB an ordered set of *at most* N users in each cell, iii) distributing the allocated power among the selected users in each PRB of each cell (i.e., the power distribution problem), and iv) selecting the MCS (Modulation and Coding Scheme) used for each transmission in each PRB in each cell. By incorporating the MCS selection to the joint scheduling definition, as opposed to using Shannon's $log_2(1 + SINR)$ capacity formula, the problem becomes even more challenging but much more realistic as will be discussed later. The contributions of this chapter are:

1. The formulation of the centralized (DL) scheduling problem for a proportional fair multiple access scheme that combines hybrid NOMA and OFDMA. The scheduler selects an ordered set of up to N users to share each PRB using power-domain NOMA. This formulation of the centralized scheduling problem accounts for *frequency-selective fading variations*, *different UE SIC capabilities* and *inter-cell interference* while jointly optimizing power allocation and user and MCS selection. Although this centralized problem cannot be used in practice as it requires a large amount of information to be exchanged in real time and is too large to be solved quickly, it can serve as an upper-bound for studying the performance of practical schemes in offline studies. Numerical results show that inter-cell interference coordination through centralized power control can improve system performance by up to 100% when compared to the case where power is allocated equally to the channels. They also show that the performance gain might be less than computed in other studies due to the heterogeneity of the UEs' SIC capabilities.

2. Regarding power allocation, a static inter-cell interference coordination scheme for hybrid NOMA-N, based on a power map, that achieves performance close to that of the theoretical centralized problem is proposed. The proposed power map is calibrated offline in a robust manner during the planning phase of the network. Rather strikingly, it offers performance that is only 15% away from the centralized upper-bound regardless of the value of N, and the variations in the operation conditions such as the user mobility and channel coherence characteristics.

3. Engineering insights into the effect of UE velocity on NOMA performance and of using Shannon's formula instead of a practical MCS scheme are presented.

4. Finally, given a static power allocation, a family of practical scheduling algorithms for hybrid NOMA-N is proposed. The algorithms are local because scheduling can be decoupled into local problems for a given static power allocation. The proposed algorithms use OMA SINR reports from UEs and therefore no additional measurements are needed with respect to OMA. Each of them exhibits a different trade-off between complexity (i.e., run-time) and performance.

This chapter is structured as follows; Section 2.2 presents the necessary background on NOMA. Section 2.3 presents a summary of the related work. The system model and the

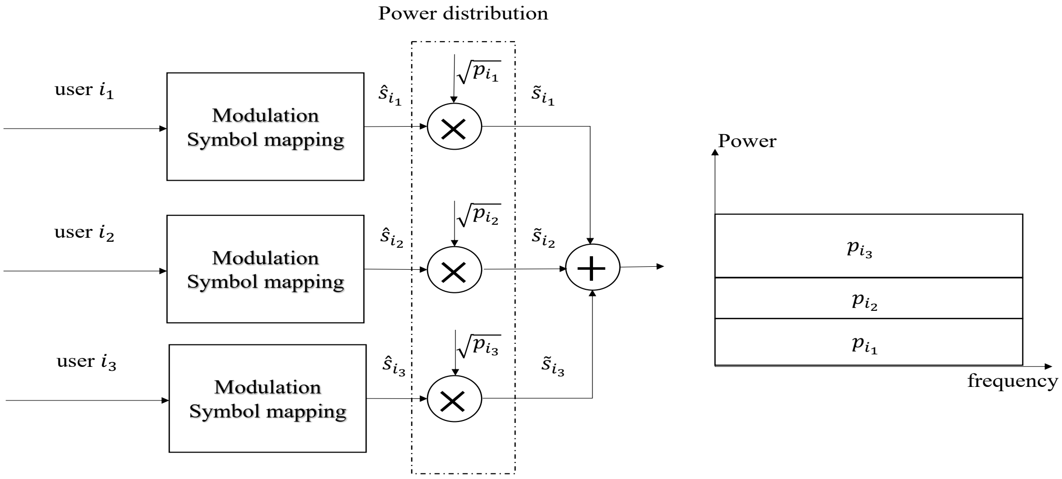

Figure 2.3: Superposition coding process setup for downlink NOMA.

main assumptions are presented in Section 2.4. In Section 2.5, we formulate and solve the centralized scheduling problem and present numerical results that show the maximum gain achievable by NOMA over OMA and by NOMA with smart power allocation compared to NOMA with simple equal power allocation. Motivated by these results, a simple power allocation scheme based on a power map is presented and evaluated in Section 2.6. In Section 2.7, a family of practical user scheduling algorithms is proposed and evaluated in terms of complexity and performance. Finally, the study is concluded in Section 2.8.

2.2 Background

NOMA (Non-Orthogonal Multiple Access) is a technology that increases the capacity of a wireless communication system by allowing multiple signals to be transmitted simultaneously using the same PRB (Physical Resource Block). Simply put, NOMA allows the base-station to communicate with multiple UEs on the same PRB at the same time. It does not require multiple antennas at the base-station or at the UEs.

NOMA can be implemented in two ways: in the power domain or the code domain. In power domain NOMA, the base-station uses varying levels of power to allow the receivers to distinguish the messages sent to different users. This causes the message sent with the

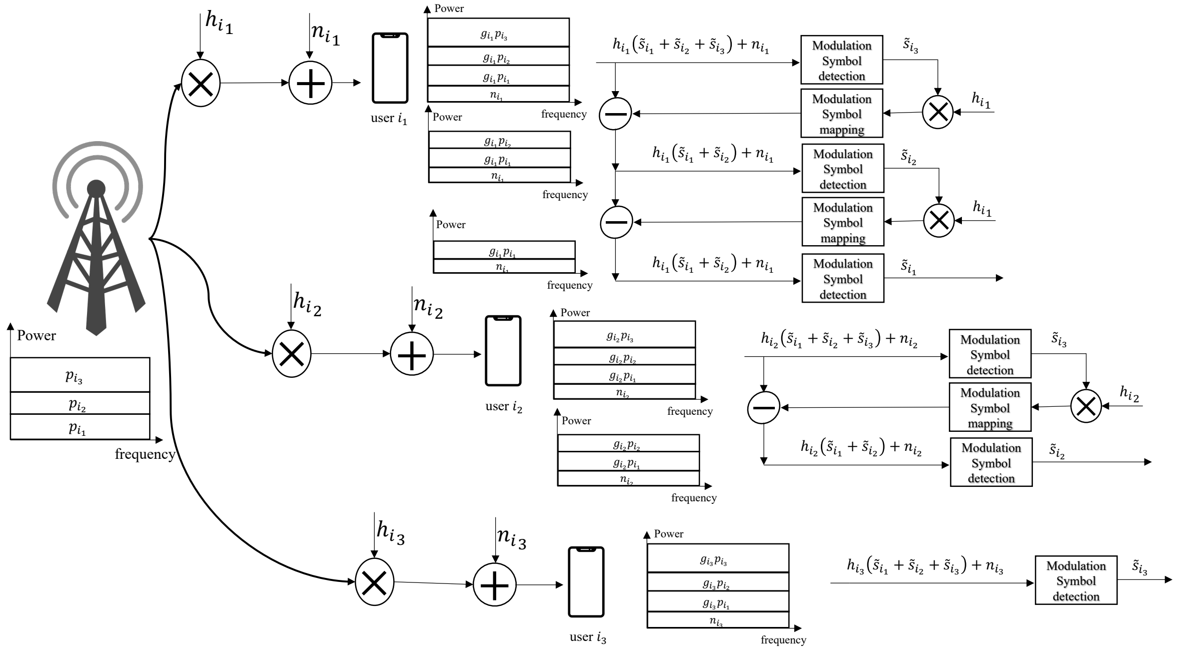

Figure 2.4: NOMA-3 setup for the downlink illustrating the SIC decoding process.

highest power to be heard the best, while the other messages may be more difficult to decode. In code domain NOMA, the base-station distinguishes the different messages by assigning different codes associated with each message. Whereas power domain NOMA is analogous to distinguishing two continuous conversations at a party based on loudness, code domain NOMA distinguishes conversations based on the language spoken.

Code-domain NOMA is more complicated than power-domain NOMA for several reasons. First, in code-domain NOMA, the interference from different users is more complex and harder to manage compared to power-domain NOMA. In power-domain NOMA, the interference from other users is simply proportional to their power levels and channel gains. However, in code-domain NOMA, the interference from other users depends on the spreading code used, as well as the channel conditions, making it more difficult to predict and manage. Moreover, the receiver design for code-domain NOMA is more complex compared to power-domain NOMA since it needs to perform additional code spreading and de-spreading operations. For these reasons the 3GPP standards mainly considers power-domain NOMA for 5G and beyond cellular systems [15].

Proceeding henceforth, the discussion will focus on power-domain NOMA. At the transmitter, a NOMA-N enabled base-station selects an ordered set $(i_1, ..., i_N)$ of N users to transmit to on a given PRB simultaneously. Following this the base-station distributes power among the selected set of users. The downlink transmission process for NOMA-3 is illustrated in Fig. 2.3.

At the receiver side, a technique known as SIC (Successive Interference Cancellation) is used to separate and decode different messages. Specifically, with the ordered set of N users, $(i_1, ..., i_N)$, the following occurs. Receiver i_n will use SIC to cancel the interference cause by i_N, then i_{N-1} and so on until it removes the interference caused by i_{n+1}. Then it decodes its message under the interference caused by users $i_{n-1}, ..., i_1$. This means that receiver i_N will decode its signal without performing any SIC operations and receiver i_1 will use SIC operations to remove every other message in the order i_N, $i_{N-1}, ..., i_2$ prior to decoding its intended message. The reception process for NOMA-3 is illustrated in Fig. 2.4.

Therefore, in the NOMA downlink, when the base-station transmits a superimposed signal to both a weak and a strong user, the base-station assigns a power level to the weak user, i.e., user with poor channel quality, such that it can decode its signal without performing SIC. The strong user can correctly decode its signal by performing a SIC operation in which the signal intended to the weak user is first subtracted, leaving only the signal intended for the strong user. This decoding order is feasible since the strong user has better channel quality than the weak user and therefore can successfully decode the message intended to the weak user and remove its impact on the total received signal.

Despite NOMA being a new multiple access technique that is based on both superposition coding at the transmitter and SIC at the receiver, these techniques are not new; their origins can be found in the existing literature on information theory [2]. The elegant idea of using superposition coding at the transmitter along with SIC at the receiver was first proposed by Cover [3] and it was shown to achieve the capacity of the broadcast channel [16]. The information theoretic broadcast channel is a communication scenario where a single transmitter (base-station) sends different messages to several receivers taking into account the limitations of the channel and the interference between the messages.

Although NOMA is a promising technique to improve the capacity of wireless communications systems as well as increase the number of supported users, it faces several challenges that need to be addressed. First it requires complex SIC receivers to decode the different messages being transmitted simultaneously. Moreover, as a result of multiplexing different messages in the power or the code domain, NOMA generates additional inter-user interference which can affect system performance and needs to be carefully managed. These challenges can be addressed using advanced signal processing techniques and efficient network operation solutions.

2.3 Related Work

Significant efforts have been made towards designing schedulers for NOMA since such designs are not trivial extensions to OMA scheduling even when the power per PRB is given. In addition to selecting users and the decoding order for a given PRB, the scheduler has to determine how to distribute the power between the scheduled users.

A plethora of (joint) resource allocation problems and algorithms for DL NOMA have been proposed with various design objectives and with different assumptions (single cell vs. multi cells, flat-fading channels, etc.). DL NOMA algorithms have been proposed for maximizing energy efficiency [17, 18, 19, 20, 21], total system-throughput [22, 23, 24] and outage minimization [25, 26]. Practical system constraints such as SIC receiver sensitivity [27, 28] and maximum power per antenna [29] have been considered. Fair resource allocation in NOMA has been considered through minimum QoS (Quality of Service) constraints [30, 31], min-max fairness [32], utility fairness, [33, 34], weighted sum-rate [35] and proportional fairness [36].

The scheduling problem is typically formulated as a non-convex optimization problem that is NP hard [37, 38]. In the special case of a single-cell network, the authors in [21, 39] used bio-inspired global optimization heuristics to jointly optimize power allocation and user selection. The joint user-selection and power distribution for NOMA is typically divided into sub-problems where each problem is solved independently (note that the power allocation problem defined as the allocation of power to different PRBs is rarely studied). For the special case of a flat-fading multi-carrier NOMA system [34, 31] have proposed a decomposition approach [40] to divide the joint problem into sub-problems that are solved independently and iteratively. This iterative process has no guarantees of solving, or even convergence, but often works well in practice.

In [41], theoretical guidelines for user-selection algorithms are studied assuming a fixed power distribution strategy and the authors focus on single carrier systems. In [42], an algorithm for finding the user-selection that maximizes the sum-rate is proposed for a given power allocation in a downlink multi-carrier single-cell system with the constraining assumption that the number of active users is N times the number of subchannels. A greedy user-selection method for weighted sum-rate maximization has been proposed in [43] along with a DC (Difference of Convex)-programming power distribution method for a multi-carrier single-cell system. In [22], K-means clustering is employed to solve the user selection sub-problem.

The power distribution sub-problem has been studied extensively for some scenarios. For a single-cell single carrier system, the authors in [33] have found a closed form solution for α-fair power distribution with the assumption that all users are scheduled for the whole time-slot. In general, for the single-cell single-carrier case, a solution can be derived by solving the KKT (Karush–Kuhn–Tucker) optimality conditions of the problem as shown in [44, 45]. In [46], the authors propose a variant of the water-filling method for power distribution. In [43], DC programming was proposed to find the proportional fair power allocation.

Most of the proposed resource allocation policies for NOMA ignore power allocation, inter-cell interference and assume single-cell systems. In [47], a case study of max-min fair NOMA is analyzed to show the limitations of such inter-cell interference agnostic schemes. In [48], an inter-cell interference coordination scheme for NOMA is presented. The main idea is to use a central controller to dynamically allocate bandwidth per cell. The scheme increases the amount of signaling in the network since it requires active coordination between different base-stations.

The majority of papers assume flat-fading. However as 5G is envisioned to support much higher speeds there is a need for NOMA schemes that can work well in frequency-selective fading scenarios. Furthermore, most prior works compute the achievable rate using Shannon's capacity formula. Not only does this lead to overestimation of the achievable system performance but it also biases user selection. Indeed, with Shannon's formula, it rarely makes sense to select less than the maximum number of users while with a practical MCS scheme, it does as will be shown later.

The novelty of our work resides in the fact that we i) study a hybrid multi-cell NOMA-N system, consider MCS selection and non-flat channels, ii) formulate and solve the most general centralized joint scheduling problem for different values of N (this centralized problem can be seen as the optimum inter-cell interference coordination scheme), iii) address the problem of power allocation through the design and the calibration of a power-map suitable for both OMA and NOMA, and iv) propose a simple practical local algorithm for user selection as an alternative to optimal user-selection, achieved by exhaustive search.

2.4 System Model and Assumptions

We consider a DL cellular system with J base-stations (the set of base-stations is denoted as $\mathcal{J} = \{1, \dots , J\}$) and a licensed band of size B Hz that is reused at each base-station. The system is OFDMA-based and the band is divided in M subchannels, each with a bandwidth $b = \frac{B}{M}$ Hz. Each base-station j has a total transmission power P_j (in Watts). Time is slotted and a frame is composed of T time-slots of duration τ sec each. The sets of all channels and times slots in a frame are denoted as $\mathcal{M} = \{1, \dots, M\}$ and $\mathcal{T} = \{1, \dots, T\}$ respectively. A PRB (Physical Resource Block) is a pair (m, t), and is the smallest scheduling unit.

We consider a realization ω that is characterized by the set of all users $\mathcal{U} = \{1, \dots, U\}$ in the system, the user association that partitions $\mathcal{U}$ into J subsets $\mathcal{U}_j$ and all the DL channel gains. Let $g_{i,j}^{m,t}$ be the channel gain between base-station j and user i on PRB (m, t). Note that we assume a general frequency-selective channel. Let $y_{i,j}$ be a binary parameter indicating if user i is associated with base-station j when $y_{i,j} = 1$ or not if $y_{i,j} = 0$. We assume that user association $y_{i,j}$ is computed beforehand and that each user i is associated with only one base-station, i.e., $\sum_j y_{i,j} = 1$ and $j(i) = \sum_{j \in \mathcal{J}} y_{i,j}$ denotes the base-station that user i is associated with.

The transmit power of base-station j on PRB (m, t) is denoted as $p_j^{m,t}$. In an OMA system, the SINR (Signal-to-Noise and Interference Ratio) experienced by user i on PRB (m, t), if the user is allocated that PRB exclusively, can be computed as:

$$\gamma_i^{m,t} = \frac{\sum\limits_{j \in \mathcal{J}} y_{i,j} p_j^{m,t} g_{i,j}^{m,t}}{\sigma^2 + \sum\limits_{j \in \mathcal{J}} (1 - y_{i,j}) p_j^{m,t} g_{i,j}^{m,t}} = \frac{S_i^{m,t}}{\sigma^2 + I_i^{m,t}}, \tag{2.1}$$

where $S_i^{m,t}$, $I_i^{m,t}$ σ^2 are the received signal, inter-cell interference and noise powers respectively.

The rate $r_i^{m,t}$, seen by user i on PRB (m, t), is a function of the SINR $\gamma_i^{m,t}$ experienced by the user, i.e., $r_i^{m,t} = f(\gamma_i^{m,t})$. In practical cellular systems, a base station has access to K MCS (Modulation and Coding Scheme) and needs to select one of them for each user it wants to transmit to, based on its radio conditions. In that case, the function $f(.)$ is a piece-wise constant increasing function [1]. From an optimization perspective, this piece-wise constant function is problematic since it is non-differentiable and quasi-concave [49]. For more on rate functions please see Section 2.9.

In DL NOMA-N, base-station j can transmit at most N messages simultaneously using the same PRB (m, t) via superposition coding. Hence, in its most general form, the DL scheduler computes, on a per PRB basis, the power allocation $p_j^{m,t}$, selects an *ordered* set of active users $(i_{(1)}, i_{(2)}, ..., i_{(N)})$ (with the appropriate SIC capabilities) to share the PRB as well as the power share of each user. Mathematically, mapping between users and selected users is represented by the binary decision variable $a_{n,i}^{m,t} \in \{0, 1\}$ which has a value of 1 when user i is selected as the n-th user, i.e. $i_{(n)}$, in PRB (m, t) and 0 otherwise. The power share of the n-th selected user $i_{(n)}$ is denoted by $x_{n,j}^{m,t} \in [0, 1]$. Note that it is possible to select less than N users in a PRB by allocating no power to some of the users in the ordered set.

Due to the nature of DL-NOMA, there are certain restrictions on the selection of the ordered user set. At the receiver side, each user in this ordered set decodes several messages in a specific order using SIC. User $i_{(N)}$ decodes its message treating every other message as a source of interference, i.e., it does not require any SIC capability. User $i_{(n)}$, with $n < N$, uses SIC to eliminate the interference induced by users $i_{(n+1)}, ..., i_{(N)}$ prior to decoding its message under the interference created by users $i_{(1)}, ..., i_{(n-1)}$, hence, it needs the capability to decode $K = N - n - 1$ other signals than its own, which we refer to as SIC-K capability User $i_{(1)}$ uses SIC-N to cancel every other message and decodes its message free from intra-cell interference. Therefore, user i, the n^{th} user in the ordered set, gets the following rate in PRB (m, t):

$$r_{n,i}^{m,t} = a_{n,i}^{m,t} f\bigg(\frac{x_{n,j(i)}^{m,t}\gamma_i^{m,t}}{\sum_{q=1}^{n-1} x_{q,j(i)}^{m,t}\gamma_i^{m,t} + 1}\bigg) \tag{2.2}$$

where recall that $j(i)$ denotes the base-station that user i is associated with. As a result, two main constraints on the ordered user set selection arise. Namely, a UE selected as the n-th active user in a NOMA-N system must: 1) have SIC hardware capability to decode $(N - n)$ messages before decoding its own message, and 2) its current channel conditions allow for successful elimination of these messages.

The SIC capability of a UE is determined by the number of interfering signals a UE can cancel before decoding its own message. A UE with SIC-K capability $(K \geq 0)$ can cancel K messages before decoding its own message. For a UE i with SIC-K capability, we define the binary capability indicator parameter $z_{i,n}$ as:

$$z_{n,i} = \begin{cases} 1 & n \geq N - K \\ 0 & otherwise, \end{cases} \tag{2.3}$$

which determines if user i can be selected or not as the n-th active user considering its SIC capability. We clarify the parameter through the following examples. Assume UE i has a SIC-1 capability and that the base-station is using NOMA-3. It would have $z_{1,i} = 0$, $z_{2,i} = 1$ and $z_{3,i} = 1$ which means that it can never be selected as the first in the order, i.e., as $i_{(1)}$. In addition, if a UE has no SIC capability $(K = 0)$ it would only have $z_{N,i} = 1$ and $z_{n,i} = 0 \;\; \forall n < N$. Note that $z_{N,i} = 1 \;\; \forall i$.

Recall that the n-th selected user needs to decode the messages intended for $i_{(n+1)}, ..., i_{(N)}$ successfully to ensure that its own message is decoded successfully. For $i_{(n)}$ to be able to decode the message intended to user $i_{(n+1)}$ the following condition must be satisfied:

$$\sum_{i \in \mathcal{U}_j} a_{n,i}^{m,t} \, f\bigg(\frac{x_{n+1,j}^{m,t} \gamma_i^{m,t}}{\sum_{q=1}^{n} x_{q,j}^{m,t} \gamma_i^{m,t} + 1}\bigg) \geq \sum_{i \in \mathcal{U}_j} a_{n+1,i}^{m,t} \, f\bigg(\frac{x_{n+1,j}^{m,t} \gamma_i^{m,t}}{\sum_{q=1}^{n} x_{q,j}^{m,t} \gamma_i^{m,t} + 1}\bigg) \quad \forall j. \tag{2.4}$$

This condition can be simplified to be: $\gamma_{i_{(n)}}^{m,t} \geq \gamma_{i_{(n+1)}}^{m,t}$ if the rate function is monotonically increasing. In short, the ordering is based on the OMA SINR reported by each user. Note that due to the frequency-selective nature of the channel a valid order in one PRB might not necessarily be valid in another PRB due to channel variations.

Next we formulate the centralized scheduling problem for the hybrid NOMA multi-cell system.

2.5 Centralized Scheduling for Hybrid NOMA

2.5.1 Problem Formulation

Recall that by scheduling, we refer to the problem of jointly selecting the per PRB power allocation $p_j^{m,t}$, the ordered set of users characterized by $(a_{n,i}^{m,t})$, the intra-PRB power distribution among the selected users $x_{n,j}^{m,t}$, and the MCS selection for each transmission in each PRB which gives the rates $r_{n,i}^{m,t}$. Scheduling is performed on a frame-basis and channel gains are not assumed flat within a frame.

Given a multi-cell system realization ω (that includes the set of users (U), the user associations given by $y_{i,j}$'s, all the channel gains $g_{i,j}^{m,t}$'s, the K_i's (i.e., the SIC capability of each user), the $z_{n,i}$'s, the P_j's, and σ^2), we formulate problem $\mathbf{P}_1(\omega)$ as follows all variables are non-negative and real except the $a_{n,i}^{m,t}$'s that are binary. The purpose of investigating centralized scheduling is to provide an upper-bound against which practical schemes can be compared.

$\mathbf{P}_1(\omega)$: Centralized NOMA-N scheduling problem

$$\max_{\substack{x_{n,j}^{m,t}, a_{n,i}^{m,t}, p_j^{m,t}, \\ \lambda_i, r_{n,i}^{m,t}, S_i^{m,t}, I_i^{m,t}}} \quad \prod_{j \in \mathcal{J}} \prod_{i \in \mathcal{U}_j} \lambda_i$$

$$a_{n,i}^{m,t} \in \{0,1\} \qquad \forall n, m, t, i \qquad (2.5a)$$

$$\sum_{i \in \mathcal{U}_j} a_{n,i}^{m,t} \leq 1 \qquad \forall n, m, t \qquad (2.5b)$$

$$\sum_{n=1}^{N} x_{n,j}^{m,t} \leq 1 \qquad \forall m, t, j \qquad (2.5c)$$

$$\lambda_i \leq \sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{M}} \sum_{n=1}^{N} r_{n,i}^{m,t} \qquad \forall i \qquad (2.5d)$$

$$r_{n,i}^{m,t} \leq a_{n,i}^{m,t} f\Big(\frac{x_{n,j(i)}^{m,t} S_i^{m,t}}{\sum_{q=1}^{n-1} x_{q,j(i)}^{m,t} S_i^{m,t} + \sigma^2 + I_i^{m,t}}\Big) \qquad \forall n, m, t, i \qquad (2.5e)$$

$$r_{n,i}^{m,t} \leq \sum_{u \in \mathcal{U}_{j(i)}} a_{\varrho,u}^{m,t} f\Big(\frac{x_{n,j(i)}^{m,t} S_u^{m,t}}{\sum_{q=1}^{n-1} x_{q,j(i)}^{m,t} S_u^{m,t} + \sigma^2 + I_u^{m,t}}\Big), \qquad \forall \varrho \in \{1, ..., n-1\}, n, m, t, i \qquad (2.5f)$$

$$S_i^{m,t} = \sum_{j \in \mathcal{J}} y_{i,j} p_j^{m,t} g_{i,j}^{m,t} \quad \& \quad I_i^{m,t} = \sum_{j \in \mathcal{J}} (1 - y_{i,j}) p_j^{m,t} g_{i,j}^{m,t} \qquad \forall i, m, t \qquad (2.5g)$$

$$\sum_{m \in \mathcal{M}} p_j^{m,t} \leq P_j \qquad \forall j, t \qquad (2.5h)$$

$$a_{n,i}^{m,t} \leq z_{n,i} \qquad \forall n, m, t, i \qquad (2.5i)$$

The objective of the centralized scheduler is to maximize the product of the throughput of all users in the multi-cell system within a given frame comprising MT PRBs to achieve per-frame proportional fairness between the users. In that case, as discussed in [50, 51, 52], the natural performance metric to compare schemes is the geometric mean of the throughput, which is the normalized product defined as: $GM(\omega) = (\prod_{j \in \mathcal{J}} \prod_{i \in \mathcal{U}_j} \lambda_i)^{1/|\mathcal{U}|}$.

Constraint (2.5b) follows from the fact that a PRB can only be allocated to a single user of each rank in the order. Constraint (2.5c) follows from the power distribution fraction definition for each user in the ordered set for a given PRB. Constraints (2.5e) - (2.5f) follow from the NOMA conditions for successful operation described in the previous section. Constraint (2.5h) states that the power per time-slot is kept below the maximum transmission power available. Finally, constraint (2.5i) ensures that a user has the SIC capability needed to perform the required interference elimination procedures prior to decoding its own message.

The problem in the current form is a MINLP (Mixed Integer Non-Linear Program) which is very hard to solve for a reasonably sized system. To solve the problem more efficiently, the following change of variable is performed:

$$\varphi_{n,i}^{m,t} = a_{n,i}^{m,t} x_{n,j(i)}^{m,t} p_{j(i)}^{m,t} \tag{2.6}$$

This is done by noting that we can move the binary allocation variable inside the rate function and thus the product of the three variables appears in the argument of the rate function. Constraint (2.5i) is dropped and the binary capability $z_{n,i}$ parameter is multiplied by the rate function to yield the same behavior as constraint (2.5i) by blocking allocation to users that lack the require SIC decoding capability.

By using the variables $\varphi_{n,i}^{m,t}$ along with $p_j^{m,t}$, the problem can be re-written as $\mathbf{P2(\omega)}$ (all variables are non-negative and real and $\epsilon > 0$ is a small constant).

$\mathbf{P}_2(\omega)$: Centralized NOMA-N scheduling problem (signomial form)

$$\max_{\varphi^{m,t}_{n,i,j},p^{m,t}_{j},\lambda_i,r^{m,t}_{n,i}} \quad \prod_{j\in\mathcal{J}}\prod_{i\in\mathcal{U}_j}\lambda_i^{1/|\mathcal{U}|}$$

$$\sum_{i\in\mathcal{U}_j}\sum_{n=1}^{N}\varphi^{m,t}_{n,i}\leq p^{m,t}_{j} \qquad \forall j,m,t \tag{2.7a}$$

$$\varphi^{m,t}_{n,i}\varphi^{m,t}_{n,v}\leq\epsilon \qquad \forall\{i,v\in\mathcal{U}_j:i\neq v\},j,m,t,n \tag{2.7b}$$

$$r^{m,t}_{n,i}\;r^{m,t}_{n,v}\;\leq\epsilon \qquad \forall\{i,v\in\mathcal{U}_j:i\neq v\},j,m,t,n \tag{2.7c}$$

$$\sum_{m\in\mathcal{M}}p^{m,t}_{j}\leq P_j \qquad \forall j \tag{2.7d}$$

$$\lambda_i\leq\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}}\sum_{n=1}^{N}r^{m,t}_{n,i} \qquad \forall i \tag{2.7e}$$

$$r^{m,t}_{n,i}\leq z_{n,i}f\Bigg(\frac{g^{m,t}_{i,j(i)}\varphi^{m,t}_{n,i}}{g^{m,t}_{i,j(i)}\sum\limits_{\nu\in\mathcal{U}_{j(i)}}\sum\limits_{q=1}^{n-1}\varphi^{m,t}_{q,\nu}+\sigma^2+\sum\limits_{\substack{k\in\mathcal{J}\\k\neq j(i)}}p^{m,t}_{k}g^{m,t}_{i,k}}\Bigg) \qquad \forall i,m,t,n \tag{2.7f}$$

$$r^{m,t}_{n,i}\leq z_{n,i}f\Bigg(\frac{\varphi^{m,t}_{n,i}\sum\limits_{u\in\mathcal{U}_{j(i)}}\varphi^{m,t}_{\varrho,u}g^{m,t}_{u,j(i)}}{\sum\limits_{u\in\mathcal{U}_{j(i)}}\Big(\varphi^{m,t}_{\varrho,u}\big(g^{m,t}_{u,j(i)}\sum\limits_{\nu\in\mathcal{U}_{j(i)}}\sum\limits_{q=1}^{n-1}\varphi^{m,t}_{q,\nu}+\sigma^2+\sum\limits_{\substack{k\in\mathcal{J}\\k\neq j(i)}}p^{m,t}_{k}g^{m,t}_{u,k}\big)\Big)}\Bigg) \qquad \forall\varrho\in\{1,...,n-1\},i,m,t,n \tag{2.7g}$$

The total transmitted DL power per PRB is the sum of the power allocated to the N multiplexed users (constraint (2.7a)). Constraints (2.7b) and (2.7c) enforce that only one user can be allocated a non-negligible power or rate at every rank in the ordered set allocated to each PRB.

Upperbound to $\mathbf{P}_1(\omega)$: $\mathbf{P}_2(\omega)$ is equivalent to problem $\mathbf{P}_1(\omega)$ for $\epsilon=0$ and it becomes an upper-bound if $\epsilon>0$ since it has a larger feasible search space. To solve the problem, we replace the piece-wise MCS function $f(.)$ by a monomial power function $\hat{f}(.)$ that upper bounds it and then $\mathrm{P}_2(\omega)$ becomes a signomial program. The single-condensation method is then used to solve $\mathbf{P}_2(\omega)$ [53], [54].

Feasible solution to $\mathbf{P}_1(\omega)$: Let the solution of $\mathbf{P}_2(\omega)$ be denoted as $(\varphi_{n,i}^{m,t})^*$ and $(p_j^{m,t})^*$. The value of the objective function is an upper bound for problem $\mathbf{P}_1(\omega)$. A feasible solution (and hence a lower bound) to $\mathbf{P}_1(\omega)$ can be obtained from the solution of $\mathbf{P}_2(\omega)$ by using the following procedure. First, a feasible user selection is computed as:

$$(a_{n,i}^{m,t})^* = \begin{cases} 1 & \text{if } \; i = \underset{i \in \mathcal{U}_j}{\operatorname{argmax}} \; (\varphi_{n,i}^{m,t})^* \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

Then, a feasible power-split is computed using

$$(x_{n,j(i)}^{m,t})^* = \frac{\sum\limits_{i \in \mathcal{U}_{j(i)}} (\varphi_{n,i}^{m,t})^*}{(p_{j(i)}^{m,t})^*}. \tag{2.9}$$

Finally, the rate $r_{n,i}^{m,t}$ is calculated by using the piece-wise constant MCS function $f(.)$.

If the upper bound and the lower bound are close, then the feasible solution is quasi-optimal. Note that a better feasible solution could be obtained from problem $\mathbf{P}_2(\omega)$ by replacing $\hat{f}(.)$ by a tight (not necessarily upper bounding) monomial approximation $f_A(.)$. More details on the rate function approximation are given in section 2.9.

2.5.2 Numerical Results

In this subsection, the upper bound and the feasible solution (lower bound) for $\mathbf{P}_1(\omega)$ for NOMA-2 and NOMA-3 are compared for different mixes of SIC capabilities. To quantify the impact of careful power allocation, the feasible solution is also compared with the one obtained with equal power allocation per PRB. Specifically, a regular (hexagonal) network with $J = 7$ identical base-stations ($P_j = 40$ W for all j) is considered with wrap-around to prevent border effects.

We consider an urban setting and the large scale fading component of $g_{i,j}^{m,t}$ is calculated using the corresponding 3GPP model [55]. The channel power gain between user i and base-station j in PRB (m, t) is calculated as:

$$g_{i,j}^{m,t} = F_{i,j}^{m,t} 10^{-PL_{i,j}/10},$$

Table 2.1: System parameters used in the simulation.

Network Model Parameters			Channel Model Parameters		
Parameter	**Symbol**	**Value**	**Parameter**	**Symbol**	**Value**
Minimum UE distance	$d_{(min)}$	35 ms	Subchannel bandwidth	b	180 KHz
Inter-site distance	D	500 m	Carrier frequency	f_c	2.5 GHz
Number of users	U	10 J	Shadowing coefficient	σ_{shadow}	6 dB
Number of base-stations	J	7	Noise spectral power density	N_o	-174 dBm/Hz
Number of subchannels	M	100	Coherence bandwidth (pedestrian)	B^p_{coh}	100 MHz
Number of time-slots	T	10	Coherence time (pedestrian)	T^p_{coh}	10 ms
Time slot duration	τ	1 ms	Coherence bandwidth (vehicular)	B^v_{coh}	2 MHz
Base-station power	P_j	40 W	Coherence time (vehicular)	T^v_{coh}	10 ms

where $F_{i,j}^{m,t}$ is the small scale fading component and $PL_{i,j}$ is the path loss which is assumed to be the same for all PRBs in the frame.

Specifically, the path-loss $PL_{i,j}$, in dB, between user i and base-station j is computed as:

$$PL_{i,j} = 13.54 + 39.08 \log_{10}\left(\sqrt{d_{i,j}^2(m) + (25 - 1.5)^2}\right) + 20 \log_{10}(f_c) + 20 + \mathcal{S}_{i,j} \tag{2.10}$$

where $d_{i,j}$ is the distance between base-station j and user i in meters, f_c is the carrier frequency in GHz and $\mathcal{S}_{i,j}$ is a log-normal random variable with a zero mean and a standard deviation $\sigma_{SF} = 6\ dB$ that accounts for the impact of large scale shadowing.

For small scale fading, independent and identically distributed (i.i.d) block Rayleigh fading is considered in which the small-scale fading parameters are constant within each coherence block, and are independent from block to block. A *coherence block* is defined as the set of adjacent PRBs in time and frequency for which the channel fading parameters remain the same. This time (resp. bandwidth) is called the channel coherence time (resp. channel coherence bandwidth). The fading component within a coherence block is an exponential random variable with a unit mean.

Two fading scenarios are considered, one for pedestrian users and one for vehicular users. For the pedestrian setting, a coherence bandwidth and time of 10 MHz and 100 ms respectively are considered. For the vehicular setting, a coherence bandwidth and time of 2 MHz and 10 ms are assumed. A user sees the same channel gains on all the PRBs of a coherence block. The feasible rate computations are based on the MCS mapping table presented in [1]. We assume that the user association is determined based on the best mean SINR. The values of the parameters are listed in Table 2.1. Note that $\sigma^2 = N_0 b$.

Five scenarios are compared, each with specific system and UE capabilities. Namely, the scenarios considered are:

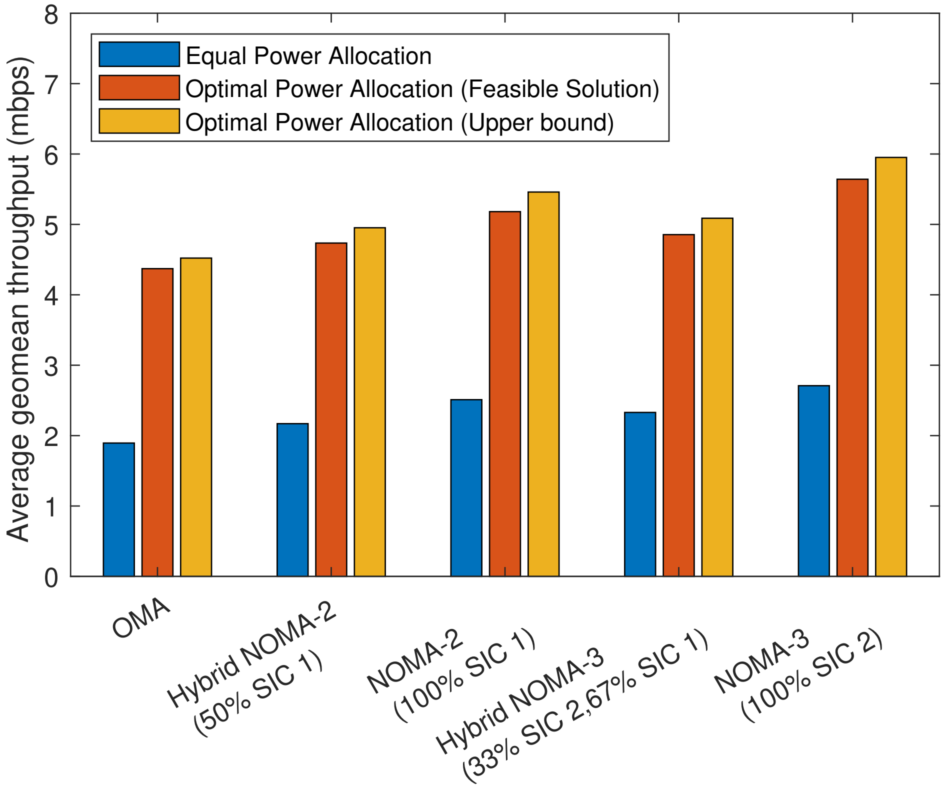

Figure 2.5: Centralized scheduling performance for the pedestrian setting, results are averaged over 100 realizations with 70 users uniformly distributed over the 7 cells.

1. OMA where the base-stations transmit only one message to one user in a PRB.
2. Hybrid NOMA-2 where the base-station can schedule two users in a PRB but only 50% of UEs are SIC-1 capable.
3. Hybrid NOMA-3 where 33% are SIC-2 UEs and 67% are SIC-1 UEs.
4. NOMA-2 $N = 2$ where all UEs are capable of doing SIC-1.
5. NOMA-3 where all UEs are capable of doing SIC-2.

Figs. 2.5 and 2.6 present the numerical results (feasible solution and upper bound) for the centralized NOMA scheduling for the pedestrian and the vehicular settings, respectively.

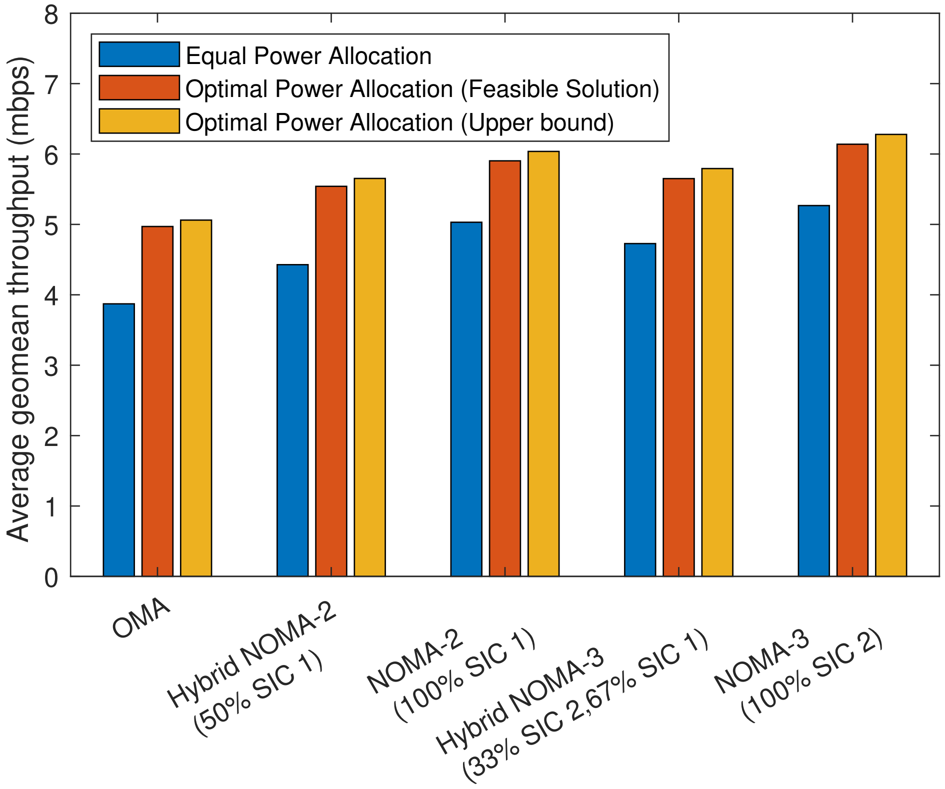

Figure 2.6: Centralized scheduling performance for the vehicular setting, results are averaged over 100 realizations with 70 users uniformly distributed over the 7 cells.

For each scenario we also show the state-of-the-art where the power is divided equally per subchannel, i.e., $p_j^{m,t} = \frac{P_j}{M} \qquad \forall\ m \in \mathcal{M},\ t \in \mathcal{T}$. This scheme is referred to as the equal power allocation scheme in the following.

The results given are averaged over 100 realizations, (which corresponds to 1000 timeslots and $100,000$ PRBs) when the total number of users in the system is 70 uniformly distributed over the 7 cells. The geometric mean throughput is used as performance indicator since it is the objective function for proportional fair scheduling.

The gap between the feasible solution and the upper bound shows that the upper bound is tight. For the pedestrian setting, the gap is 3.4%, 5.3%, 5.4% for the centralized OMA, NOMA-2 and NOMA-3 schemes respectively. Similarly, for the vehicular setting, the gap is 1.8%, 2.2%, 2.3% for the centralized OMA, NOMA-2 and NOMA-3 schemes respectively.

Likewise, the gap in the cases of hybrid NOMA-2 and NOMA-3 the gap is 4.6% and 4.8% respectively at low UE speed and the gap is 2.0% and 2.1% respectively at vehicular speed. Hence, the centralized scheduling problem has been solved to quasi-optimality.

The first observation is that, if we take OMA with equal power allocation as the baseline, a larger gain can be obtained by improving power allocation (and remaining with OMA) than using NOMA-2 (or 3) with equal power allocation. In fact, with optimal power allocation, the gain due to NOMA decreases (the gain of NOMA-2 over OMA is about 31% and 22% with equal and optimal power allocation respectively). The second observation is that NOMA-3 does not bring a significant improvement over NOMA-2. The fact that power allocation can significantly improve performance motivates the derivation of simple and efficient power allocation schemes in the next section. We also note that the mobility model has a significant impact on the gains achievable via NOMA. Indeed, the centralized scheduling, with optimal power allocation, for NOMA-2 (resp. NOMA-3) brings a gain (compared to equal power allocation) of 108% (resp. 106%) for the pedestrian case and only 17% (resp. 17%) for the vehicular case. This is mainly due to the longer fading duration for low mobility users that can be counteracted by well-thought power allocation and interference coordination. Finally, the fact that not all users have the right SIC capability impacts NOMA's achievable performance.

The results so far are based on centralized scheduling and do not show large gains for NOMA over OMA (for a given power allocation scheme). However, these results are based on independent realizations as oppose to a dynamic scenario where users come and go. To validate these conclusions in a more dynamic setting, we will first propose a realistic (static) power allocation scheme. Indeed, a dynamic coordinated power allocation is too difficult to manage in practice. We will show next that the practical static power allocation scheme ensures operation simplicity, is easy to calibrate, and produces performance that is relatively close to the performance of the upper bound centralized scheduling. However, we have one additional engineering insight to discuss before doing so regarding the use of a practical MCS function instead of Shannon's formula.

In Fig. 2.7 (resp. Fig. 2.8), we compare how often the scheduler would select 1 user instead of 2 (resp. 1 user or 2 users instead of 3) when Shannon's $\log_2(1+SINR)$ formula is used as opposed to when the practical MCS is used for NOMA-2 (resp. NOMA-3). The results show that with Shannon's formula, the scheduler will always try to select the maximum number of users. However, with the practical MCS scheme, the scheduler will select a single user for 20% of the PRBs for NOMA-2 and will select 1 or 2 users for 50% of the PRBs for NOMA-3. Hence, using Shannon's formula to perform user selection can be misleading.

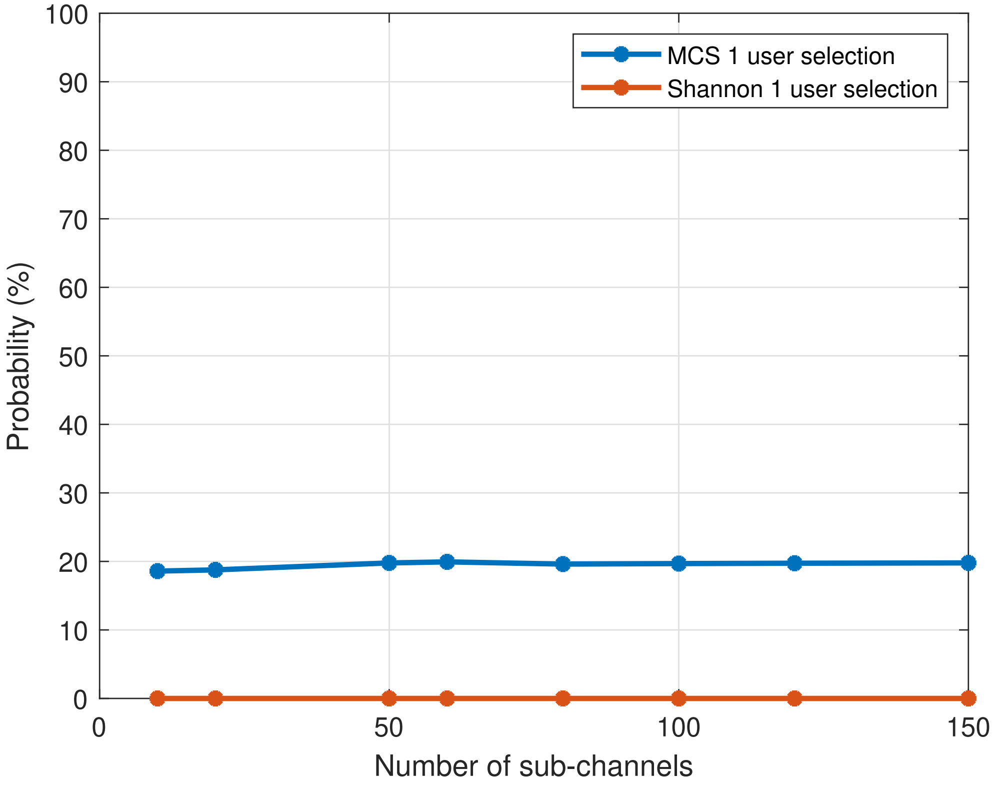

Figure 2.7: Probability of scheduling one user with Shannon formula (Red) and practical MCS rate function (blue), assuming NOMA-2 and all users are at least SIC-1 capable.

2.6 Static Power Allocation via Power Map for NOMA

It is important to note that once the power per PRB is fixed, as in the case of the equal power allocation scheme, each base-station can determine its scheduling locally (i.e., the selection of users in each PRB, the power distribution within a PRB among the selected users, and the MCS selection for each transmission in each PRB), making the scheduling problem much simpler. Indeed, in the centralized problem above, the scheduling in each cell is only coupled to scheduling in other cells because of power allocation. In this section, a simple scheme is presented that sets the power allocation *statically* (but not equally) for NOMA by pre-computing beforehand (i.e., in the planning stage) a power map, using many realizations. This power map is then published and used by each base station during scheduling. Combined with a practical local user scheduling algorithm, to be presented

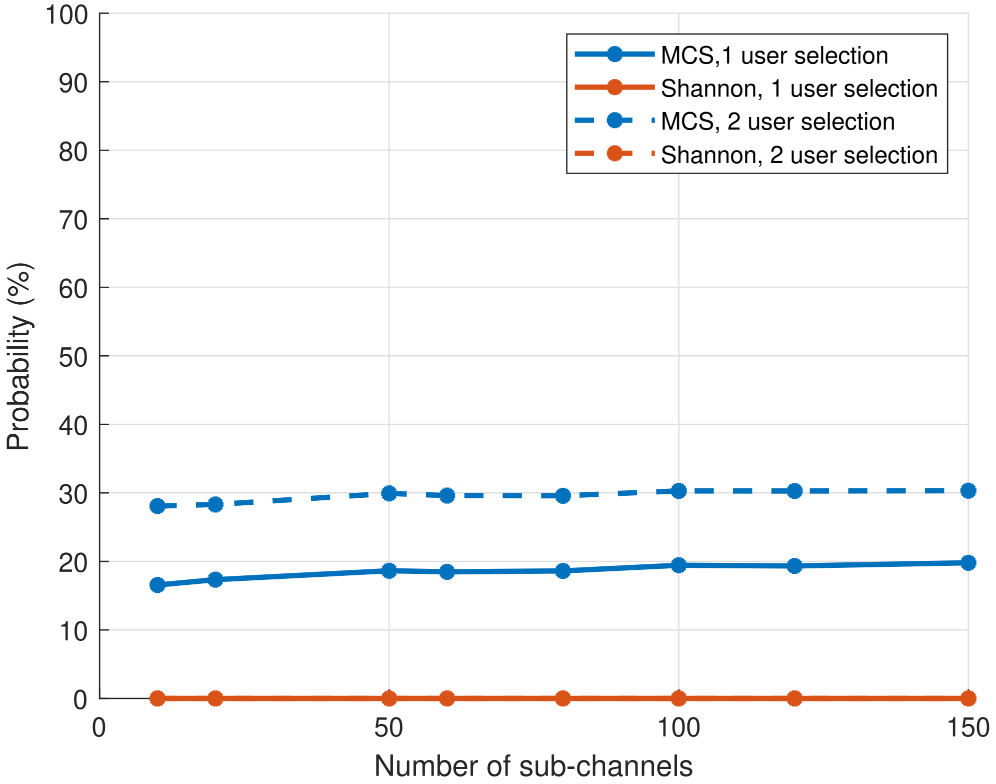

Figure 2.8: Probability of selecting one user (solid) and two users (dashed) with Shannon's formula (Red) and practical MCS rate function (blue) assuming NOMA-3 and all users are SIC-3 capable.

next, the proposed power map improves the performance of NOMA compared to the state of the art that uses equal power allocation and approaches the performance of the centralized upper-bound presented earlier.

Specifically, while $p_j^{m,t}$, the power used by base-station j in PRB (m,t), is a variable in $\mathbf{P}_1(\omega)$, it will be set offline (and remains the same for all frames) using a process described next and become an input to the scheduling problem. The set of $p_j^{m,t}$ for all j, m, t is called a power map. Note that equal power allocation can be seen as a special power-map. It was shown in [56], that carefully calibrated power maps for OMA are robust to changes in operating conditions such as the number of users per cell and the channel conditions and yield excellent performance. To the best of our knowledge, this work shows the first attempt at generalizing the concept of power map to NOMA.

To compute our power map, the M subchannels are divided into L *interleaved* sub-bands, each with an equal number of subchannels, denoted as M_l. In cases where the number of subchannels M is not divisible by L, a minor adjustment is made to keep the number of channels in all interleaved sub-bands roughly the same. For each base-station, each PRB in a sub-band is allocated the same power level.

Base-stations in the system are divided into L groups (similar to frequency reuse coloring). Let $s(j)$ be the group number for base-station j. All base-stations in the same group use the exact same transmit power in PRB (m, t). We assume that the power used on a PRB can only take one of L values. Let $\mathcal{L} = \{\beta_1, \ldots, \beta_L\}$ be the set of per channel power levels where $0 \leq \beta_l \leq P_j$ for $l = 1, \ldots, L$ (our purpose is to compute these values offline). The power levels are allocated cyclically to each PRB with a different initial level per group, i.e., each group uses a different level in PRB $(1, 1)$. For example, referring to Fig. 2.9 where $L = T = 3$ and $M = 5$, the base-stations in group 1 would use power level β_1 on PRB $(1, 1)$, β_2 on PRB $(2, 1)$, etc. Base-stations in group 2 starts the cycle with power β_2 on PRB $(1, 1)$, etc. Thus, the transmit power $p_j^{m,t}$ used by base-station j on PRB (m, t) is statically set at β_l where

$$l = 1 + \big((s(j) - 1) + (m - 1) + (t - 1) \mod L\big). \tag{2.11}$$

The interleaving helps reduce long-term fades by providing more diverse opportunities to be utilized by a smart scheduling algorithm. With this scheme each cell sees some PRBs with improved SINR (with respect to the case with equal power allocation) by reducing the inter-cell interference (of course in return, each cell also sees some PRBs with worse SINR). As an example, assume that $(\beta_1, \beta_2, \beta_3) = (0.04, 0.06, 0.9)$ which makes all the PRBs assigned β_3 see very good SINRs (since the BSs in other groups transmit with low power). These PRBs could be used for edge users while the users closer to the center can make use of the PRBs where the inter-cell interference is larger. Recall that the PRBs are dynamically allocated to users by a local scheduler.

The objective is to obtain a power map that is robust, i.e., performs well for a large range of realizations. To do so, a methodology similar to the one proposed in [8] is followed. A set Ω of test realizations is considered to calibrate, for a given L, the power map (i.e., obtain the set $\mathcal{L}$). Given a pre-selected partition of the base-stations into L groups, the objective is to select $\mathcal{L}$ that maximizes the ensemble average of the geometric mean $GM(\omega)$ over the test realizations.

Specifically, given a set of realizations Ω where each realization ω includes the set of user $\mathcal{U}(\omega)$, the $y_{i,j}(\omega)$'s, all the channel gains $g_{i,j}^{m,t}(\omega)$'s, the $K_i(\omega)$'s (i.e, the SIC capability

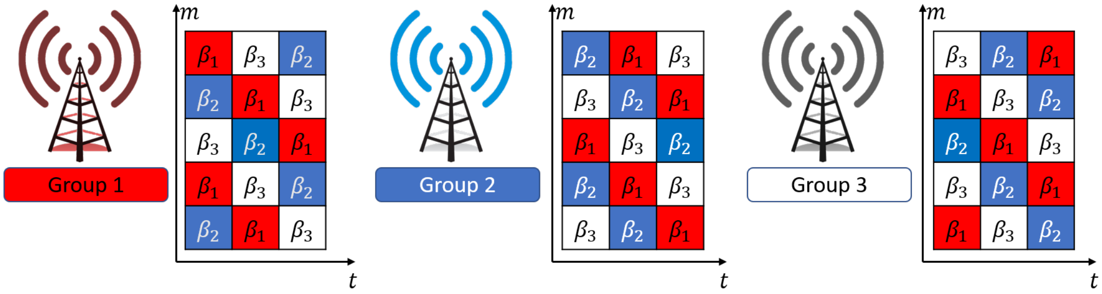
Group 1
m
β_1 β_3 β_2
β_2 β_1 β_3
β_3 β_2 β_1
β_1 β_3 β_2
β_2 β_1 β_3
t
Group 2
m
β_2 β_1 β_3
β_3 β_2 β_1
β_1 β_3 β_2
β_2 β_1 β_3
β_3 β_2 β_1
t
Group 3
m
β_3 β_2 β_1
β_1 β_3 β_2
β_2 β_1 β_3
β_3 β_2 β_1
β_1 β_3 β_2
t

Figure 2.9: Example power map with $L = 3$, $M = 5$ and $T = 3$.

of each user), the $z_{n,i}(\omega)$'s), the $P_j(\omega)$'s, the $s(j)$'s (i.e, the group number per base-station), L and σ^2, we formulate problem $\mathbf{P}_3(\omega)$ as below. All variables are non-negative and real except the $a_{n,i}^{m,t}$'s that are binary.

$\mathbf{P}_3(\Omega)$: Power map calibration with L levels for NOMA-N.

$$\max_{x^{m,t}_{n,j}(\omega)a^{m,t}_{n,i}(\omega),\beta_l} \quad \frac{1}{\sum\limits_{\omega\in\Omega}|\mathcal{U}(\omega)|}\sum_{\omega\in\Omega}|\mathcal{U}(\omega)|GM(\omega)$$

$$GM(\omega)=\Big(\prod_{j\in\mathcal{J}}\prod_{i\in\mathcal{U}_j(\omega)}\lambda_i(\omega)\Big)^{\frac{1}{|\mathcal{U}(\omega)|}} \qquad \forall\omega \tag{2.12a}$$

$$a^{m,t}_{n,i}(\omega)\in\{0,1\} \qquad \forall i,m,t,n,j,\omega \tag{2.12b}$$

$$\sum_{i\in\mathcal{U}_j}a^{m,t}_{n,i}(\omega)\leq 1 \qquad \forall j,m,t,n,\omega \tag{2.12c}$$

$$\sum_{n\in\mathcal{N}}x^{m,t}_{n,j}(\omega)\leq 1 \qquad \forall j,m,t,\omega \tag{2.12d}$$

$$\lambda_i(\omega)\leq\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}}\sum_{n=1}^{N}r^{m,t}_{n,i}(\omega) \qquad \forall i,j,\omega \tag{2.12e}$$

$$r^{m,t}_{n,i}(\omega)\leq a^{m,t}_{n,i}(\omega)f\Big(\frac{x^{m,t}_{n,j(i)}(\omega)S^{m,t}_i(\omega)}{\sum_{q=1}^{n-1}x^{m,t}_{q,j(i)}S^{m,t}_i(\omega)+\sigma^2+I^{m,t}_i(\omega)}\Big) \qquad \forall i,m,t,n,\omega \tag{2.12f}$$

$$r^{m,t}_{n,i}(\omega)\leq\sum_{u\in\mathcal{U}_j}a^{m,t}_{\varrho,u}(\omega)f\Big(\frac{x^{m,t}_{n,j(i)}(\omega)S^{m,t}_u(\omega)}{\sum_{q=1}^{n-1}x^{m,t}_{q,j(i)}S^{m,t}_u(\omega)+\sigma^2+I^{m,t}_u(\omega)}\Big) \quad \forall\varrho\in\{1,...,n-1\},i,m,t,n,\omega \tag{2.12g}$$

$$S^{m,t}_i(\omega)=\sum_{j\in\mathcal{J}}y_{i,j}(\omega)p^{m,t}_j g^{m,t}_{i,j}(\omega) \qquad \forall i,m,t,\omega \tag{2.12h}$$

$$I^{m,t}_i(\omega)=\sum_{j\in\mathcal{J}}\big(1-y_{i,j}(\omega)\big)p^{m,t}_j g^{m,t}_{i,j}(\omega) \qquad \forall i,m,t,\omega \tag{2.12i}$$

$$\sum_{l\in\mathcal{L}}\beta_l M_l\leq P_j \qquad \forall j \tag{2.12j}$$

$$p^{m,t}_j=\beta_{1+(s(j)+m+t-3) \mod L} \qquad \forall j,m,t,\omega \tag{2.12k}$$

$$a^{m,t}_{n,i}\leq z_{n,i} \qquad \forall i,n,m,t \tag{2.12l}$$

The objective function of $\mathbf{P}_3(\Omega)$ is the weighted average of the geo-mean throughputs computed over the calibration set of realizations Ω. The weight is the number of users in a given realization since the geo-mean throughput decrease as the number of users increase and this modification prevents the solver from returning a solution biased towards lightly loaded scenarios.

This calibration problem can be transformed into a signomial program and solved the same way as for the centralized problem (though it is larger). Note that this problem has to be solved *offline* in a planning phase and hence its complexity is not a major issue.

2.6.1 Numerical Results

We consider the same network as before and focus on the power map computations. We show that the power map computed for OMA for the pedestrian setting works well for OMA in the vehicular setting and for NOMA in both settings. The performance results for two distinct single settings, i.e., a pure pedestrian and a pure vehicular are shown. Next, a mixed setting consisting of both pedestrian and vehicular users is studied. Similar conclusions have been seen in irregular (non-hexagonal) network deployments.

We test the power map approach for $L = 2, 3$, and 5. The power levels are calibrated by using a calibration set Ω containing 60 realizations with different average numbers of users per base-station, i.e., 5, 10 and 15 users. We test the power maps with another set of 100 realizations. All shown results are averaged over these 100 test realizations.

Power map robustness

For a given L, 6 power maps have been computed (one for OMA and pedestrian, one for OMA and vehicular, one for NOMA-2 and pedestrian). The power maps for all the cases are given in Table 2.2. The six power maps are very similar and hence, for a given L, the power map obtained for OMA and pedestrian was used in all settings going forward. To test the robustness of the power map calibrated for OMA for the pedestrian setting, we apply it to the 5 other scenarios and show the differences in performance in Table 2.2. Strikingly, the results show that the losses are minimal and this power map is robust to the setting and the scheme.

Performance for two distinct "pure" settings

Fig. 2.10 (resp. Fig. 2.11) shows the results for the pure pedestrian (resp. vehicular) setting for OMA, NOMA-2 and NOMA-3. For each case, for an average of 10 users per cell, there

Table 2.2: Values of the β_l's for different schemes and settings. The last column shows the performance loss when using the OMA pedestrian power map instead of the power map specifically computed for a setting and a scheme.

L	**Scheme**	**Pedestrian**		**Vehicular**	
		$[\beta_l]/\frac{P_j}{M}$	**Performance Loss**	$[\beta_l]/\frac{P_j}{M}$	**Performance Loss**
2	OMA	[0.1 0.9]	-	[0.12 0.88]	0.1%
	NOMA-2	[0.1 0.9]	0%	[0.16 0.84]	0.2%
	NOMA-3	[0.1 0.9]	0%	[0.16 0.84]	0.3%
3	OMA	[0.04 0.06 0.9]	-	[0.06 0.09 0.85]	0.7%
	NOMA-2	[0.04 0.06 0.9]	0%	[0.06 0.09 0.85]	0.6%
	NOMA-3	[0.04 0.06 0.9]	0%	[0.07 0.09 0.84]	0.6%
5	OMA	[0.001 0.05 0.049 0.05 0.85]	-	[0.05 0.05 0.10 0.05 0.75]	0.3%
	NOMA-2	[0.001 0.05 0.049 0.05 0.85]	0%	[0.05 0.05 0.05 0.05 0.80]	0.2%
	NOMA-3	[0.001 0.05 0.049 0.05 0.85]	0%	[0.05 0.10 0.05 0.05 0.75]	0.2%

are 5 results corresponding to equal power allocation, optimal power allocation (referring to the upper bound computed by solving the centralized problem) and three power maps (for $L = 2, 3$ and 5). For each value of L, a power map is computed for OMA pedestrian and is used for OMA, NOMA-2 and NOMA-3 for both settings. Given a power map, the scheduling for each realization is done optimally and locally as explained previously.

Focusing first on Fig. 2.10, there is a significant improvement in performance when using power maps instead of equal power allocation for all cases. For OMA, the gains are 49.7%, 70.4% and 89.1% for $L = 2, 3$ and 5 respectively with respect to equal power allocation. For NOMA-2 (resp. NOMA-3) the gains are $43\%, 59\%, 72\%\%$ $(40\%, 54\%, 67\%)$ with respect to equal power NOMA-2 (resp. NOMA-3). The gaps between the optimal power allocation and the best power map are 25%, 20% and 19% for OMA, NOMA-2 and NOMA-3 respectively.

Comparing NOMA-2 (resp. NOMA-3) over OMA, a gain of 38.5% (resp. 49%) with respect to equal power allocation and 21.1% (resp. 29.6%) is obtained when a power map with $L = 5$ levels is applied. The gain of NOMA-3 over NOMA-2 is between 7.6% for equal power allocation and 6.9% for the power map with $L = 5$.

Focusing now on Fig. 2.11, it is seen that the achieved rates are higher due to the shorter fade duration and that the gains due to the power map over equal power are less significant. For OMA, the gains with respect to equal power are 9.9%, 13.5%, 15.4% for $L = 2, 3$, and 5 respectively. For NOMA-2 (resp. NOMA-3), the gains with respect to equal power, are $3.7\%, 6\%, 6.6\%$ (resp. $2.3\%, 3.2\%, 3.9\%$). The gap between the optimal power allocation and the best power map is: 22.1%, 11.4% and 11.1% for OMA, NOMA-2 and NOMA-3 respectively. Comparing NOMA-2 (resp. NOMA-3) over OMA, a gain of

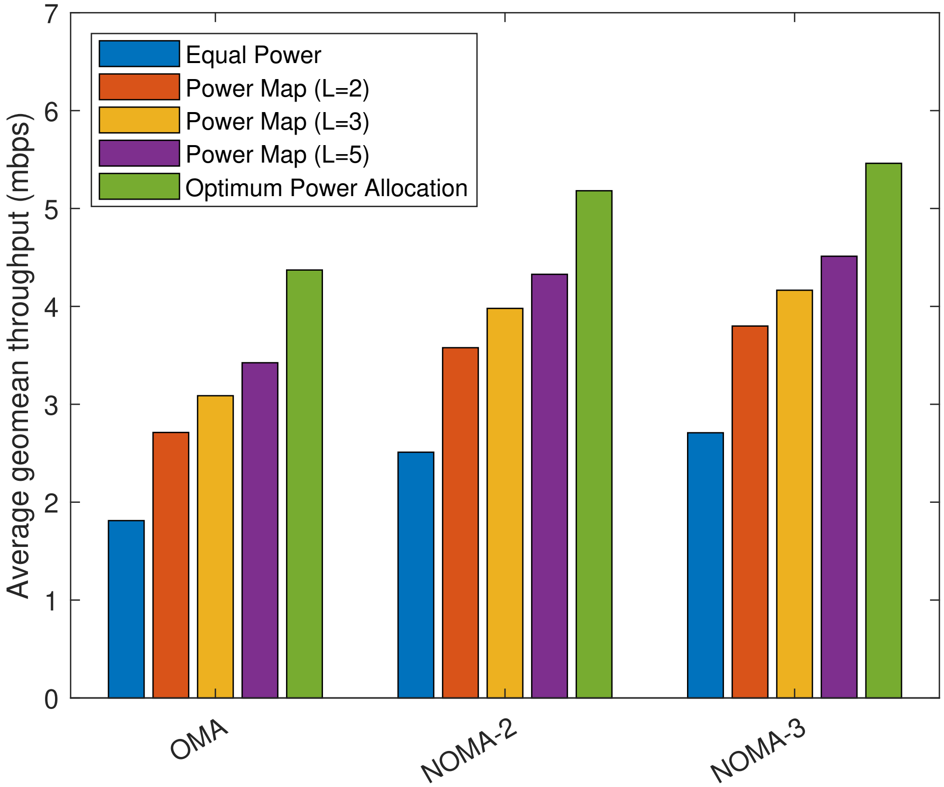

Figure 2.10: Power map performance for OMA, NOMA-2, NOMA-3 for the pedestrian setting when the average number of users per cell is 10 and all users have full SIC capability.

29.9% (resp. 36.1%) with equal power allocation and 20.3% (resp, 22.4%) with a power map with $L = 5$ levels are observed. The gain of NOMA-3 over NOMA-2 is between 4.7% for equal power and 2.9% for power map with $L = 5$.

The following results show that the benefits of NOMA versus OMA and of power map versus equal power are very dependent on the setting. In a real scenario, there would be a mix of pedestrian and vehicular users. Hence, mixed settings are studied next.

Performance for mixed settings

In the following, a mixed setting is assumed where 20% of the users are vehicular users, i.e, their channels have smaller coherence blocks while the remaining users are pedestrians.

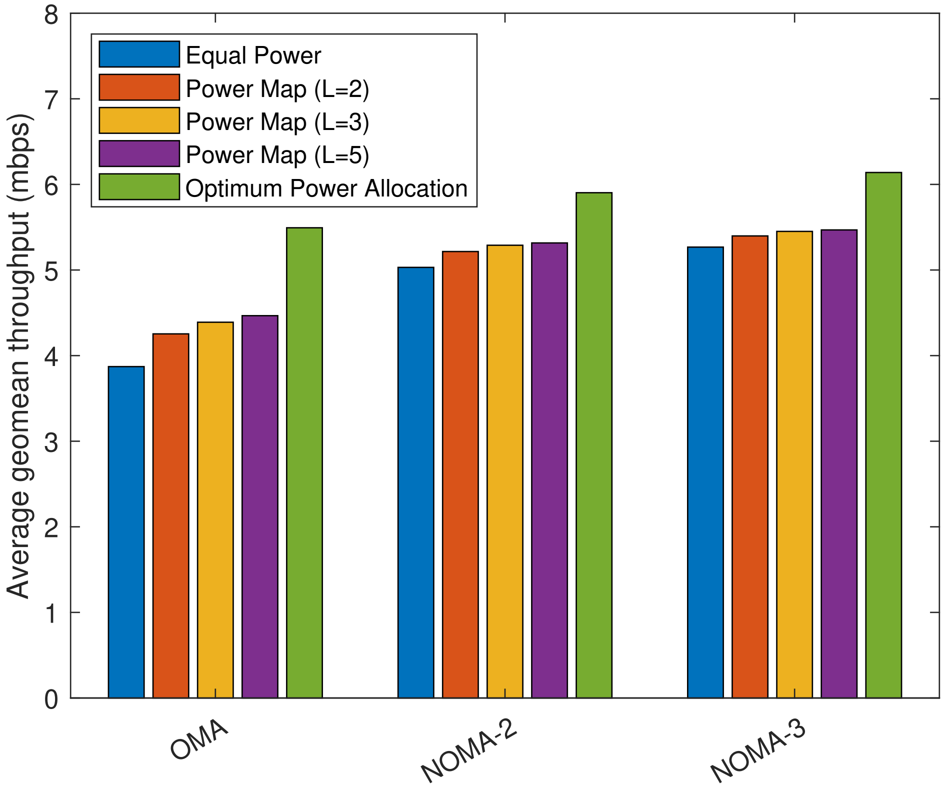

Figure 2.11: Power map performance for OMA, NOMA-2, NOMA-3 for the vehicular setting when the average number of users per cell is 10 and all users have full SIC capability.

We compare the performance of OMA, NOMA and hybrid NOMA for this setting.

Fig. 2.14 gives the gain of NOMA-2 and NOMA-3 over OMA when using the same power map with $L = 5$ as a function of the average number of users per cell for an hybrid case and a non-hybrid one. The hybrid NOMA-2 case has only 50% of the UEs capable of doing SIC and for the hybrid NOMA-3 case, each UE is equally likely to have SIC-2, SIC-1 or no SIC capabilities. The results show that the gain increases with the number of users per cell since more users means more options for the scheduler to make use of NOMA and enhance the system performance. The gain of using NOMA-3 over NOMA-2 also increases with the number of users. Finally, a hybrid mix of SIC capability has a significant impact on the performance.

Figs. 2.12 and 2.13 compare the performance gain of NOMA-2 with power map with

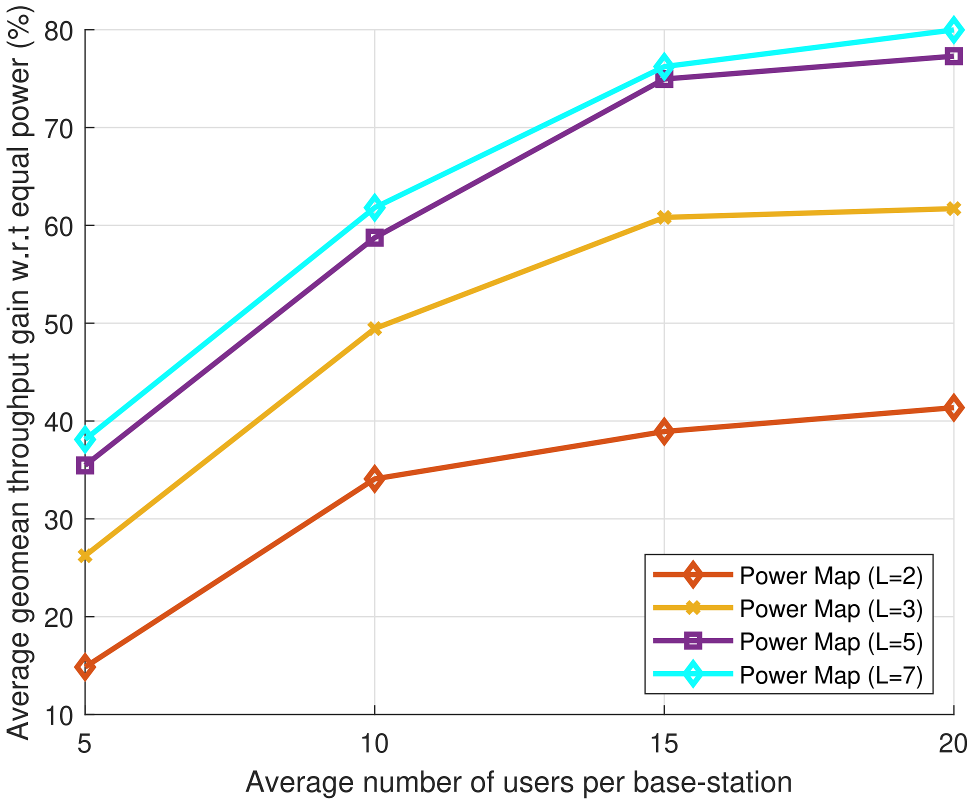

Figure 2.12: NOMA-2 geometric mean throughput gain for power map relative to equal power allocation versus the average number of users per cell for the mixed setting.

$L = 2, 3$, and 5 over NOMA-2 with equal power allocation. The gains are significant. As expected, the more power levels the better the performance since the system have more degrees of freedom to optimize over. The gains also increase with the number of users until they reach a plateau. Similar results were obtained for NOMA-3 as seen in Fig. 2.13. We also performed a similar numerical campaign for the case of an irregular non-hexagonal network. It led to similar conclusions.

In Fig. 2.15, the geometric mean throughput for hybrid NOMA-2 assuming different SIC-1 users ratio per cell for different power maps. We conclude the following about hybrid NOMA. First, the geometric mean throughput increases with the SIC capability probability of the served users, where we define this probability to be the ratio of users with SIC 1 capability to the number of served users. Second, the proposed power map

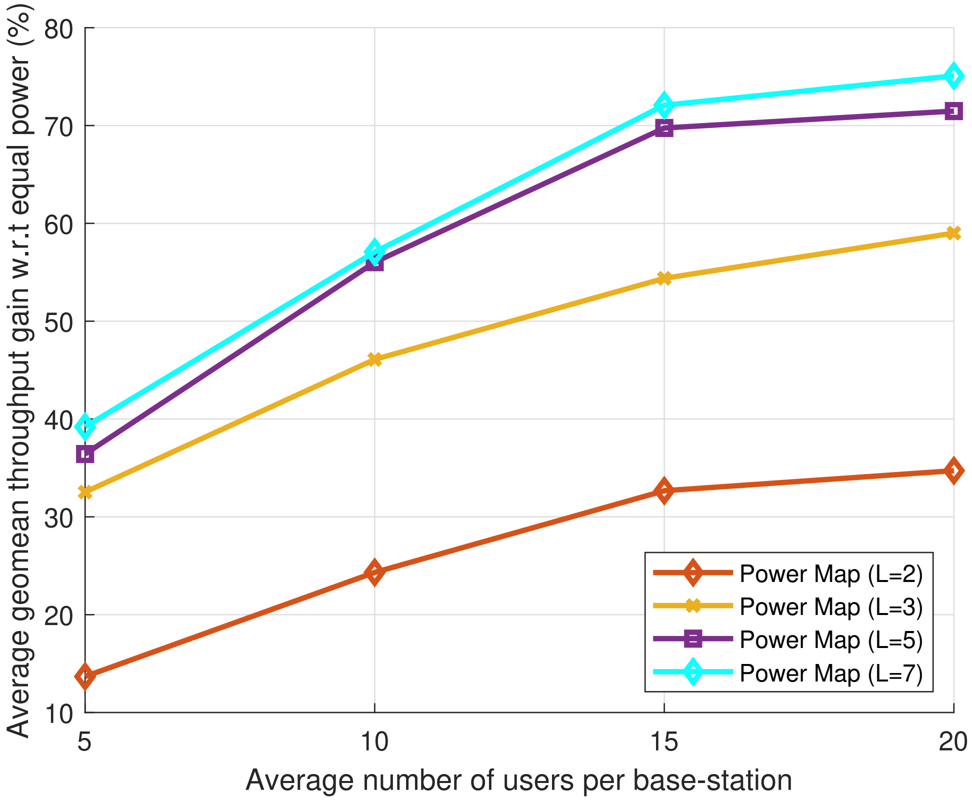

Figure 2.13: NOMA-3 geometric mean throughput gain with power map relative to equal power allocation versus the average number of users per cell for the mixed setting.

brings an approximately equal gain to OMA, Hybrid NOMA and NOMA which means that the present solution is robust to the choice of multiple access scheme as well.

In summary, NOMA-2 is likely good enough since NOMA-3 provides only 8% more gain compared to NOMA-2 considering the associated hardware complexity, and that NOMA-2 with a power map with $L = 5$ performs significantly better than NOMA-2 with equal power allocation (the state-of-the-art). In the next section, NOMA-N ($N = 2, 3$) is used with a power map and a practical local scheduling scheme is proposed for NOMA-2 and NOMA-3 that is of low complexity and provides good performance.

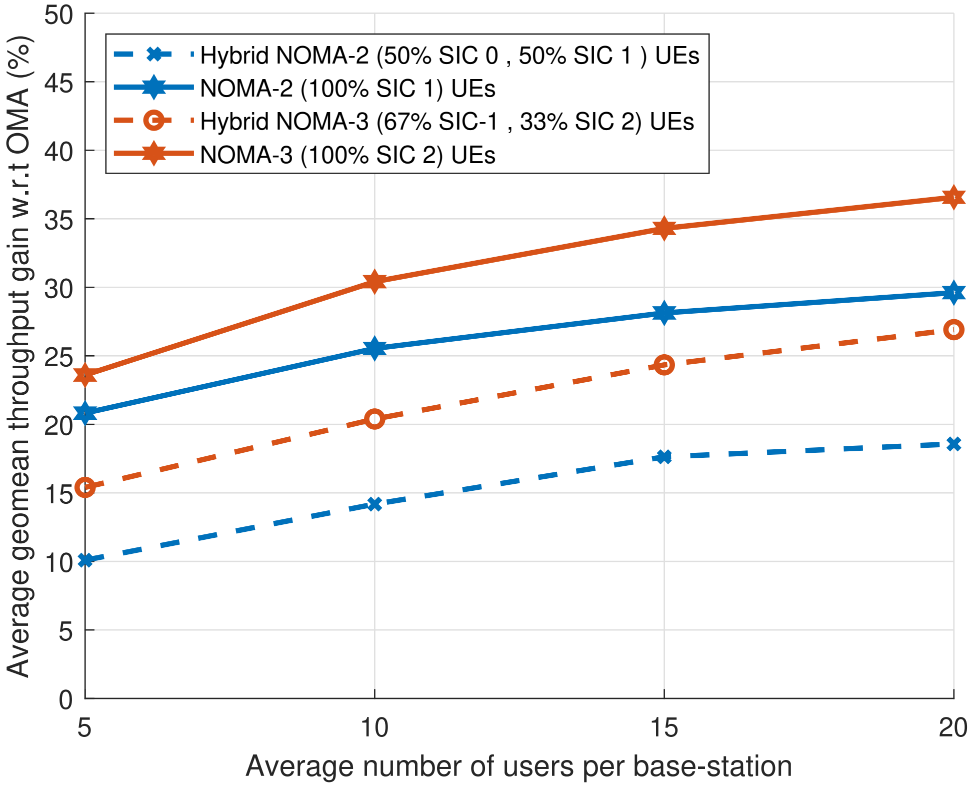

Figure 2.14: NOMA vs. OMA gain in geometric mean throughput with a power map with $L = 5$ versus the average number of users per cell for the mixed setting

2.7 Practical Local Scheduling for NOMA with Power Map

2.7.1 Design of Online Algorithms

Assuming a pre-computed static power allocation as presented in Section 2.6 (this includes equal power allocation), the centralized scheduling problem $\mathbf{P}_1(\omega)$ can be *decoupled* into local scheduling problems, to be performed at each base-station independently, without loss of optimality. These local problems only require the OMA SINR estimates per PRB at each base-station (hence NOMA does not incur a larger signaling overhead). In the following, the local hybrid NOMA-N scheduling problem is formulated on a given frame

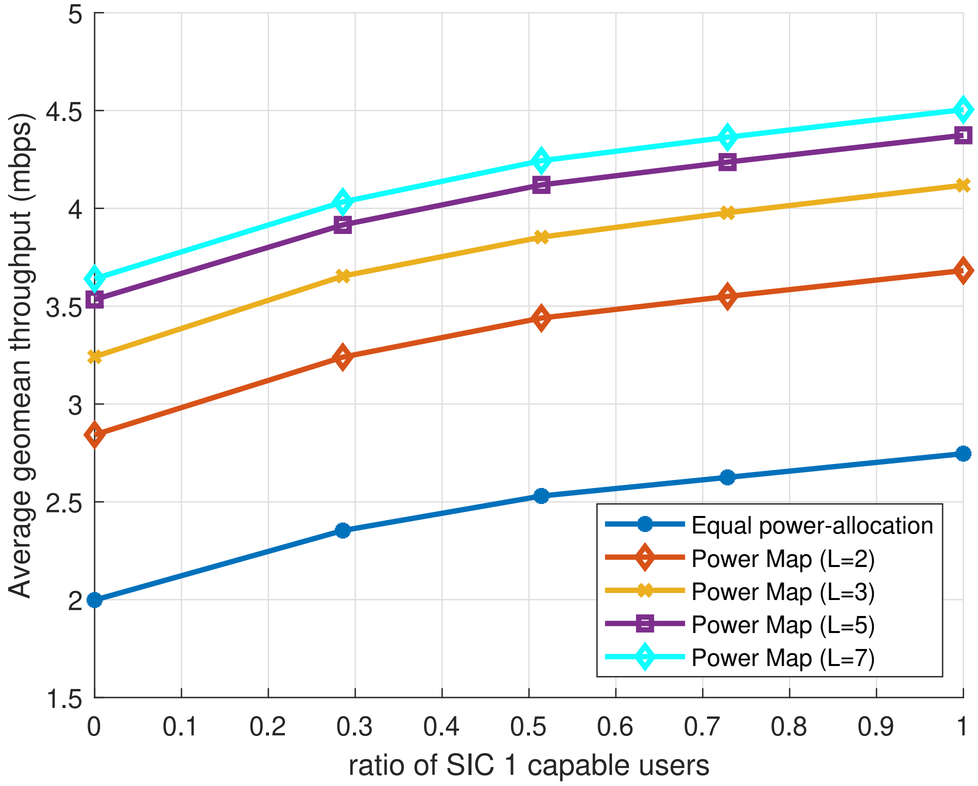

Figure 2.15: Comparison of hybrid NOMA-2 geometric mean throughput for various various SIC-1 capable users ratios, assuming an average of 10 users per base station.

as $\mathbf{P}_4(\delta)$ for a given single cell system realization δ which contains all the per PRB SINRs, γ_i^c's. In this problem, for ease of notation, the PRB double index (m, t) is replaced by a single index $c = m + M(t-1)$ and the *local* scheduling problem is formulated in terms of the OMA SINR values γ_i^c directly. Furthermore, as the scheduling problem is now a local problem, the index j is omitted for simplicity. Note also that all variables except the $(a_{n,i}^c)$'s are non-negative and real. $\mathbf{P}_4(\delta)$ is a much smaller MINLP than $\mathrm{P}_1(\omega)$. However, it cannot be solved fast enough to be practical. Instead of trying to compute the schedule for one frame at a time, we will follow a sequential approach. A scheduling decision is computed for each PRB sequentially in a myopic fashion while keeping track of history to provide fairness on a period possibly larger than a frame. This is in line with state of the art OMA schedulers that are opportunistic, sequential and aim to maximize proportional fairness at each step considering the past but without considering the future (i.e., the rest

$\mathbf{P}_4(\delta)$: Local NOMA-N scheduling problem (per frame)

$$\max_{x_n^c, a_{n,i}^c, \lambda_i, r_{n,i}^c} \prod_{i \in \mathcal{U}} \lambda_i$$

$$a_{n,i}^c \in \{0,1\} \qquad \forall n, c, i \tag{2.13a}$$

$$\sum_i a_{n,i}^c \leq 1 \qquad \forall n, c \tag{2.13b}$$

$$\sum_{n=1}^{N} x_n^c \leq 1 \qquad \forall c \tag{2.13c}$$

$$a_{n,i}^c \leq z_{n,i} \qquad \forall n, c, i \tag{2.13d}$$

$$\lambda_i \leq \sum_c \sum_{n=1}^{N} r_{n,i}^c \qquad \forall i \tag{2.13e}$$

$$r_{n,i}^c \leq a_{n,i}^c f\Big(\frac{x_n^c}{\sum_{q=1}^{n-1} x_q^c + 1/\gamma_i^c}\Big) \qquad \forall n, c, i \tag{2.13f}$$

$$r_{n,i}^c \leq \sum_u a_{\varrho,u}^c f\Big(\frac{x_n^c}{\sum_{q=1}^{n-1} x_q^n + 1/\gamma_u^c}\Big) \qquad \forall n, c, i, \varrho \in \{1, ..., n-1\} \tag{2.13g}$$

of the frame). Consider PRB c and let the rate seen by user i over the window including PRB c and the past W PRBs, be

$$\lambda_i(c) = W \bar{R}_i^c + r_i^c, \tag{2.14}$$

where r_i^c is the potential rate user i can receive if allocated PRB c (this rate will depend on its place in the ordered set), W is the fairness window chosen by the operator, and $\bar{R}_i^c$ is the average per PRB rate seen by user i in the past W PRBs. Then the optimum scheduling decision in PRB c is the solution to a problem with the following objective function:

$$\begin{aligned}
\underset{a^c_{n,i},x^c_n}{\operatorname{argmax}} \log(\lambda_i) &= \underset{a^c_{n,i},x^c_n}{\operatorname{argmax}} \sum_{i\in\mathcal{U}} \log\bigl(W\bar{R}^c_i + r^c_i\bigr) \\
&= \underset{a^c_{n,i},x^c_n}{\operatorname{argmax}} \sum_{i,n} \log\left(W\bar{R}^c_i + a^c_{n,i} f\Bigl(\frac{x^c_n}{\sum_{q=1}^{n-1} x^c_q + 1/\gamma^c_i}\Bigr)\right) \\
&= \underset{a^c_{n,i},x^c_n}{\operatorname{argmax}} \sum_{i,n} \log\left(1 + \frac{a^c_{n,i}}{W\bar{R}^c_i} f\Bigl(\frac{x^c_n}{\sum_{q=1}^{n-1} x^c_q + 1/\gamma^c_i}\Bigr)\right) \\
&\approx \underset{a^c_{n,i},x^c_n}{\operatorname{argmax}} \sum_{i,n} \frac{a^c_{n,i}}{\bar{R}^c_i} f\Bigl(\frac{x^c_n}{\sum_{q=1}^{n-1} x^c_q + 1/\gamma^c_i}\Bigr) \qquad (2.15)
\end{aligned}$$

where we have used the approximation $\log(1+x) \approx x$ for $x \ll 1$ since $W \gg 1$, and we have removed W from the optimization since it is a constant common to all users. Hence, in PRB c, the scheduler would solve the sequential problem $\mathbf{P}_5(\delta)$ below.

$\mathbf{P}_5(\delta)$**:** Local sequential NOMA-N scheduling problem (for PRB c)

$$\max_{a^c_{n,i},x^c_n} \sum_{i,n} \frac{a^c_{n,i}}{\bar{R}^c_i} f\Bigl(\frac{x^c_n}{\sum_{q=1}^{n-1} x^c_q + 1/\gamma^c_i}\Bigr) \qquad s.t.\ (2.13a) - (2.13d)$$

Problem $\mathbf{P}_5(\delta)$ is still an MINLP but of reduced size. Note, that in practice, in OMA local scheduling algorithms that are also sequential, to avoid tracking the detailed rate history of each user, the exact average rate $\bar{R}^c_i$ over the window of size W is replaced by the exponential moving average of the rates over the period, i.e., it is updated as

$$\bar{R}^{c+1}_i = \Bigl(1 - \frac{1}{W}\Bigr)\bar{R}^c_i + \frac{1}{W}\sum_{n=1}^{N}(a^c_{n,i})^* f\Bigl(\frac{(x^c_n)^*}{\sum_{q=1}^{n-1}(x^c_q)^* + 1/\gamma^c_i}\Bigr) \qquad (2.17)$$

where $(a^c_{n,i})^*$ and $(x^c_n)^*$ denote the solution selected by the local scheduler. We adopt the same practice in the following.

A local scheduler that solves $\mathbf{P}_5(\delta)$ performs user selection and power distribution jointly and might not be fast enough (it has to be solved for each PRB). Thus, the problem is further decoupled into two sequential independent problems: *user selection* followed by

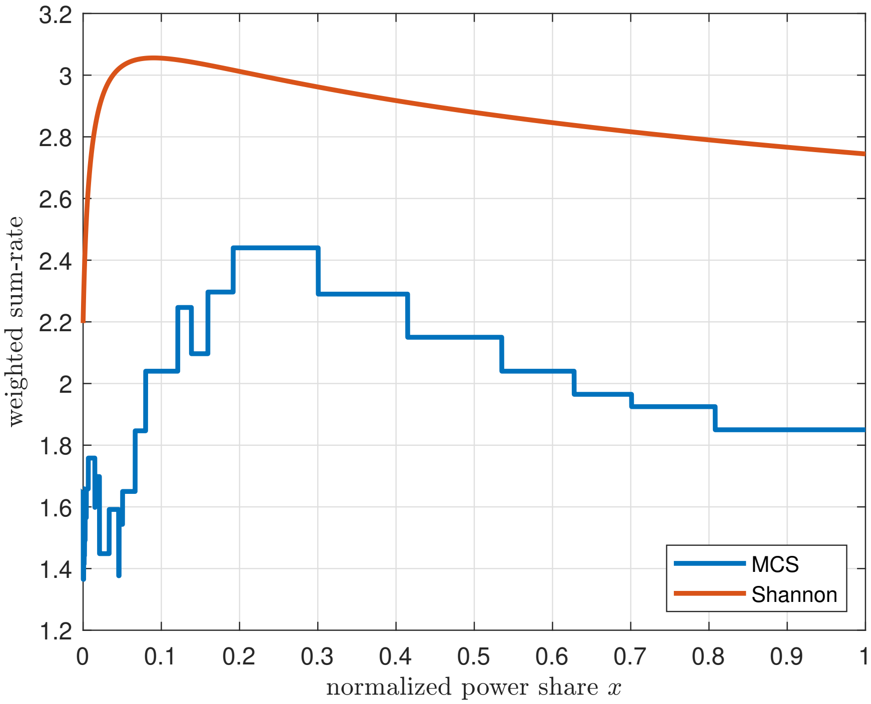

Figure 2.16: An example of a possible problem with the Shannon-based power distribution scheme to be used as an approximation for the MCS-based rate function generated assuming $\bar{R}_{i_1} = 3$, $\bar{R}_{i_2} = 2$, $\gamma_{i_1} = 300$ and $\gamma_{i_2} = 20$.

power distribution given the selected users. The goal of the *user selection* problem is to propose a set of tuples $(i_1, \ldots, i_N)$, while the goal of the power distribution is to compute for each proposed tuple, the optimal choice of power levels, i.e. the x_n values. An optimal solution to $\mathbf{P}_5(\delta)$ can only be determined using an Exhaustive Search (ES) user selection [57], i.e., proposing all possible tuples and finding for each tuple the unique combination of x_n. Greedy search (GS), presented in [43], finds a good sub-optimal user-selection for NOMA-2 by only proposing all the pairs of users that contain the user with the highest $\frac{f(\gamma_i^c)}{\bar{R}_i^c}$. This method, along with an optimal power distribution, achieves a quasi-optimal performance, however, it is complex and cannot easily be extended to higher order NOMA.

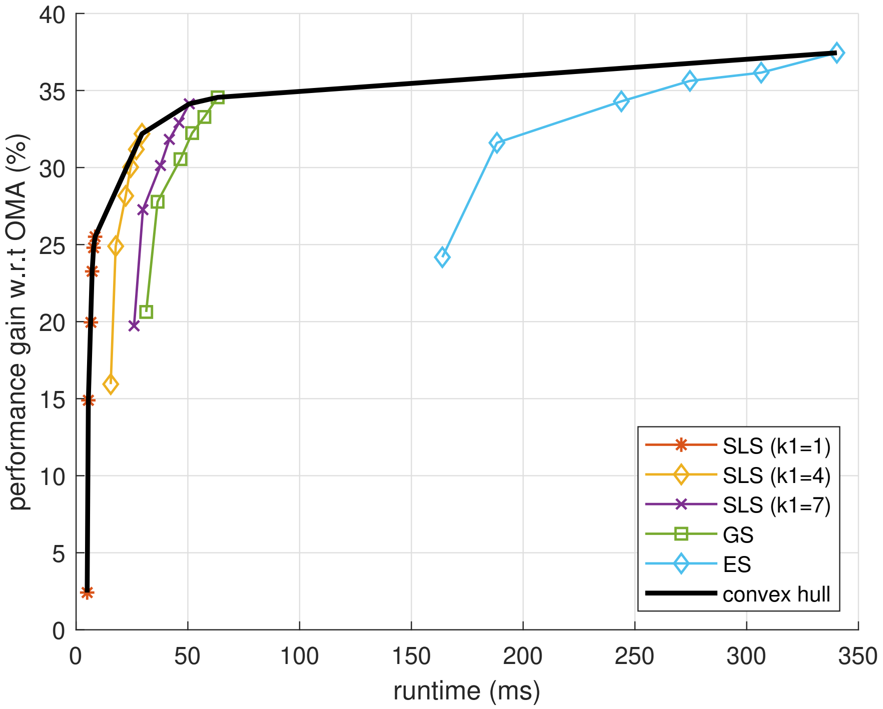

Figure 2.17: The achieved performance compared to OMA assuming equal power allocation and an average of 10 users per cell vs. runtime. For each user selection scheme, power distribution with $k_2 \in \{0, 2, 4, 8, 12, 16\}$ is tested.

```
Algorithm 1: Simplified Local Scheduling (SLS) for NOMA-2 in PRB $c$
Given: $k_1$, $k_2$, $(\gamma_i^c)$, $(\Gamma_v)$, $(\bar{R}_i^c)$; // returns a sorted list of users
sorted_users = Sort($f(\gamma_i^c)/\bar{R}_i^c$)); // builds the chain of $k_1$ pairs to test
chain = BuildUserPairChain (sorted_users,$k_1$); // returns the first pair,
    i.e., $(i_1, i_2)$
(test_pd, test_score) = ComputePowerDistribution (test_pair) ;
(selected_score, selected_pair, selected_pd) = (test_score, test_pair, test_pd) ;
number_of_tested_pairs = 1 ;
// end of initialization
while number_of_tested_pairs $\leq k_1$ do
    test_pair = GenerateNextPair (test pair, Chain);
    (test_pd, test_score) = ComputePowerDistribution (test_pair) ;
    if test_score > selected_score then
        (selected_score, selected_pair, selected_pd) = (test_score, test_pair, test_pd) ;
    end
    number_of_tested_pairs = number_of_tested_pairs + 1;
end
compute $\bar{R}_i^{c+1}$ using eq.(2.17);
return (selected pair, selected pd) ;
Function ComputePowerDistribution $(i_1, i_2)$:
    $x$ = Solve (2.18); // value of (2.15) given pair and $x$
    test_score = $\frac{x\gamma_{i_1}^c}{\bar{R}_{i_1}^c} + \frac{1}{\bar{R}_{i_2}^c}\frac{(1-x)}{x+1/\gamma_{i_2}^c}$ ; // computes the discontinuity points
    $\xi[v] = \Gamma_v/\gamma_{i_1}\ \forall v$ ; // sort $\xi[v]$ (nearest to farthest from $x$)
    $\xi'$ = Sort( $||\xi[v] - x||$) ;
    for $index \in \{1, ..., k_2\}$ do
        $y = \xi'[index]$;
        next_score = $\frac{y\gamma_{i_1}^c}{\bar{R}_{i_1}^c} + \frac{1}{\bar{R}_{i_2}^c}\frac{(1-y)}{y+1/\gamma_{i_2}^c}$ ;
        if next_score > test_score then
            ($x$, test_score) = ($y$, test_score) ;
        end
    end
    test_pd = $x$ ;
    return (test_pd, test_score) ;
end
```

We propose a family of solutions for sequential user-selection and power distribution, called Simplified Local Scheduling SLS(k_1, k_2), that strikes different complexity and performance tradeoffs based on the selection of integers k_1 and k_2. We describe SLS(k_1, k_2) for NOMA-2 in the following. The extension to NOMA-3 is straightforward. In a given PRB c, the user-selection part of SLS starts by sorting users in terms of their ratio of achievable OMA rate to the average received rate so far, i.e., $f(\gamma_i^c)/\bar{R}_i^c$ and thus for simplicity we describe SLS assuming $f(\gamma_{i_1}^c)/\bar{R}_{i_1}^c \geq \ldots \geq f(\gamma_{i_U}^c)/\bar{R}_{i_U}^c$. SLS starts with the pair (i_1, i_2) then tests an additional $k_1 - 1$ extra pairs to select the one achieving the highest objective value. The $k_1 - 1$ extra pairs are selected following this *order:* (i_1, i_3),(i_1, i_4), $\ldots$, (i_1, i_U), (i_2, i_3), $\ldots$,(i_{U-1}, i_U), which we call a chain in the following. It should be noted that both ES and GS can be considered as special cases of SLS and that, for NOMA-2, ES corresponds to $k_1 = U(U-1)$ while GS corresponds to $k_1 = U - 1$. For each selected pair, power distribution is performed as described next.

For the *power-distribution* problem, the optimal solution for $\mathbf{P}_5(\delta)$ for a given user-selection (u_1, u_2), assuming a piece-wise rate function, can only be determined using an exhaustive search. In [58], the authors have proposed the following power-distribution strategy

$$\sum_{q=1}^{n} x_q^c = \begin{cases} \frac{\bar{R}_{u_{n+1}}^c/\gamma_{u_{n+1}}^c - \bar{R}_{u_n}^c/\gamma_{u_n}^c}{\bar{R}_{u_n}^c - \bar{R}_{u_{n+1}}^c} & \bar{R}_{u_n}^c > \bar{R}_{u_{n+1}}^c \\ 1 & \bar{R}_{u_n}^c \leq \bar{R}_{u_{n+1}}^c \end{cases} \tag{2.18}$$

which is optimal assuming an unbounded log rate function. However, it is sub-optimal for practical MCS as indicated by the results in the following subsection. This is because, by not taking the MCS levels into account, we are not distributing the power efficiently as seen in the example in Fig. 2.16 where we show the value of the objective function in one PRB for specific values of the rates $\bar{R}$ and the SINRs as a function of $x_1 = x$. The value of x at which the objective function is optimal for Shannon is very bad for the practical case of the piece-wise constant rate function. Hence, using the value given by (2.18) may be sub-optimal and a search around this value can help find a better solution. Note that while the objective function is smooth with Shannon, it is piece-wise constant in practice and hence, it is enough to compute the objective function at the discontinuity points. A discontinuity point is a value of x for which one of the two selected users sees an abrupt change of rate (or similarly, the objective function sees a discontinuity). There are at most $2V$ such points.

In PRB c, for a given pair (i_1, i_2), the power distribution component of SLS starts by computing an initial estimate of the normalized power ratio x^c based on (2.18) and tries the k_2 discontinuity points nearest to the initial estimate x^c. To eliminate redundancies, we only compute V of these discontinuity points corresponding to $\xi_v = \frac{\Gamma_v}{\gamma_{i_1}}$ for all v. SLS sorts these based on the Euclidean distance from the initial estimate x_c, computes the weighted sum-rate for the first k_2 of those discontinuity points and selects the point achieving the largest value. Algorithm 1 on page 48 presents the different steps of SLS(k_1, k_2) in more details.

SLS(k_1, k_2) computes in each PRB for each of the k_1 user-pairs, k_2 values of the weighted sum-rate which greatly reduces the complexity with respect to the ES-based optimal solution that corresponds to SLS$(U(U-1), V)$. We show next how one can strike different trade-offs in terms of performance and run-time by adjusting with k_1 and k_2. Note that, with a slight abuse of notation, we call ES(k_2) as SLS$(U(U-1), k_2)$ and GS(k_2) as SLS$(U-1, k_2)$).

2.7.2 Numerical Results

The performance of SLS(k_1, k_2) is evaluated in the same mixed setting described in Section 2.6. The system is simulated for a duration of 100 frames for 100 different realizations of a multi-cell systems with $J = 7$ base-stations and an average of 10 users per base-station. The UE traffic is generated using the full-buffer model, which is the recommended model for assessing throughput and spectral efficiency by 3GPP [59].

To quantify the complexity/performance trade-offs associated with SLS, we characterized the complexity by the runtime of the C language implementation of the algorithms averaged over 1000 realizations on an Intel core i7-9750H machine clocked at 2.60 GHz. For a given k_1, we compute the average over 100 realizations of the geometric mean throughput for different values of k_2 and record the average run-time. We then plot the gain with respect to OMA versus the run-times instead of with respect to k_2 in Fig. 2.17. The results show i) the high runtime cost of ES and to a smaller extent of GS; ii) the importance of power distribution in terms of run time (this is often ignored in previous works); iii) that SLS provides a continuum of trade-offs depending on the values of k_1 and k_2; iv) that the majority of the NOMA gain with respect to OMA can be achieved in around 20ms; further computation time brings limited additional gain. It is also observed that power distribution based strictly on (2.18) (corresponding to the lowest point in each curve) performs poorly. Finally, SLS can achieve 95% of the optimum performance much faster than ES(k_2) and even GS(k_2).

Table 2.3: Available rates and the corresponding SINR thresholds [1].

SINR Threshold (dB)	-6.5	-4	-2.6	-1	1	3	6.6	10	11.4	11.8	13	13.8	15.6	16.8	17.6
Efficiency (bits/symbol)	0.15	0.23	0.38	0.6	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.9	4.52	5.12	5.55

2.8 Conclusion

In this work, we studied hybrid NOMA-N in frequency selective fading channels while assuming a practical MCS rate function. We formulated a centralized scheduling problem to compute an upper-bound on practical schemes from which we conclude that the geometric mean throughput can be improved by up to 108% with optimal power allocation and scheduling with NOMA-2 over OMA. Motivated by such results, a practical static power allocation is proposed. Strikingly, it performs only 15% worse than the centralized upper-bound irrespective of the multiple access scheme selection (i.e., OMA, NOMA-2 or NOMA-3) and mobility and channel models. Finally, a family of practical (local) scheduling schemes denoted as SLS(k_1, k_2) is proposed. SLS (k_1, k_2) is a generalization of previous user-selection algorithms such as Exhaustive Search and Greedy Search, is suitable for hybrid NOMA-N and introduces a scalable power distribution scheme suitable for practical (i.e., piece-wise constant) rate function.

2.9 Rate functions

The rate $r_i^{m,t}$ seen by user i associated with base-station $j(i)$ on PRB (m, t) is a function of the SINR $\gamma_i^{m,t}$ experienced by the user on that PRB. In practice, the rate function is step-wise. For example, the one given by 3GPP for LTE [1] is provided in Table 2.3 and illustrated in Fig. 2.18. 3GPP provides a table of the modulation coding schemes (MCS) along with the range of SINR values for the latest standard release. Mathematically, it can be written as $r_i^{m,t} = f_{3GPP}(\gamma_i^{m,t})$ a piece-wise constant increasing function with:

$$f_{3GPP}(\gamma) = \begin{cases} 0 & \gamma < \Gamma_1 \\ a_1 & \Gamma_1 \leq \gamma < \Gamma_2 \\ \vdots & \vdots \\ a_K & \Gamma_K \leq \gamma \end{cases} \tag{2.19}$$

where K is the number of available MCSs, a_k is the spectral efficiency in bits/s/Hz for MCS k and the Γ_k's are the SINR thresholds in dB. Let the set $\mathcal{K} = \{1, 2, \ldots, K\}$ be the

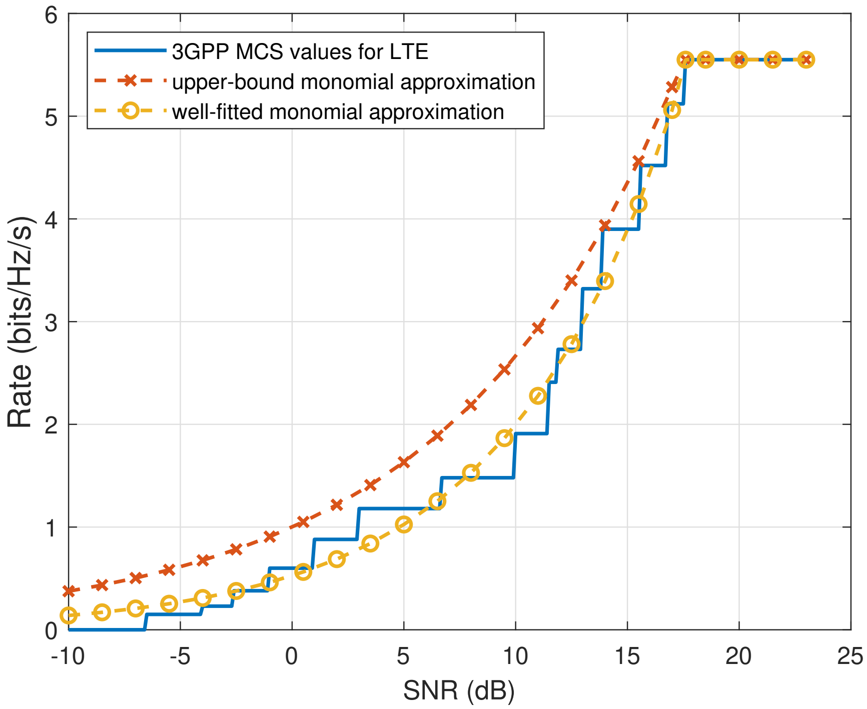

Figure 2.18: Comparison of various rate functions.

set of all possible MCSs. Γ_k is the minimum required SINR for a user to decode a signal with MCS k and receive a rate a_k successfully. We emphasize that Γ_1 is the minimum SINR required for getting a non-zero rate.

Unfortunately, this function is difficult to use in an optimization problem due to its discontinuities and non-convexity. A work-around suggested in [56] is to replace this function by a tight upper bound monomial approximation, i.e., $f_{pow}(\gamma) = \min\big(\eta\gamma^{\alpha}, a_K\big)$.

It is important to note that these approximations are only valid as long as $\eta\gamma^{\alpha}$ does not exceed a_K. Hence, the approximation needs to be truncated accordingly. A comparison of the various rate functions is presented in Fig. 2.18.

In this chapter, the upper bounding monomial approximation $f_{UB}(\gamma) = \min(5.55, \gamma^{0.43})$ to the piece-wise rate function is used to derive an upper bound solution to the centralized joint problem. Additionally, the well-fitted monomial approximation $f_{fit}(\gamma) = \min(5.55, 0.53\ \gamma^{0.58})$ is used to derive a feasible solution for the problem.

Chapter 3

Resource allocation for MU-MIMO systems

In this chapter, we investigate the performance and operation of the downlink of a massive multi-user MIMO single cell system when using ZFT (Zero Forcing Transmission) precoding and the user equipment (UE) has a single antenna.

3.1 Motivation

Massive MU-MIMO (Multi-User MIMO) [12] is a game-changing technology and a strong catalyst for next generation networks that could not otherwise keep up with the ever-growing demand for mobile data. When combined with OFDMA, a base-station equipped with a large number of antennas can simultaneously serve many end-users using the same time-frequency resource. Specifically, in OFDMA, the available bandwidth is divided into sub-channels and time is divided into time-slots where each sub-channel and time-slot pair is called a Physical Resource Block (PRB). OFDMA MU-MIMO enabled-systems can then transmit to multiple users at once in each PRB. However, to reap the benefits of MU-MIMO in practice, well-designed resource allocation processes are needed to efficiently operate the network in real-time. Hence, the trade-off to strike when designing these processes is between efficiency and runtime.

Network operation, i.e., RRM (Radio Resource Management), is the problem of allocating different radio resources such as power, frequency and antennas to different users in a fair manner. Proportional fairness in rates [51] is most commonly used by operators.

With proportional fairness, network resources are allocated in such a way that no single user can increase their rate without decreasing another user's rate by at least the same percentage. A proportional fair set of rates maximizes the geometric mean of the users rates [50].

Specifically, the ultimate goal in this chapter is to design practical algorithms for operating a massive MU-MIMO downlink system serving single antenna UEs using ZFT precoding. To achieve this, we first study the best performance that can be achieved by the system when there is no constraint on the runtime, and we call this an offline study. This offline study gives us insights on the operation and sets a performance target for our practical algorithms.

The best achievable performance is the solution of a challenging problem that jointly optimizes 1) power allocation, 2) user selection, 3) precoding, 4) power distribution and 5) Modulation and Coding Scheme (MCS) selection such that the overall proportional fairness in rate is maximized over a certain horizon. Power allocation is the process that determines the amount of power allocated to a sub-channel by the base-station. Typically, power is allocated equally across all sub-channels on the downlink. With the help of this assumption, the solution to the original problem is approximated by solving a sequence of smaller per PRB problems weighted sum-rate as discussed in Section 3.5.

Focusing now on one of these per PRB problems, the processes to jointly optimize are 1) user selection, 2) precoding, 3) power distribution among the selected users, and 4) MCS selection. This joint optimization problem is NP-hard [60, 61] and cannot be directly solved except for very small systems. However, for a given user selection, precoding can be done first and then, given precoding, the remaining processes can be formulated into a smaller non-convex power distribution problem that can be solved (this is only possible because of the choice of ZFT precoder – not all precoders allow the decoupling of precoding and power distribution). Assuming that this power distribution problem can be solved easily (which is discussed in the next paragraph) for a given user-set, then the original PRB problem can be solved by searching over all user-sets, which is cumbersome except for small systems. *One of the contributions of this chapter* is to explore smart search techniques that limit the number of user-sets that we try. Note that the common belief is that in a massive MIMO system, all active users can be transmitted to in a PRB. While this is indeed possible, we will show that this is not a good idea and we propose a user selection based on grouping users in random groups of size K_{RR} and we discuss a way to determine K_{RR} based on the setting.

We now focus on the joint per PRB problem given a user-set. In our system, there is one transmitted signal stream per selected user (since UEs have a single antenna) and with ZFT, these streams are designed such that there is no inter-user interference. Note that ZFT precoding can easily be computed [61, 62] and the relative gap between the performance of ZFT and capacity diminishes as the number of antennas increase [12]. These reasons make ZFT precoding a strong candidate for practical deployments. Following ZFT precoding, each users sees an effective SISO channel[1] with no inter-user interference.

The next step is power distribution, where the power allocated to a PRB is distributed among the selected users in a manner that maximizes the weighted sum-rate. If a selected user u has an effective channel η_u in the PRB under consideration and is allocated power p_u, then its SNR[2] will be $\eta_u p_u / \sigma^2$ where σ^2 is the noise power. Its rate is determined by the so-called rate function that translates the SNR into a rate. The exact rate function is piecewise constant made of L levels corresponding to the L Modulation and Coding Schemes (MCS) of the system (please see Section 3.4 for more details).

The power distribution problem can then be formulated directly with the exact piecewise rate function or using an approximate rate function which is smooth and concave (e.g., the Shannon rate function). The approximation ignores the discrete nature of the MCS rate function and the fact that there is a certain SNR threshold Γ_1 below which no rate can be received but it allows the problem to be solved using the well-known water-filling method [63, 49, 64]. While there is a significant body of work on MU-MIMO, virtually all use the approximation approach. We show that this is a critical flaw of these studies since due to the threshold Γ_1, many users are often effectively assigned a zero rate even if they have been assigned some power. In that case, power is wasted. *Yet another contribution* is to formulate and solve the power distribution problem using the exact piecewise rate function and hence, get an optimal solution to the joint problem of power distribution and MCS selection.

Irrespective of the rate function being used for power distribution, it is possible that some selected users see an SNR lower than Γ_1, though the approach using the exact rate function would allocate a zero power to those users. In that case, those users will not be transmitted to and thus, they should not be included in the precoding phase, since it makes the effective channels of other users worse than necessary. At the very least, another iteration is needed where these users are removed and precoding is performed on the reduced user-set. Of course, we could remove only a subset of the users that see a zero-rate at a given time to perhaps give non-zero rate to the others in a subsequent iteration. We will show that even a single additional iteration brings a very large performance gain.

[1] We use the terms effective channel and effective SISO channel interchangeably.

[2] Recall that, thanks to ZFT precoding, the interference is been cancelled and hence in the remainder of this chapter, we will work with SNR.

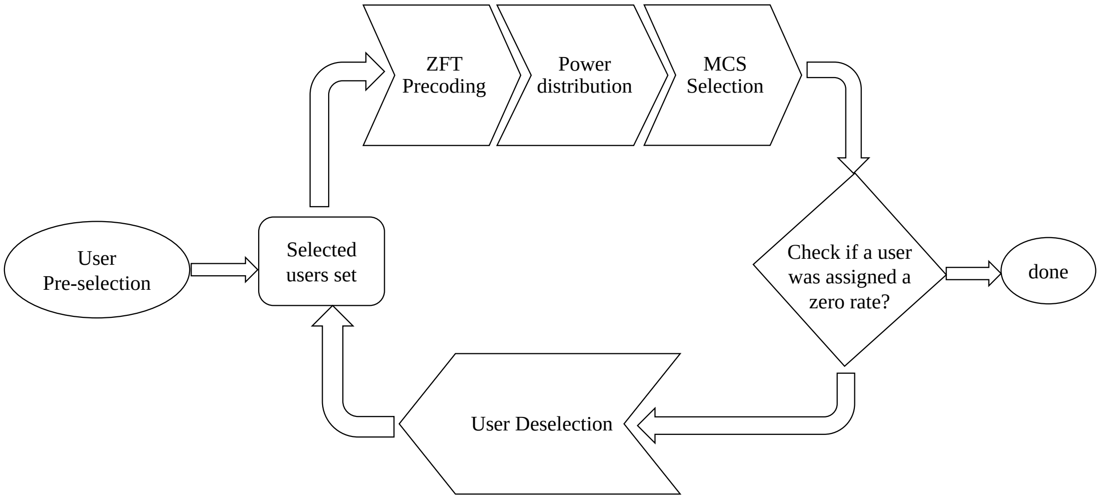

Figure 3.1: The sequence of processes performed per PRB assuming a user-set is preselected.

Thus, by starting with a certain user-set and then removing some or all users with zero rate, the overall process becomes iterative. With extra iterations, we will get better performance at the expense of additional computational complexity. *Another contribution* is to propose and evaluate an iterative solution with a simple de-selection policy when several users see a zero-rate in the first iteration.

A summary of the RRM process sequence, following an initial user selection, is illustrated in Fig. 3.1. Initially a user-set is pre-selected for which a ZFT precoder is computed and then power is distributed among the selected users. The resulting rates after MCS selection are checked to see if there are any users assigned a zero rate and if so, some or all of those users are removed to improve the system performance. This proposed iterative method is shown to outperform the benchmark that does not perform similar iterations.

Finally, based on all the insights obtained from the offline study, we first propose *(yet another contribution)* an MCS-aware greedy algorithm, inspired by the offline study, to solve the power distribution problem quickly using the exact piecewise rate function and show how the end result is quasi optimal and much better than the water-filling approach used in the literature.

The final contribution is the proposal of a real-time network operation solution that is almost as fast as the state of the art benchmark adopted from the literature and provides performance close to the target performance obtained in the offline study. Our solution includes random grouping of users in groups of fixed size K_{RR}, our new MCS-aware greedy power distribution algorithm and a single additional iteration to remove all zero-rate users (we call this Full Drop (FD)). The improvement in performance of our integrated proposed solution with respect to the benchmark can be as high as 233%. We also quantify the performance gain of each of these new steps.

In summary, the main messages are:

1. Because of the existence of an SNR threshold Γ_1 below which non zero rate cannot be obtained, power distribution might yield zero-rate users that have to be deselected to improve the effective channels of other users. Dropping all the zero-rate users at once is found to be good enough.

2. Although in massive MIMO networks all users could be selected, this approach yields poor performance. Astonishingly, random user grouping, with a relatively small group size compared to the number of antennas, is shown to, when used with the proper power distribution and user de-selection method, not only achieve comparable geometric mean throughput performance to the best feasible solution computed offline, but it reduces runtime as well.

3. In order to compute the offline performance target for a given PRB, we propose a set of greedy search methods (to search for a good user-set) as well as an approach based on Branch-Reduce-and-Bound (BRB), where the latter yields an upper-bound to the achievable performance and all yield feasible solutions to the problem. While all the proposed search methods find good performing user-sets in different ways, many of them achieve performance comparable to the feasible solution returned by BRB. This achieved performance is shown to be quasi-optimal for small systems (with less than 30 users) since it is within 10% of the upper-bound computed with BRB. We found out that a particular greedy search method yields good performance within a reasonable runtime (for planning purposes) even if its complexity renders it impractical for real-time deployments.

4. We show that the power distribution computation done under the assumption that the rate function is approximated by Shannon's formula yields wrong insights, in addition to being 22% sub-optimal. Instead of spreading the power thin as recommended with Shannon, the power is focused on a smaller set of users. Our alternative MCS-aware power distribution uses a simple greedy power distribution algorithm and is found to be quasi-optimal.

The remainder of this chapter is structured as follows. A review of the necessary background is presented in Section 3.2. Related work is surveyed in Section 3.3 where the main distinguishing features of our study are highlighted. In Section 3.4, the system model and our notations are presented as well as the joint optimization problem that we seek to solve. We start our offline study in Section 3.5 with the aim of finding the best feasible solution to the resource management problem in an offline setting with no-constraint on runtime. Following this, in Section 3.6, we focus on practical algorithms for real-time network operations, beginning with a description of the state of the art benchmark. Then we develop our practical scheme that provides excellent performance/runtime tradeoffs, i.e., its performance is close to the performance obtained in the offline study with runtime comparable to that of the state of the art benchmark. Numerical results showing the superiority of the proposed scheme compared to the benchmark are presented and discussed in Section 3.7. Finally, the chapter is concluded in Section 3.8.

3.2 Background

3.2.1 Beamforming

Because of its simplicity and flexibility, OFDMA (Orthogonal Frequency Division Multiple Access) is the de facto multiple access mechanism in 4G and 5G cellular systems. In OFDMA, the base-station divides the available frequency and time into physical resource blocks (PRBs) and dynamically assigns PRBs to end users, as shown in Fig. 3.2. The PRB is the smallest scheduling unit. The maximum number of end users scheduled on a PRB in a SISO (Single Input - Single Output) system is one, i.e., it is a SUT system. In the following, we discuss the usage of multiple antennas in order to improve the overall system performance and to allow for MUT transmission.

Beamforming is a wireless communication technology that directs and shapes radio wave signals transmitted and received by multiple antennas. The purpose of beamforming is to increase signal strength while decreasing interference, resulting in improved wireless connection reliability and performance.

Traditionally, omni-directional antennas were used in wireless communication. Because these antennas broadcast signals in all directions, the transmissions are weaker and more susceptible to interference. Sectorization was proposed to overcome these shortcomings. The coverage area was divided into smaller sectors, each of which was covered by a directional antenna.

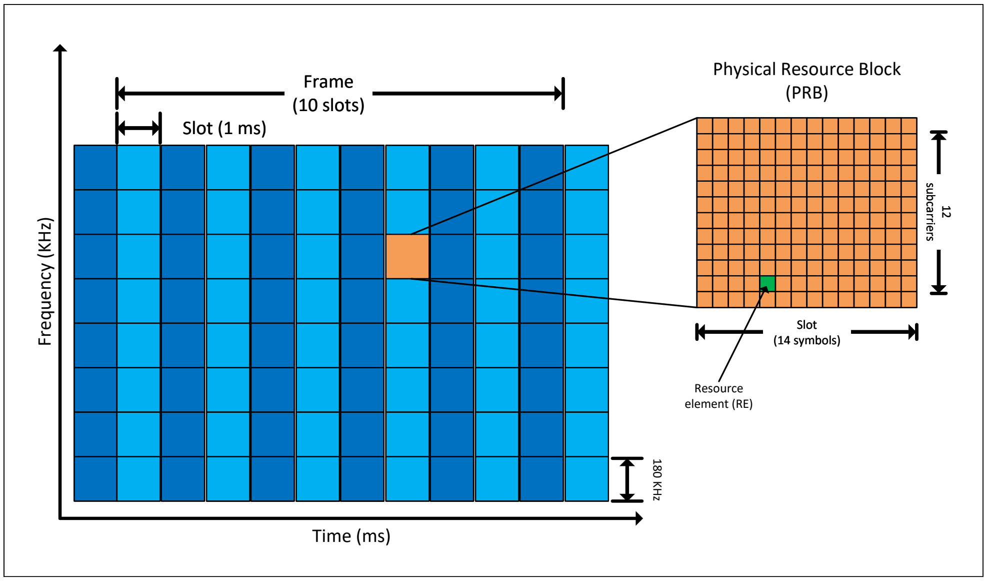

Figure 3.2: OFDMA frame structure for 5G NR standard.

Sectorization was a huge success but it has some limitations. One of them was the antenna's fixed beam direction. The beam direction could not be dynamically altered to match the positions (and the movements) of the receivers, resulting in poor signal quality and increased interference in many cases. Interference from adjacent sectors may also arise as result of overlapping beam patterns. Moreover, the capacity of the network was also limited by the number of sectors and the number of users that could be served by each sector.

Steering antennas was the way forward but originally it was done through a cumbersome mechanical process. To remedy these shortcomings, smart antennas were developed.

The introduction of smart antennas was a major game changer. Smart antennas are multi-antenna arrangements, often known as antenna arrays with multiple antenna elements, that use appropriate signal processing techniques to create and steer beams. The intelligence of the smart antennas is not in the actual antenna, but rather in signal processing. These antennas employ signal processing techniques to dynamically direct radio waves towards the receiver. This enables the antenna to track the receiver's movement and alter the beam direction accordingly. As a result, the link is more robust and reliable, and the limitations of sectorization are addressed.

Adjusting the direction of the signal transmitted from an antenna array can be performed electronically by controlling the phase and the amplitude, i.e., the beamforming

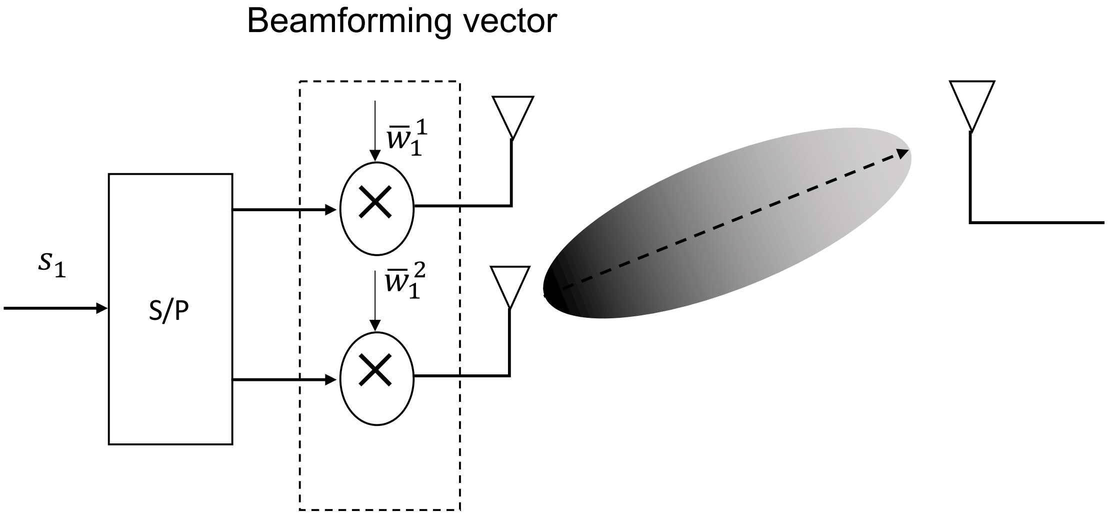

Figure 3.3: An illustration of the beamforming process.

weights, of the signals transmitted by each antenna element. Tuning the phase and amplitude can be performed dynamically in real-time, allowing the radiated radio signal direction to be changed quickly and more accurately. This process of directing the radio signal energy in a specific spatial direction by tuning the phase and the amplitude of the signals transmitted from an antenna array is referred to as beamforming. An illustration of the beamforming process is provided in Fig. 3.3.

Analog, digital, and hybrid beamforming are three types of beamforming techniques used in wireless communication systems. Analog beamforming is the simplest form of beamforming, in which the beamforming weights are applied to the signal at the RF (radio frequency) level using analog components such as phase shifters and power amplifiers. Analog beamforming is less complex and cheaper than digital beamforming, but it has limited flexibility in terms of the beam patterns since it uses analog components, such as phase shifters and power amplifiers, that are designed for specific beam patterns and frequency ranges making them less agile in adapting to changing channel conditions.

More importantly, analog beamforming uses a single RF processing chain. An RF processing chain refers to a sequence of signal processing components and processes used to preprocess and transmit radio signals. It includes all the single processing stages from the digital base-band output of the physical layer to the antenna, including, digital to analog conversion, upconversion, amplification, ... etc. In analog beamforming, a single RF processing is used meaning that only one downlink signal can be transmitted at a time. In contrast, digital and hybrid beamforming allow for the simultaneous transmission of multiple downlink signals from multiple RF processing chains. Hence, MUT transmission is not possible with analog beamforming.

Digital beamforming applies the beamforming weights in the digital domain before they are converted to the analog RF domain. Digital beamforming allows for more precise control over the signal transmission direction since digital signal processing provides the ability to generate a wider range of beamforming patterns. This can be employed to improve the communication reliability, reduce interference and increase the system capacity. However, digital beamforming requires having a number of RF chains equal to the number of available antenna elements making it a more costly solution, especially when the number of antennas is large like for mmWave bands.

Hybrid beamforming is a combination of analog and digital beamforming. In this technique, the phase and amplitude, i.e. the beamforming weights, adjustments are applied to the signals using both analog and digital components. Hybrid beamforming achieves a trade-off with better flexibility than analog in terms of the beamforming patterns that can be generated and lower complexity and cost compared to pure digital beamforming. This is achieved by having a number of RF chains lower than the number of available antenna elements. For further discussion on hybrid beamforming and its different implementation structures and their design trade-offs, the interested reader is referred to [65, 66].

3.2.2 Evolution of MIMO in Cellular Systems

Smart antennas, i.e., combining multiple antenna arrays with beamforming, can be utilized for various objectives; such as: 1) improve robustness to fading, 2) increased coverage, 3) interference suppression, and 4) increase the number of served users. The different antenna elements of a smart antenna can be employed to offer spatial diversity, and thus enhance the system's fading resilience. By coherently forming an antenna pattern that maximizes the receive power at a specific location, coverage may be extended. Alternatively, the antenna pattern can be designed to reduce interference. Multiple antennas can increase the number of served users by decreasing interference and/or serving multiple users located in different directions. It should be emphasized that it is not possible to have all those benefits simultaneously to their fullest extent. For example, smart antennas can be utilized to reduce interference in order to either improve the quality of a single link or have more users in the system or to have a trade-off between the two.

The ability to independently control the data stream fed to each element of the antenna array enables MIMO (Multiple Input - Multiple Output) wireless communications.

In MIMO, which was first proposed by Winters [10], multiple data streams are transmitted simultaneously over multiple antenna elements which boosts the wireless connection robustness and improves the link capacity. The development of practical MIMO techniques over the last 20 years has been one of the primary enablers of 4G/LTE and more recently 5G/NR.

The key enabler of MIMO is precoding. Precoding is a signal processing technique used to shape the transmission of multiple data streams over multiple antennas in a MIMO systems. It involves applying a set of weights, precoding weights, to each data stream before it is transmitted over the wireless link. Additionally, where optimizing the beamforming weights implies that they are chosen to maximize the SINR at the receiver, various cost functions could be defined for optimizing precoding weights as there is a trade-off between different streams.

Precoding is a generalization of beamforming to multiple data streams, where precoding weights serve a similar role as beamforming weights in beamforming. While beamforming is typically used to control the direction and shape of the beam pattern for a single data stream, precoding is used to control the direction of multiple data streams transmitted simultaneously.

During the early phases of MIMO in 4G networks, precoding was performed using a code-book based approach. In this approach, a code-book is predefined, containing a set of precoding matrices. The precoding matrix for a given transmission is selected from this code-book and applied to the data streams prior to transmission. This method is simple and computationally efficient, as the number of precoding matrices in the code-book is limited.

However, the limited number of precoding matrices available for selection results in sub-optimal performance. To overcome these limitations, more advanced precoding methods, such as linear precoding and non-linear precoding, have been introduced. These methods can provide better performance and reliability but at the cost of increased computational complexity.

Precoding can be either linear or non-linear. Linear precoding is a method of precoding in which the precoded data streams are obtained as linear combination of the original data streams. In other words, the precoded data streams are the results of applying a linear precoding matrix to the original data stream as illustrated in Fig. 3.4. Linear precoding is more computationally efficient than non-linear streams and as the number of antennas increase the gap in performance vanishes. With precoding, the precoding matrix is designed based on the channel information and other system constraints, rather than being selected from a predefined code-book. This approach provides better performance as the precoding

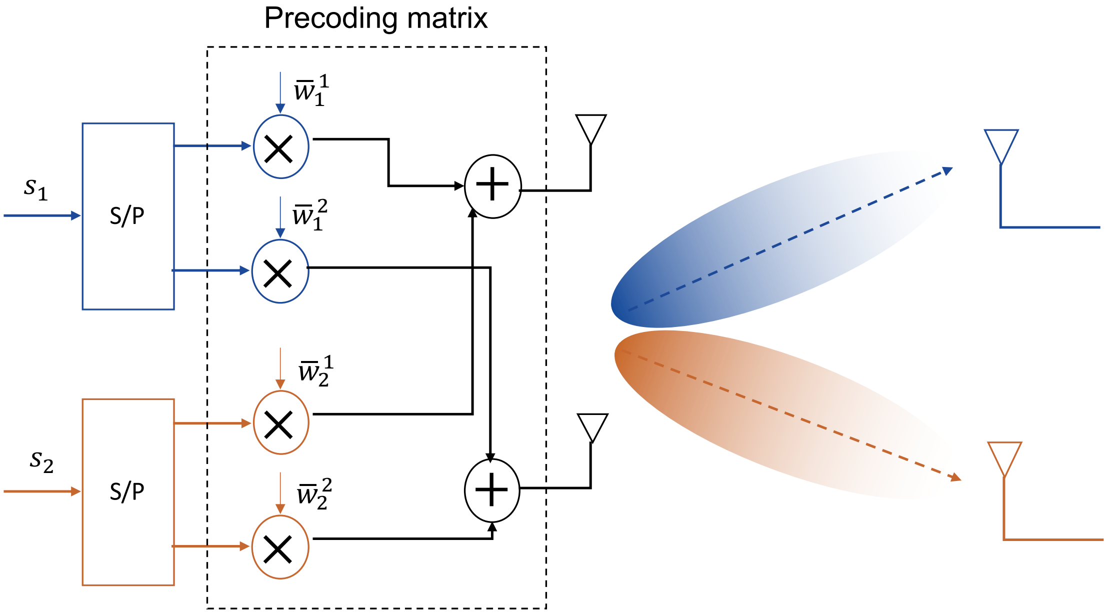

Figure 3.4: An illustration of the linear precoding process.

matrix is optimized for the given channel conditions and/or the system objectives, but it also requires more computational resources and system complexity.

In a cellular environment, MIMO can be utilised in either a SU-MIMO (Single-User MIMO) or a MU-MIMO (Multi-User MIMO) operation mode. In SU-MIMO, the base station employs all of its antenna elements to serve one end user per PRB, whereas in MU-MIMO, the base station connects with several end users simultaneously. An illustration of both concepts is presented in Fig. 3.5.

Because MU-MIMO systems are more flexible, they outperform SU-MIMO counterparts. However, designing practical schemes for MU-MIMO systems is significantly more challenging than developing practical schemes for SU-MIMO systems due to the lack of cooperation amongst end users when decoding in the downlink. In SU-MIMO, all antennas belong to the same UE and thus a single decoder can jointly decode all streams. This is not possible in MU-MIMO since each UE cannot make use of the received copies at other UEs.

In Fig. 3.6, an illustration of a general downlink MU-MIMO system is presented. At each PRB, the base-station selects a number of users to serve as well as the number of streams per user. The users' data streams are precoded using a linear precoding matrix

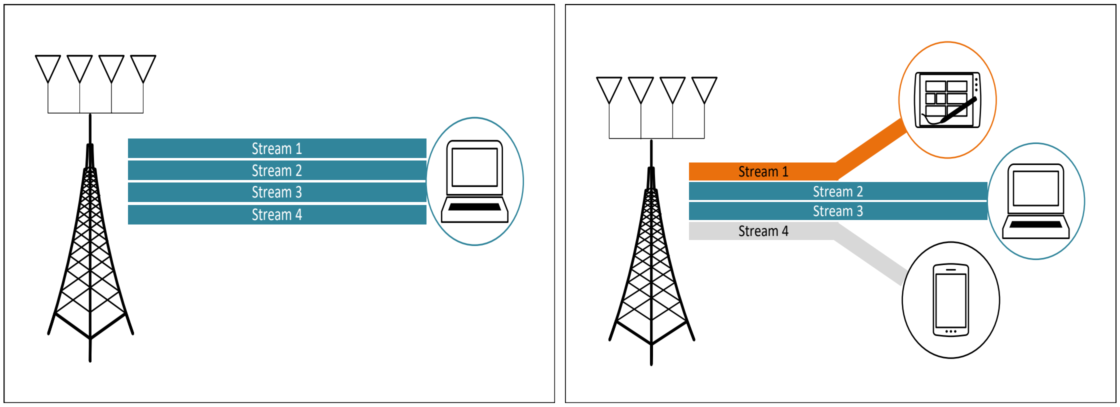

Figure 3.5: An illustration of SU-MIMO (left) and MU-MIMO (right).

before being fed to the base station's transmit antennas. Each user uses their antenna elements to receive the data streams intended for them.

Massive MIMO is a game-changing technology that allows a large number of end users to communicate simultaneously using the entire allocated frequency spectrum. This aggressive spatial multiplexing is enabled by having a greater number of antenna elements than end-users.

In massive MIMO, the large number of antennas at the base-station provides the high antenna array gain and a lot more degrees of freedom for beamforming. This enables the base-station to direct more energy toward the intended user and reduce the interference to other users. This results in improved spectral efficiency and increased coverage as well as reduced power consumption for the users. Moreover, as the number of antennas increase, the channel robustness to small scale fading increases, a property referred to as channel hardening. Furthermore, channel hardening simplifies the processing at the base-station as it allows linear precoding to achieve near-optimal performance as the channel gains between the base-station and the UE become approximately constant [12, 13, 67].

In conclusion, Massive MIMO has the potential to revolutionize wireless communications by providing high capacity at a reasonable level of complexity, making it a key technology for 5G and beyond cellular systems. However, in order to achieve these gains in practice, efficient network operation algorithms are required.

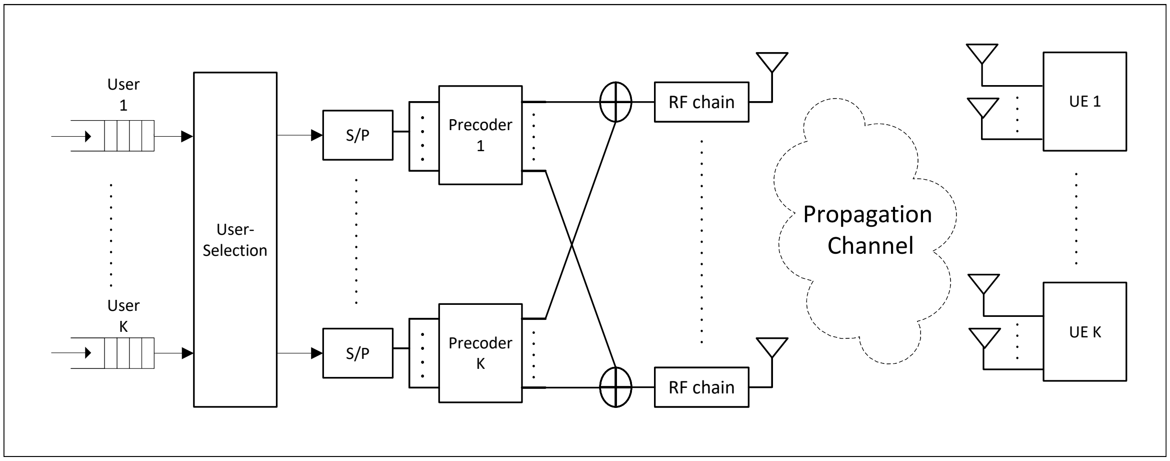

Figure 3.6: Typical MU-MIMO system block diagram.

3.3 Related Work

Because spectrum is scarce, it must be efficiently used, and therefore resource allocation is one of the most essential operational mechanisms in cellular networks. An efficient use of radio resources is one that maximizes an objective function, carefully chosen to trade-off the rates of the users and fairness, such that no user is starved. As mentioned previously, proportional fairness [6] is a common fairness criteria in which the operator chooses a resource allocation at which no relative increase in one user's rate can be achieved without lowering the rate of another user by the same relative amount or more.

As mentioned in the introduction, resource allocation for a massive MIMO cellular system is composed of multiple steps: *power allocation*, *user selection*, *precoding*, *power distribution* and *MCS selection*. In terms of power allocation for OFDMA systems, the defacto approach is to perform equal power allocation per PRB. Equal power allocation, along with approximating the proportional fairness objective as a weighted sum of the users' rate allows the decomposition of the original problem as a sequence of per PRB problems. Although recent works [68, 7] have shown that using power maps can be a practical and more efficient alternative to equal power allocation in multi-cell systems, power maps have not been extended to MIMO multi-cell networks. Since in this chapter we focus on operating a single-cell MU massive MIMO system, we assume equal power allocation, and power map design for MU massive MIMO systems is left for future work.

User selection is, in principle, optional in massive MIMO networks since by definition, there are many more antennas at the base-station than users to be served. Nevertheless, user selection can drastically improve performance and fairness and, depending on system assumptions, the optimal number of scheduled users can be much smaller than the number of users. In [69, 67], it was found that to maximize capacity, the number of users should be around 20% to 40% of the number of antennas However, this result was found under the assumption that the optimization goal is the sum spectral efficiency without any consideration of fairness or discrete MCS. Furthermore, in [70], it is shown that users that experience similar channels (which happens with non-negligible probability [71]) should not be simultaneously scheduled. The authors, however, consider max-min fairness and neglect the impact of discrete MCS.

Several studies have investigated the user selection for MU-MIMO systems [60]. Finding the optimal user selection is NP-hard[60], and it can only be done in reasonable time in extremely specific scenarios involving a small number of users and antennas. As a result, the standard approach in the literature is to design sub-optimal heuristic algorithms that strike a balance between complexity and performance. The use of general purpose optimization meta-heuristics such as genetic [72] and particle swarm [73] have been proposed to solve the user selection sub-problem. However, these algorithms have significant complexity and are thus impractical for MU massive MIMO systems, which have tight runtime constraints. For MU-MIMO systems, greedy opportunistic techniques such as in [74, 75, 76] have been developed assuming a Shannon rate function, where the base-station starts with an initial guess for the user-set and then iteratively updates the solution as long as the objective improves. Furthermore, it was shown in [77] that, compared to [74, 75, 76], starting with a full set of users and gradually removing one user at a time achieves a slightly better sum-rate while greatly reducing user selection complexity by using vectorized operations for precoder re-computation after removing users. In this work, we also consider different greedy searches. However, we focus on proportional fairness rather than sum-rate and we use a different stopping criteria. Unlike previous works that used incremental performance as a stopping criteria, i.e., the search is stopped when a change would result in performance degradation, we propose to keep searching until there are no users left to drop and taking the best out of all test user selections.

Once a user-set has been selected, the next sub-problem to be tackled is precoding. While the optimal choice of precoding is known to be non-linear dirty-paper-coding with successive interference cancellation at the receiver [78], this approach is too complex for practical deployment. Furthermore, the computation of optimal linear precoding vectors is generally an NP-hard problem [61] and can only be obtained using techniques such as branch and reduce [79]. Alternatively, sub-optimal linear precoding can be derived using techniques such as fractional programming [80, 81] or through the combination of line-search techniques with convex optimization [82]. In practice, linear precoding heuristics such as Maximum Ratio Transmission (MRT), Zero-forcing Transmission (ZFT) and Regularized Zero-forcing Transmission (R-ZFT) [83] are quick to compute. ZFT is particularly attractive since it nullifies inter-user interference and thus greatly simplifies the power distribution sub-problem. For these reason, in this work we focus exclusively on ZFT systems.

After determining the user selection and ZFT precoding, the base-station must solve the power distribution sub-problem which determines the power allocated to each selected user in a PRB based on the selected objective function. The selected objective is a function of the power allocated to each user that reflects the chosen trade-off between fairness and spectral efficiency. To solve the power distribution work, the defacto approach is to assume that the rate is an increasing concave function of the assigned power which allows the usage of the well known water-filling algorithm [64, 49]. It should be noted that this simple approach is only applicable with ZFT precoding and in general power distribution cannot be decoupled from precoding and would thus require more sophisticated optimization techniques such as geometric programming [84] or fractional programming [80]. Although several works have investigated power distribution, all assume a Shannon rate function which is a poor approximation to the piece-wise constant rate function obtained from practical MCS. In this work, we avoid this pitfall and design an MCS-aware solution that tackles the problem directly without approximation.

Finally, MCS selection refers to the process by which the base-station selects an MCS for each user in a PRB based on their SNR. To the best of our knowledge, this process is not taken into account in prior studies. Nevertheless, we will show that its effects cannot be neglected. In particular, as there are a limited number of MCSs to select from and each has a corresponding SNR threshold below which a user cannot correctly demodulate and decode, the actual rate-function is piece-wise constant. There is therefore an SNR threshold below which even the lowest rate MCS cannot be demodulated and decoded. Thus, below this SNR, the delivered rate to a user is zero, a point that is always overlooked when using the Shannon rate function. Thus, one can identify a set of users that would then receive a rate of zero in the PRB and that can therefore be dropped from the PRB without reducing performance. Indeed, one could deselect these users, and perform precoding and power distribution calculations anew. This is an optional iterative process that can improve overall performance, as will be shown in Section 3.7. Alternatively, one can skip any iterations and "waste" the power assigned to these users at the benefit of reduced computational complexity.

Although several works have already investigated the different steps considered during the operation of a MU massive MIMO network, these studies have not considered the discrete nature of the rate function. Moreover, most of the related works focus on single-carrier and do not consider the massive MIMO OFDMA case. The results of our investigation indicate that combining grouping with simple deselection strategies is an effective operation strategy that strikes a favourable balance in terms of both performance and runtime. Furthermore, our proposed approach achieves in real-time performance that is close to the best offline feasible solution to joint optimization problem.

3.4 System Description and Assumptions

Notation: matrices and vectors are set in upper and lower boldface, respectively. The operators $(.)^T$, $(.)^H$, $(.)^\dagger$ and $||.||$ denote, the transpose, the Hermitian transpose, the pseudo inverse and the Euclidean norm. Caligraphic letters, such as $\mathcal{A}$, denote sets and $|\mathcal{A}|$ denotes set cardinality of set $\mathcal{A}$. Finally, $\mathbb{C}$ denotes the complex numbers.

In this work, we assume that the base-station uses massive MU-MIMO which implies that there are more antennas (M) than users (U). Moreover, the used precoding is ZFT since it simplifies the design of network operation algorithms and because its relative gap to the optimal precoding diminishes as the number of antennas grows [12, 13]. In addition, we assume digital beamforming, i.e., one RF processing chain per antenna, is used since it allows the base-station to change precoding on a per PRB basis. Furthermore, we assume a full-buffered traffic model, i.e., the base-station always has data to send to every user. Finally, we assume perfect channel state estimation.

We consider the downlink of an OFDMA single-cell system with C subchannels, a total power budget per time-slot of P^{total} and M antennas at the base-station. In OFDMA, the smallest resource unit is called a PRB (Physical Resource Block) and is indexed by a frequency subchannel c and time-slot t pair (c, t). These radio resources are used to serve a set of U single-antenna users, $\mathcal{U}$. We assume that each PRB is allocated a power $P = P^{\mathsf{total}}/C$ (corresponding to equal power per PRB allocation). For a given PRB (c, t) the complex channel coefficient between the m-th antenna at the base-station and user u is $g_u^m(c, t)$. We denote the channel vector between the base-station and user u as $\mathbf{g}_u^{c,t} = \left[g_u^1(c,t), g_u^2(c,t), \ldots, g_u^M(c,t)\right]^T \in \mathbb{C}^{M\times 1}$ and the channel matrix as $\mathbf{G}(c,t) \in \mathbb{C}^{U\times M}$:

$$\mathbf{G}(c,t) = \left[\, \mathbf{g}_1^{c,t}, \mathbf{g}_2^{c,t}, \ldots, \mathbf{g}_U^{c,t}\right]^T. \tag{3.1}$$

3.4.1 Precoding

Given a set of selected users $\mathcal{X} = \{u_1, \ldots, u_{|X|}\} \subset \mathcal{U}$ for a given PRB, the base-station performs ZFT. The precoding vector for a selected user u and PRB is denoted as $\overline{\mathbf{w}}_u^{c,t} =$

$[\overline{w}_u^1(c,t), \overline{w}_u^2(c,t), \ldots, \overline{w}_u^M(c,t)] \in \mathbb{C}^{M \times 1}$. The precoding matrix $\overline{\mathbf{W}} \in \mathbb{C}^{M \times |\mathcal{X}|}$ is defined as:

$$\overline{\mathbf{W}}(\mathbf{c}, \mathbf{t}) = \left[\overline{\mathbf{w}}_{u_1}^{c,t}, \overline{\mathbf{w}}_{u_2}^{c,t}, \ldots, \overline{\mathbf{w}}_{u_{|\mathcal{X}|}}^{c,t}\right]. \tag{3.2}$$

Let $s_u^{c,t}$ be a transmitted symbol to user u on PRB (c,t). We assume that the transmitted symbols are normalized to unit average power, i.e., $\mathbb{E}[|s_u^{c,t}|^2] = 1$. We denote the vector of transmitted symbols as $\mathbf{s}^{c,t} = [s_{u_1}^{c,t}, s_{u_2}^{c,t}, \ldots, s_{u_{|X|}}^{c,t}]$. Therefore, the allocated power to user $u \in \mathcal{X}$ on a given PRB is: $p_u^{c,t} = ||\overline{\mathbf{w}}_u^{c,t}||^2$ and we have

$$\overline{\mathbf{w}}_u^{c,t} = \sqrt{p_u^{c,t}}\mathbf{w}_u^{c,t}, \tag{3.3}$$

where $\mathbf{w}_u^{c,t} = \frac{\overline{\mathbf{w}}_u^{c,t}}{||\overline{\mathbf{w}}_u^{c,t}||}$ is the normalized precoding vector of $\overline{\mathbf{w}}_u^{c,t}$.

With a linear precoder, the transmitted signal vector from all base-station antennas on a given PRB is defined as:

$$\overline{\mathbf{s}}^{c,t} = \sum_{u \in \mathcal{X}} \overline{\mathbf{w}}_u^{c,t} s_u^{c,t} = \overline{\mathbf{W}}\ \mathbf{s}^{c,t}. \tag{3.4}$$

The received symbol $y_u^{c,t}$ at user u is then:

$$\begin{aligned} y_u^{c,t} =& (\mathbf{g}_u^{c,t})^T \overline{\mathbf{s}}^{c,t} + n_u^{c,t} \\ =& (\mathbf{g}_u^{c,t})^T \overline{\mathbf{W}}(c,t)\ \mathbf{s}^{c,t} + n_u^{c,t} \\ =& (\mathbf{g}_u^{c,t})^T \overline{\mathbf{w}}_u^{c,t} s_u^{c,t} + \sum_{v \in \mathcal{X}, v \neq u} (\mathbf{g}_u^{c,t})^T \overline{\mathbf{w}}_v^{c,t} s_v^{c,t} + n_u^{c,t}, \end{aligned} \tag{3.5}$$

where $n_u \sim \mathcal{N}_{\mathcal{C}}(0, \sigma^2)$ is a complex circularly-symmetric Gaussian random variable that models additive white Gaussian noise (AWGN) at user u. Assuming the system uses ZFT, we have, for each two selected users u and v:

$$(\mathbf{g}_u^{c,t})^T \mathbf{w}_v^{c,t} = 0, \qquad u \neq v \wedge u, v \in \mathcal{X}, \forall c, t \tag{3.6}$$

and the SNR[3] $\gamma_u^{c,t}$ for user u at PRB (c,t) is then

$$\gamma_u^{c,t} = p_u^{c,t} ||(\mathbf{g}_u^{c,t})^T \mathbf{w}_u^{c,t}||^2 / \sigma^2. \tag{3.7}$$

[3] The fact that ZFT precoding cancels inter-user interference means that the rate is governed by the signal to noise ratio.

Table 3.1: Available rates and the corresponding SNR thresholds [1].

SNR Threshold (dB)	-6.5	-4	-2.6	-1	1	3	6.6	10	11.4	11.8	13	13.8	15.6	16.8	17.6
Efficiency (bits/symbol)	0.15	0.23	0.38	0.6	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.9	4.52	5.12	5.55

3.4.2 Rate function assumptions

In practice, a base station has access to L modulation and coding schemes and needs to select one of them for each user it wants to transmit to, based on the user's radio conditions. In that case, the rate function $f(.)$ that maps SNR γ_u of user u to the rate r_u, corresponding to the highest allowed Modulation and Coding Scheme (MCS) for a given target Block Error Rate (BLER), is a piece-wise constant increasing function described as:

$$f(x) = B_c \times (\ e_1 \mathbb{1}_{[\Gamma_1,\Gamma_2)}(x) + \cdots + e_L \mathbb{1}_{[\Gamma_L,\infty)}(x)\ ) \tag{3.8}$$

where $\mathbb{1}_A(x)$ denotes the indicator function with a value of 1 if $x \in A$ and zero otherwise, Γ_l is the SNR decoding threshold for MCS l, i.e., the minimum required SNR for using MCS level l, and e_l is the spectral efficiency of MCS l measured in bit/s/Hz. An example of the SNR thresholds and spectral efficiencies is provided in Table 3.1 [1]. From an optimization perspective, this piece-wise constant function is problematic since its derivative is 0 everywhere except at its points of discontinuity [49]. We also note that if the received SNR is less than Γ_1 then the received rate is zero. i.e., $f(x) = 0\ \forall x < \Gamma_1$.

Most of the literature assumes that the relationship between the SNR experienced by a user and the rate received follows the Shannon capacity formula, i.e., $f(x) = B_c \log_2(1+x)$. This approximation simplifies the problems by replacing the actual rate function with a smooth increasing concave function. Some works have proposed a different approximation by adding a penalty parameter [85] or by using different approximation such a monomial [7, 8]. However, as illustrated in Fig. 3.7, which compares the exact rate function and Shannon's capacity formula, Shannon's formula ignores both the minimum SNR required for getting a non-zero rate as well as it overestimates the achieved rate. With Shannon, a small positive increase in SNR yields a positive increase in rate and any strictly positive SNR yields a strictly positive rate which is not true for the exact rate function.

3.5 Optimal Resource Allocation: The Offline Study

In this section, our objective is to find a method for offline benchmarking of a massive MU-MIMO OFDMA downlink system. This method will be used as an offline performance

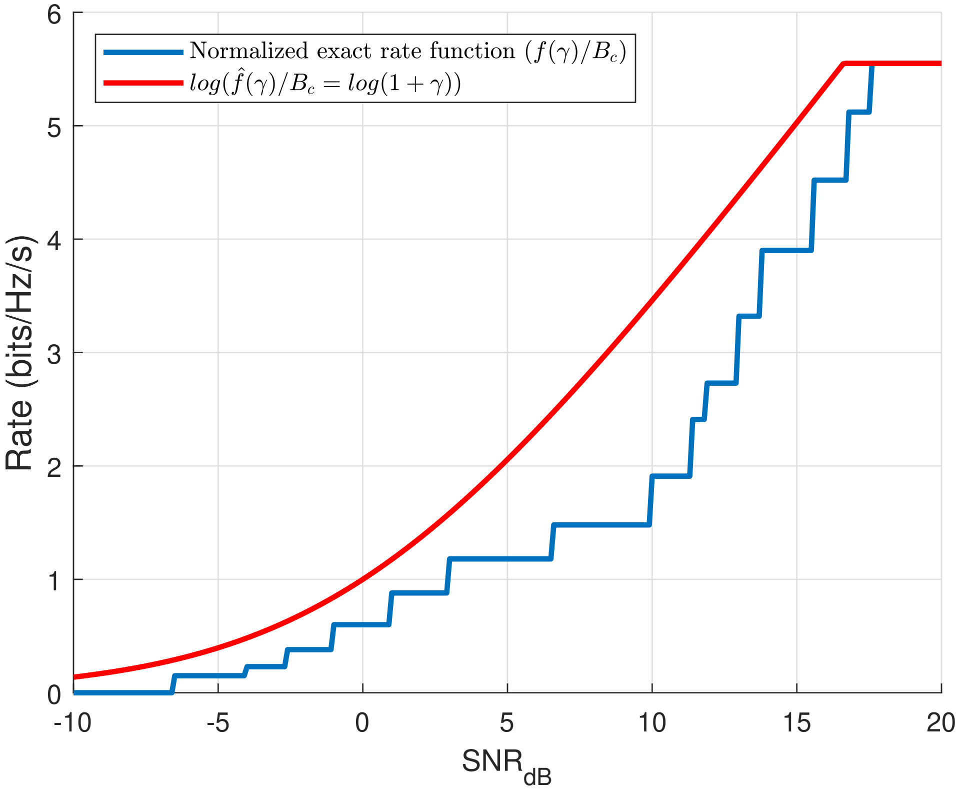

Figure 3.7: Comparing the actual rate function [1] with Shannon capacity formula in terms of the spectral efficiency (bits/Hz/s) at different SNR values (dB).

target for online systems that must compute decisions in milliseconds or less, as discussed in 3.6. In order to be able to do so effectively, we need an offline method that can compute quasi-optimal solutions in a reasonable time (e.g. a few minutes). More will be said on that later.

Ideally, the base-station would find the resource allocation by solving the optimization problem $\mathbf{P}_0(\omega)$ that strikes a trade-off between fairness and performance [6]. Problem $\mathbf{P}_0(\omega)$ below is the problem of jointly optimizing user selection, ZFT precoding, user power distribution and MCS selection over T time-slots (i.e., CT PRBs). Specifically, this problem is for a single-cell OFDMA massive MIMO system realization ω that is characterized by a given set of users $\mathcal{U}$ (such that $U \leq M$), the channel gains per PRB $\{\mathbf{g}_u^{c,t}\}$, per PRB power budget P, σ^2, the piece-wise constant rate function $f(.)$ and a large positive constant B.

$\mathbf{P}_0(\omega)$: Joint user selection, ZFT precoding and power distribution optimization for WSR maximization over a horizon of W PRBs of a single-cell multi-carrier DL system, given the rate function $f(.)$, the channel gains $\mathbf{g}_u^{c,t}$, the noise variance σ^2, power allocated per PRB P and a large positive constant B.

$$\max_{x_u^{c,t}, p_u^{c,t}, \mathbf{w}_u^{c,t}, r_u^{c,t}} \prod_{u=1}^{U} \lambda_u$$

$$\text{s.t. } \lambda_u = \sum_c \sum_t r_u^{c,t} \qquad \forall u \tag{3.9a}$$

$$r_u^{c,t} = f(\gamma_u^{c,t}) \qquad \forall u \tag{3.9b}$$

$$\gamma_u^{c,t} = p_u^{c,t} ||(\mathbf{g}_u^{c,t})^T \mathbf{w}_u^{c,t}||^2 / \sigma^2 \qquad \forall u, c, t \tag{3.9c}$$

$$\sum\nolimits_{u=1}^{U} p_u^{c,t} \leq P \qquad \forall c, t \tag{3.9d}$$

$$x_u^{c,t} \in \{0, 1\} \qquad \forall u, c, t \tag{3.9e}$$

$$p_u^{c,t} \leq x_u^{c,t} P \qquad \forall u, c, t \tag{3.9f}$$

$$||(\mathbf{g}_v^{c,t})^H \mathbf{w}_u^{c,t}||^2 \leq (2 - x_u^{c,t} - x_v^{c,t}) B \qquad \forall u, v \neq u, c, t \tag{3.9g}$$

$$||\mathbf{w}_u^{c,t}||^2 = 1 \qquad \forall u, c, t \tag{3.9h}$$

The binary indicator variable $x_u^{c,t}$ indicates whether user u is selected to be transmitted to or not in PRB (c, t), modelling the user selection decision. Precoding on a PRB is modelled using the complex vector $\mathbf{w}_u^{c,t}$ for user u. The variable $p_u^{c,t}$ denotes the power allocated to user u in the PRB, modelling the power distribution step. The MCS selection step is reflected in the variable $r_u^{c,t}$ which denotes the selected rate for a user u in terms of its SNR $\gamma_u^{c,t}$ through the piece-wise constant rate function $f(.)$. The SNR is determined directly from the allocated power and ZFT precoder vector as described in (3.9c). Constraint (3.9d) states that the total of the power distributed to all users in a PRB should be no more than the power P allocated to the PRB (c, t). Constraint (3.9f) ensures that the power allocated to the PRB in consideration, P, will only be distributed to selected users, i.e., users with a binary variable $x_u^{c,t}$ equal to 1. Furthermore, constraint (3.9g) ensures that the precoder zeros the intra-cell interference from a selected user to all other selected users.

The objective of the joint problem $\mathbf{P}_0(\omega)$ is to maximize the geometric mean of the total throughput received by each user λ_u over the horizon of CT PRBs. This objective is chosen since it is equivalent to achieving proportional fairness in the received rates [51, 6, 50]. Proportional fairness strikes a favourable balance between fairness and cell capacity maximization since it ensures that no one user can unilaterally improve their rate without degrading the throughput of another user by at least the same percentage.

Since the power allocation to each PRB is given, the optimization problem is only coupled because the rate a user gets in the time horizon is the aggregate over all the CT PRBs and the objective is to maximize the geometric mean of the users rates. The larger the T, the larger the fairness window. However, using a large window would imply knowing channel state information (CSI) for a large window, which is unreasonable. Hence, typically T is chosen to be the length of a frame with a proper mechanism to keep track of recent past history. However, for a practical solution, computation would have to be done on a PRB level to be fast. In that case $\mathbf{P}_0(\omega)$ has to be transformed. Specifically, the problem is approximated as a sequence of per PRB weighted sum-rate (WSR) maximization problems. This approximation is in line with the state of the art schedulers that are opportunistic, sequential, and aim to maximize proportional fairness at each step considering the past but without considering the future, i.e., they are myopic. Specifically, on PRB (c, t), let the rate received by user u in a window of $W + 1$ PRBs finishing at PRB (c, t) be,

$$\lambda_u^{c,t} = WR_u^{c,t} + r_u^{c,t}, \tag{3.10}$$

where $r_u^{c,t}$ is the rate that user u will receive in PRB (c, t), $W + 1$ is the selected fairness window and $R_u^{c,t}$ is the per PRB average rate received by user u in the past W PRBs which is assumed to be known. Then the optimum resource allocation decision for PRB (c, t) is the solution to an optimization problem with the following objective:

$$\begin{aligned}
\prod\nolimits_u (WR_u^{c,t} + r_u^{c,t})^{1/U} &\equiv \sum\nolimits_u \log\big(WR_u^{c,t} + r_u^{c,t}\big) \\
&\equiv \sum\nolimits_u [\log\big(WR_u^{c,t}\big) + \log\big(1 + r_u^{c,t}/WR_u^{c,t}\big)] \\
&\equiv \sum\nolimits_u [\log\big(1 + r_u^{c,t}/WR_u^{c,t}\big)] \\
&\simeq \frac{1}{W}\sum\nolimits_u r_u^{c,t}/R_u^{c,t},
\end{aligned}$$

where $\equiv$ indicates that the optimization problem is equivalent in the sense that the same resource allocation maximizes both sides of the equation. The last step comes from $\log(1 + x) \simeq x$ when x is small. Maximizing a WSR $\sum_u \theta_u^{c,t} r_u^{c,t}$ with the weights chosen as

$\mathbf{P}_1(\omega)$: Joint user selection, ZFT precoding and power distribution optimization for WSR maximization in one PRB of a single-cell multi-carrier DL system, given the weights θ_u, the rate function $f(.)$, the noise variance σ^2, the channel gains $\mathbf{g}_u$, the power allocated per PRB P and a large positive constant B.

$$
\begin{aligned}
\max_{x_u, p_u, \mathbf{w}_u, r_u} & \sum\nolimits_{u=1}^{U} \theta_u r_u \\
\text{s.t. } & r_u = f(\gamma_u) && \forall u && \text{(3.11a)} \\
& \gamma_u = p_u ||\mathbf{g}_u^T \mathbf{w}_u||^2 / \sigma^2 && \forall u && \text{(3.11b)} \\
& \sum\nolimits_{u=1}^{U} p_u \leq P && && \text{(3.11c)} \\
& x_u \in \{0, 1\} && \forall u && \text{(3.11d)} \\
& p_u \leq x_u P && \forall u && \text{(3.11e)} \\
& ||\mathbf{g}_v^H \mathbf{w}_u||^2 \leq (2 - x_u - x_v) B && \forall u, v \neq u && \text{(3.11f)} \\
& ||\mathbf{w}_u||^2 = 1 && \forall u && \text{(3.11g)}
\end{aligned}
$$

$\theta_u^{c,t} = 1/R_u^{c,t}$ asymptotically maximizes the long term utility function, i.e., the geometric mean of the rates or the proportional fairness in rates.

Hence $\mathbf{P}_0(\omega)$, can be decoupled into a sequence of weighted sum-rate maximization problems. The resulting per PRB problem, $\mathbf{P}_1(\omega)$, given below, is then a joint user selection, ZFT precoding, user power distribution and MCS selection problem on a given PRB. It should be noted that the explicit dependence on (c, t) for all variables is omitted for simplicity of notation.

Problem $\mathbf{P}_1(\omega)$ is a non-convex Mixed Integer Nonlinear Programming (MINLP) with a large number of variables since in massive MIMO the number of antennas is large and consequently, the number of users can also be quite large. Solving this problem exactly has an exponential worst-case runtime and thus, we present several method for obtaining good feasible solutions in reasonable time.

First, we note that given a user selection, i.e., given (x_u)'s, the ZFT precoder can be computed optimally in closed form as shown in [62]. Specifically, given a user selection characterized by the vector $\mathbf{x} = [x_1, x_2, \ldots, x_U]$, we define $\mathbf{G}_\mathbf{x}$ as a sub-matrix constructed from the channel matrix $\mathbf{G}$ by selecting the u^{th} row if $x_u = 1$ and dropping the row

$\mathbf{P}_2(\omega)$: Equivalent joint power distribution for a given user selection and the effective SISO channel η_u gains generated by ZFT precoding defined with a general rate function given $f(.)$, θ_u and P.

$$\begin{aligned} &\max_{p_u} \sum\nolimits_u \theta_u f(p_u \frac{\eta_u}{\sigma^2}) \\ &\text{s.t. } \sum\nolimits_u p_u \leq P, \end{aligned} \tag{3.14a}$$

otherwise. The ZFT precoder for a selected user u is then

$$\mathbf{w}_u = \frac{\left[\mathbf{G}_{\mathbf{x}}^H(\mathbf{G}_{\mathbf{x}}\mathbf{G}_{\mathbf{x}}^H)^{-1}\right]_{\pi(u)}}{\left|\left|\left[\mathbf{G}_{\mathbf{x}}^H(\mathbf{G}_{\mathbf{x}}\mathbf{G}_{\mathbf{x}}^H)^{-1}\right]_{\pi(u)}\right|\right|}, \tag{3.12}$$

where $[\mathbf{A}]_{\pi(u)}$ represents the $\pi(u)$-th column of matrix $\mathbf{A}$ and $\pi(u)$ denotes the row number of $\mathbf{g}_u^T$ in $\mathbf{G}_{\mathbf{x}}$. This is to say that the ZFT precoder for a user is computed based on the pseudo-inverse of the channel matrix of the selected users. It should be noted that the computational complexity of finding the ZFT precoders is $\mathcal{O}(UM^2)$.

Following the computation of the normalized ZFT precoders, the SNR for user u can be computed as:

$$\gamma_u = p_u\eta_u/\sigma^2, \tag{3.13}$$

where p_u is the power allocated to user u and $\eta_u = ||\mathbf{g}_u^T\mathbf{w}_u||^2$, the equivalent SISO channel gain experienced by user u after ZFT precoding. The dependency of η_u on the user-selection $\mathbf{x}$ is omitted for brevity.

For a given user selection, $\mathbf{P}_1(\omega)$ simplifies to a smaller per PRB power distribution and MCS selection problem $\mathbf{P}_2(\omega)$, with an objective obtained by combining (3.17a), (3.9b) and (3.13). Since the rate function is piece-wise constant, the power distribution problem, for a given user selection, can be described as the ILP (Integer Linear Program) shown in $\mathbf{P}_3(\omega)$ where a_u^l is a binary indicator variable showing whether user u is allocated MCS level l. Constraint (3.15a) ensures that the total power required never exceeds the available power on the PRB and constraint (3.15b) indicates that only one MCS is allocated to each user.

$\mathbf{P}_3(\omega)$ is a ILP, with at most LM variables, and although generally ILPs can be hard to solve, several commercial solvers such as CPLEX [86] and MATLAB's *intlinprog* [87]

$\mathbf{P}_3(\omega)$: ILP defining the joint power distribution and MCS selection for a given user selection and the effective SISO channel gains generated by ZFT precoding.

$$\max_{a_u^l \in \{0,1\}} \sum_{u \in \mathcal{X}} \sum_{l \in \{1,\ldots,L\}} \theta_u e^l a_u^l$$

$$\text{s.t.} \sum_{u \in \mathcal{X}} \sum_{l \in \{1,\ldots,L\}} a_u^l \frac{\Gamma^l}{\eta_u/\sigma^2} \leq P \tag{3.15a}$$

$$\sum_{l \in \{1,\ldots,L\}} a_u^l \leq 1 \qquad \forall u \in \mathcal{X} \tag{3.15b}$$

can be used to solve small ILPs to optimality in a reasonable time. Since our first goal in this study is to estimate the maximum achievable performance in an offline setting where runtime is not a limiting factor we will rely on using commercial ILP solvers for solving $\mathbf{P}_3(\omega)$. This formulation will also serve as an inspiration to our proposed greedy algorithm for power distribution presented in Section 3.6.

Therefore, given a user selection we can compute precoding optimally in closed form and then find the optimal power distribution and MCS selection using an ILP solver. Thus the per PRB joint resource allocation can be solved via a search problem over $\{x_u\}$. While this problem can be solved optimally by a brute-force exhaustive search, this approach does not scale. Therefore, we propose using a global search using BRB (Branch-Reduce-and-Bound) to get an upper-bound as well as a lower-bound, and local greedy search methods to compute good feasible solutions since BRB turned out to be too slow for our planning purposes for $U \geq 30$.

3.5.1 Branch-Reduce-and-Bound (BRB) based search

First, we notice that a user is selected if it is allocated a non-zero rate or equivalently an SNR greater that Γ_1, where Γ_1 is the required SNR for decoding the lowest MCS. This means that

$$x_u = \mathbb{1}_{r_u > 0}, \tag{3.16}$$

and x_u is non-decreasing in r_u. As a result, the joint resource allocation problem with discrete MCS can be described in terms of rates, as shown in Problem $\mathbf{P}_4(\omega)$, where the

$\mathbf{P}_4(\omega)$: Equivalent joint network operations written solely in terms of the assigned rates per user as a variable.

$$\max_{r_u \in \{0, e_1, \ldots, e_L\}} \sum_{u \in \mathcal{U}} \theta_u r_u \tag{3.17a}$$

$$\text{s.t.} \sum_{u \in \{u | r_u > 0\}} \frac{f^{-1}(r_u)}{\eta_u(\mathbf{x}(\mathbf{r}))/\sigma^2} \leq P \tag{3.17b}$$

constraint (3.17b) means that the summation of the required power for all users with non-zero rate should not exceed the total power allocated to the PRB. Eq. (3.17b) is the result of combining both (3.13) with (3.9b). The notation $\eta_u(\mathbf{x}(\mathbf{r}))$ emphasizes that the equivalent SISO channel gains are a function of the allocated rates $\mathbf{r}$ since the binary user selection vector $\mathbf{x}$ is determined by $\mathbf{r}$ following (3.16). Note that the selected rate r_u, for user u, is constrained to the discrete set of allowed rates defined by the used MCSs.

Moreover, constraint (3.17b) is an non-decreasing function of the rate. This is because $f^{-1}(r_u)$ is non-decreasing in r_u and $\eta_u(\mathbf{x}(\mathbf{r}))$ is non-increasing in r_u. The latter follows noting that $\mathbf{x}$ is a binary vector and:

$$\eta_u(\mathbf{x}) = \max_{||\mathbf{w}_u||=1} ||\mathbf{g}_u^T \mathbf{w}_u||^2 \tag{3.18a}$$

$$\text{s.t.} \ ||\mathbf{g}_v^T \mathbf{w}_u||^2 = 0 \qquad \forall \ v \text{ s.t. } x_v = 1, v \neq u \tag{3.18b}$$

and thus, increasing r_u to a positive value results in more constraints that cannot increase the maximum.

$\mathbf{P}_4(\omega)$ is therefore a discrete monotonic program where both the objective and the constrains are non-decreasing functions of the discrete rate variable. Despite the fact that discrete monotonic programs are non-convex, their structure allows for the use of BRB (Branch-Reduce-and-Bound) algorithm [88] which provides a sequence of upper bounds as well as lower bounds (feasible solutions) at each iteration until they converge to the optimal solution. In contrast to exhaustive search, BRB has a better than exponential average case complexity. As a result, even with a limited number of iterations, we can obtain more meaningful insights than a brute-force exhaustive search.

Ideally, BRB could be the only technique used for offline performance benchmarking since it provides a quasi-optimal[4] solution. However, as the problem size gets large, it becomes too slow and cannot find a good feasible solution in a reasonable time. However, we will show that greedy search methods are alternatives that can find feasible solutions faster and as good as BRB. Nevertheless, only by using BRB can we determine an upperbound.

3.5.2 Greedy search

In order to find a good feasible solution faster, we consider various greedy search solutions. The common feature of all the proposed methods is that we initially select a user-set and iteratively refine the set by either adding or deleting one user at a time. After several iterations, the best performing user-set, out of all tried user-sets, is returned.

The first approach is denoted as Greedy-up-all-the-way (GUAW). As shown in Fig. 3.8, the search starts at the root with an empty user-set $\mathcal{X} = \{\phi\}$ and progresses upwards in the tree of all possible user-sets until we reach its peak, i.e., the user-set containing all users. At each search iteration, we try adding one user to the current set and we compute the weighted sum-rate. After that, we retain the one with the highest weighted sum-rate and go to the next iteration. The algorithm is stopped when there are no more users to add and the best performing user-set is returned and thus it always tries $U(U-1)/2$ user-sets.

Alternatively, we can search the tree downwards using a Greedy-down-all-the-way (GDAW) algorithm. In Fig. 3.9, a sample path of one run of the GDAW is presented. Initially, the search is started with the user-set containing every user. Following this, one user is removed in each search iteration. The algorithim is stopped when there are no more users to drop. GDAW can potentially use various criteria for removing users. The proposed criteria are:

1. **GDAW:** In this case, we drop the weakest user based on its current experienced SNR, i.e., we drop the user $v = \operatorname{argmin}_{u \in \mathcal{X}} \gamma_u$, where γ_u is the experienced SNR of user u assuming optimal power distribution and MCS selection for the user-set $\mathcal{X}$. We try a total of U users sets with GDAW.

2. **GDAW + FD:** Initially, we perform the same computations done as with GDAW. Afterwards, for each user-set, we drop all users with zero rate (we call it full-drop),

[4] For runtime purposes, BRB is typically stopped at a reasonable tolerance, e.g. 10% gap between upper and lower bounds, despite the fact that with a discrete search space it will find eventually the exact optimum solution.

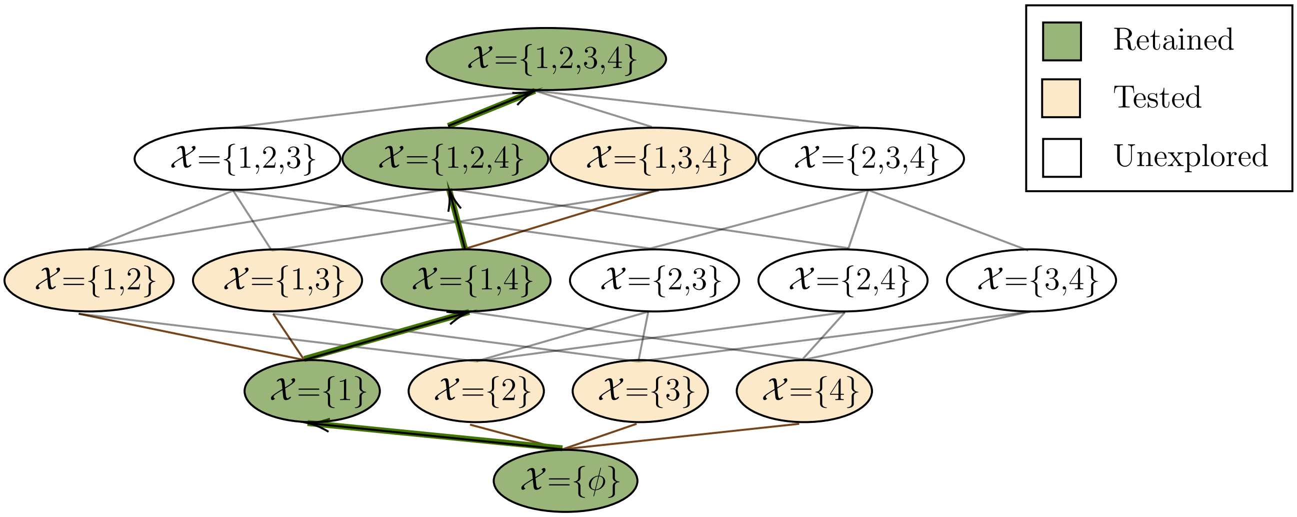

Figure 3.8: An example of a sample path performed by GUAW.

recompute precoding and optimal power distribution and MCS selection. Thus, instead of performing the computations for U user-sets, we compute for $2U$ user-sets at most.

3. **GDAW with look ahead (GDAW w LA):** In this cases, similar to GUAW, we compute the weighted sum-rate achieved by dropping each user in the current user-set. The user drop resulting in the highest weighted sum-rate is retained and is excluded from the following search iterations. We try $U(U-1)/2$ user-sets which is the same number of user-sets as with GUAW.

It should be emphasized that although the various greedy search techniques only scan a portion of the user-set tree, i.e., they do not explore all nodes and thus, they are generally providing sub-optimal solutions. However, as will be seen, some of the search methods end up performing similarly, suggesting that these simple search methods are quasi-optimal.

3.5.3 Comparing the greedy search to BRB

We compare the average weighted sum-rate value computed by both BRB and the proposed greedy search methods over 100 instances of $\mathbf{P}_1(\omega)$ as well as their runtimes. As a benchmark, we include the exact optimal solution for small problems (up to 15 users)

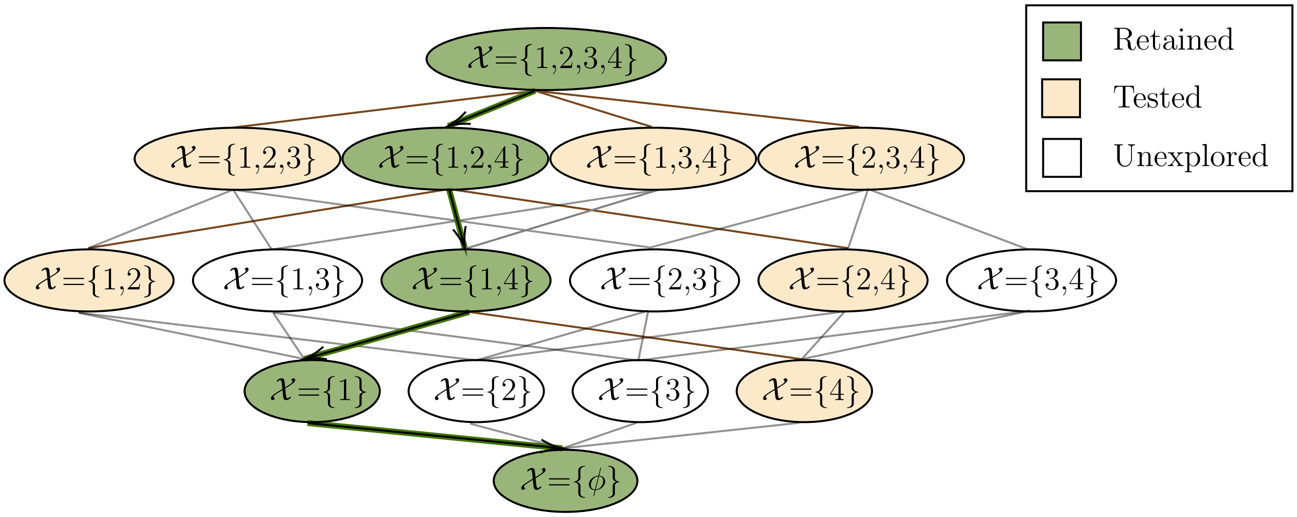

Figure 3.9: An example of a sample path performed by GDAW.

computed by exhaustive search on the user-sets. An instance of $\mathbf{P}_1(\omega)$ is characterized by a set of Independent and Identically Distributed (IID) uniformly generated random weights and a realization i.e., a channel matrix. The weights are normalized to one after generation. The channel matrix is composed of two parts, specifically,

$$\mathbf{g}_u = \sqrt{\beta_u}\tilde{\mathbf{g}}_u, \tag{3.19}$$

where β_u is a log-normal random variable representing the large-scale fading, i.e., path loss and shadowing, and $\tilde{\mathbf{g}}_u$ is a random vector representing the small-scale fading. $\tilde{\mathbf{g}}_u$ is generated following a Rician fading process. The generation method of β_u and $\tilde{\mathbf{g}}_u$ is the same as the one mentioned in Chapter 2 which is also described in [55, 82].

Figures 3.10 (resp. 3.11) compare the average weighted sum-rate of the different methods as a function of the number of users assuming the allocated power per PRB to be $10mW$ and $M = 100$ (resp. 64) antennas. First, we note that for a number of users less than 15, greedy-search methods are quasi-optimal since they achieve the same WSR as exhaustive search. Regarding BRB, we limit the computation time by limiting the number of iterations to $200,000$ with a goal of providing a gap between the upper-bound and the lower bound of at most 10%. Clearly, BRB can achieve these this gap for a number of users less that 30. However, as the number of users grows beyond 30, the BRB process is very slow at reducing the gap between its upper and lower bounds. This is visible since the upperbound is more than 10% away from the lower bound indicating that the algorithm

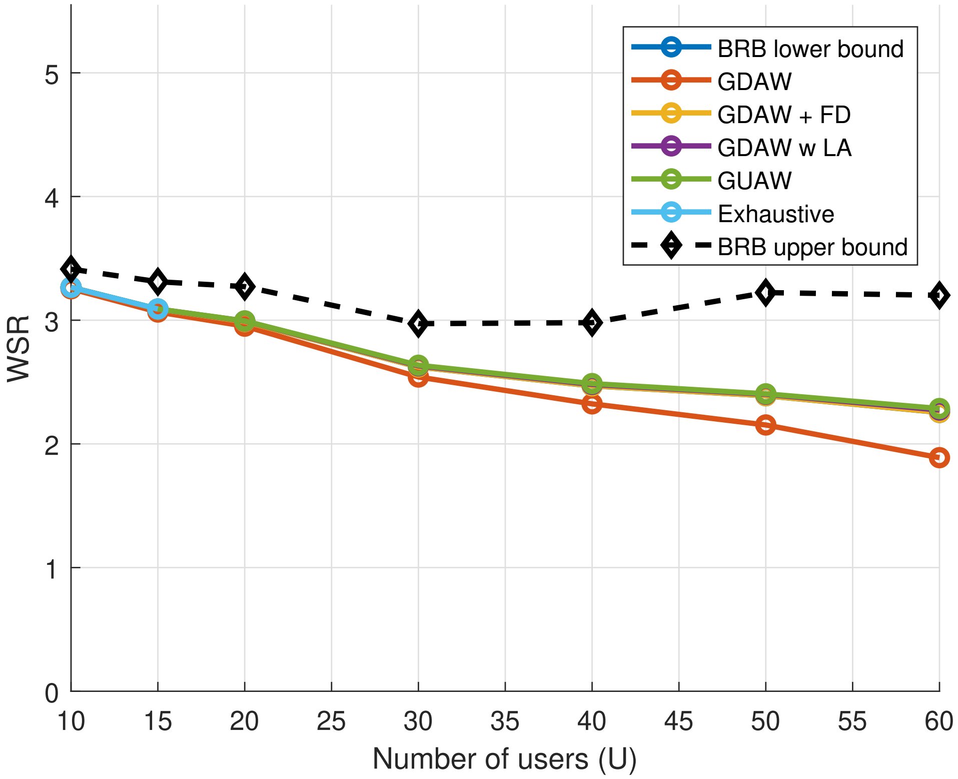

Figure 3.10: Weighted sum-rate achieved by exhaustive search, greedy search and bounds returned via BRB global search for $M = 100$ antennas and $P = 10\ mW$. The bounds are returned after $200,000$ iterations per instance.

stopped because it reached the maximum number of iterations which is equivalent to a runtime of 18.2 hours per instance for the case of $U = 40$. Interestingly, all greedy-search methods except GDAW as well as BRB yield similar results in terms of feasible solutions suggesting that the upper-bound is likely loose and that GDAW + FD is sufficient and they require only a few minutes per instances.

In summary, in this section, we have started with formulating a long-term horizon problem $\mathbf{P}_0(\omega)$ which maximizes proportional fairness by jointly optimizing user-selection, power distribution and MCS selection. We approximated $\mathbf{P}_0(\omega)$ by a sequence of weighted sum-rate problems. The per PRB problem $\mathbf{P}_1(\omega)$ is inline with the state of the art opportunistic schedulers. Then, we showed that $\mathbf{P}_1(\omega)$ for a given user selection can be

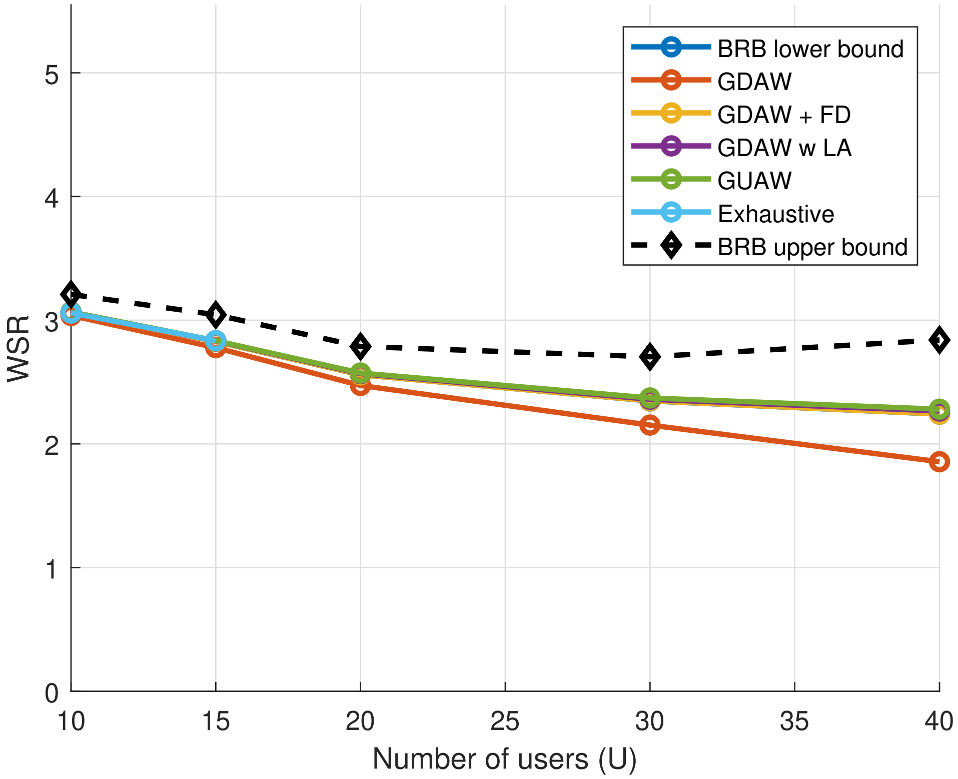

Figure 3.11: Weighted sum-rate achieved by exhaustive search, greedy search and bounds returned via BRB global search for $M = 64$ antennas and $P = 10\ mW$. The bounds are returned after $200,000$ iterations per instance.

further simplified into a smaller ILP $\mathbf{P}_3(\omega)$. Thus, solving $\mathbf{P}_1(\omega)$ optimally requires solving an ILP $\mathbf{P}_3(\omega)$ for every possible user-set which is unrealistic except for small number of users ($U \leq 15$). In order to find solution for larger systems ($U > 15$), we considered BRB (Branch-Reduce-and-Bound) which is used to find a upper-bound as well as an lower bound and thus if the gap is small (e.g., within 10%), it shows that the solution is quasi-optimal. This is useful for cases with $U \leq 30$. However, since the method takes a huge number of iterations to find a feasible solution, i.e., a lower bound, we also use greedy search methods to find good feasible solutions faster. Except for GDAW, all of the approaches for finding a feasible solution converge to similar performance; thus, for the following, we would rely on GDAW + FD for providing an offline target performance because it achieves the same performance as other methods with much lower computational cost.

3.6 Proposed Solution for Realtime Smart Network Operation

So far, we have seen that using a GDAW + FD for iterative user-selection with optimal power distribution obtained by exactly solving an ILP yields good feasible solutions that perform similarly to much more involved strategies and is shown to be quasi-optimal for up to 30 users. However, this method has a reasonable but still high runtime making it impractical for real-time operation. Thus, we use it to provide an offline performance target for practical schemes.

For real-time network operation, since ZFT precoding can be computed optimally in closed form for a given user selection, we need to develop efficient methods for performing user selection, power distribution and MCS selection. For a given user selection, we first propose a simple heuristic for jointly performing power distribution and MCS selection as an alternative to solving the ILP described by $\mathbf{P}_4(\omega)$. Second, we present a heuristic for user selection. But first we present the benchmark method inspired from the literature for network operation that we will use to assess the performance of the proposed method.

3.6.1 The Benchmark

For massive MIMO systems, we typically have a number of antennas much larger than the number of active users. Therefore most papers and the benchmark assume that all users are selected. Then precoding is performed according to (3.12). Finally, power distribution is computed using the waterfilling algorithm in [63, 64, 49]. Water filling is the solution to the power distribution, $\mathbf{P}_2(\omega)$, under the assumption that the relationship between the rate and SNR is determined via Shannon's capacity formula, i.e., $f(\gamma) = \log_2(1+\gamma)$. Finally, MCS selection is then performed by mapping the resulting SNR to the highest allowable MCS level. Note that this last step is typically ignored in most related works.

As a result, there is no attempt in the benchmark to address the issue that users may see a zero-rate when their SNR is less than Γ_1. By removing users with zero rate, better effective channels are obtained for the other selected users and power can be used more effectively. This approach ling power distribution from the MCS selection will be shown to be sub-optimal and a significant gain can be achieved at no extra complexity by switching to the proposed greedy based method described next. This result is in line with previous works [7, 89, 56, 68, 82] that show that a significant gain can be obtained by designing algorithms that are aware of the nature of the underlying physical layer limitations and the discrete nature of the rate function.

3.6.2 A Greedy Algorithm for Joint Power Distribution and MCS Selection

For power distribution and MCS selection, we replace the ILP described by $\mathbf{P}_3(\omega)$ with our proposed greedy algorithm that takes the two specific characteristics of the rate function as an inherent part of the problem instead of ignoring them. Hence, our proposed power distribution algorithm is MCS-aware. The algorithm greedily selects the user with the largest marginal improvement at each step, defined as the ratio of the gain in weighted sum-rate obtained by increasing the user's rate to the next MCS level to the required additional power for decoding this level.

The proposed heuristic algorithm is outlined in Algorithm 2. The notation $l(u)$ is used to denote the current MCS level selected for user u and it is computed as

$$l(u) = \sum_{l=0}^{L} a_u^l l, \tag{3.20}$$

where a_u^l is the binary indicator variable with a value of 1 if user u is allocated MCS level l and zero otherwise.

In each iteration, the algorithm selects the user with the largest marginal improvement $\frac{\theta_u(e^{l(u)+1}-e^{l(u)})}{(\Gamma^{l(u)+1}-\Gamma^{l(u)}))/\eta_u}$ to assign it a higher MCS than its currently allocated level $l(u)$. If the remaining power does not permit allocating a higher MCS, the user is removed from the set of users under-consideration. In addition, a user is removed from the set of users under-consideration if it is allocated the highest possible MCS. The algorithm keeps iterating until no power remains or no user remains in the set of users under-consideration.

The computation of the marginal improvement ratios and choosing the highest among them are the key computational challenges of the algorithm. The algorithm takes a maximum of UL iterations, which occurs when every user gets the maximum MCS. Since for each iteration, we need to find a maximum out of U values, the worst case complexity of the proposed algorithm is $\mathcal{O}(U^2)$ since L is a constant factor. This approach is comparable to that required by waterfilling [64] based approaches for solving $\mathbf{P}_3(\omega)$ which solves the power distribution problem.

To illustrate the advantage of our proposed MCS-aware greedy power distribution solution over waterfilling we plot the weighted sum-rate per PRB when all users are selected and the power distribution is computed using Shannon-based waterfilling, ILP, and our suggested greedy algorithm against the number of users in Fig. 3.12. The results are computed as the average weighted sum-rate achieved over $10,000$ instances of $\mathbf{P}_1(\omega)$, each

Algorithm 2: A greedy algorithm for joint power distribution and MCS selection given a user-set $\mathcal{X}$

```
**Given:** $\theta_u, e^l, \Gamma^l, \eta_u, P, \mathcal{X}$;
**Initialization:** $l(u) = p(u) = 0 \qquad \forall u \in \mathcal{X}$; $P^{rem} = P$; $\mathcal{A} = \mathcal{X}$;
**while** $P^{rem} \geq 0$ *and* $|\mathcal{A}| > 0$ **do**
    $v = \underset{u\in\mathcal{X}}{\text{argmax}} \frac{\theta_u(e^{l(u)+1}-e^{l(u)}\Gamma^{l(u)})}{(\Gamma^{l(u)+1}-\Gamma^{l(u)})/\eta_u}$;
    **if** $P^{rem} - (\Gamma^{l(v)+1} - \Gamma^{l(v)})/\eta_v \leq 0$ **then**
        $P^{rem} = P^{rem} - (\Gamma^{l(v)+1} - \Gamma^{l(v)})/\eta_v$;
        $l(v) = l(v) + 1$;
        **if** $l(v) = L$ **then**
            $\mathcal{A} = \mathcal{A} - \{v\}$;
        **end**
    **else**
        $\mathcal{A} = \mathcal{A} - \{v\}$;
    **end**
**end**
$p(u) = \Gamma^{l(u)}/\eta_u \qquad \forall u \in \mathcal{X}$;
return $l(u), p(u)$ ;
```

characterized by a set of iid uniformly generated random weights and a channel matrix. Although, selecting all users is sub-optimal, this is done to fairly compare different power distribution methods. The ILP yields a solution that is 22.50% better than waterfilling. The gap between our proposed greedy algorithm and ILP is negligible making our algorithm quasi-optimal. As a result, we consider this algorithm to be an important contribution because the performance results are strikingly better for our MCS-aware greedy algorithm and the runtimes of both the waterfilling and greedy algorithms are comparable, as will be discussed later.

3.6.3 User Selection

For the user selection problem, we propose a simple heuristic denoted as grouping with full-drop. Our user selection is decoupled into two stages 1) user grouping, and 2) user deselection via full-drop. In the first stage, a group of K_{RR} users are selected at random. The group size K_{RR} is determined empirically during the network planning phase and it is found to be much less than the number of the antennas.

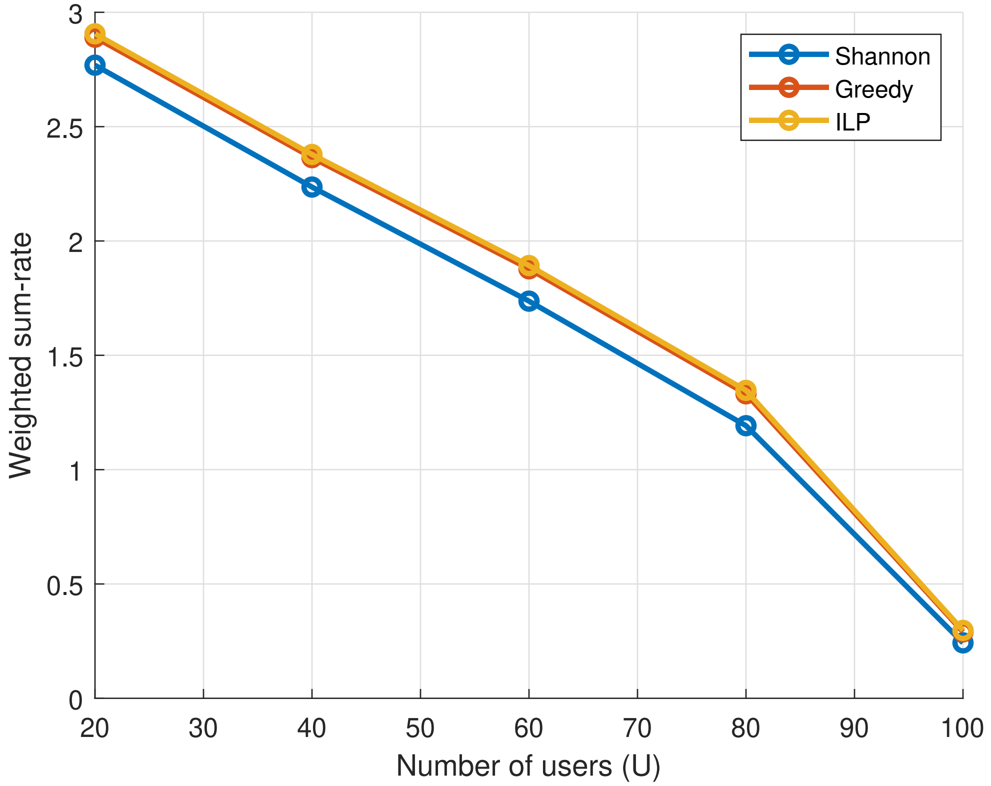

Figure 3.12: A comparison of different power-distribution methods in terms of the average weighted sum-rate per PRB assuming all users are selected for $M = 100$ and $R = 200$ m $(10,000$ instances per point).

It should be emphasised that we merely limit the amount of users who may be chosen, and select a subset of the users allocated to the group at random. This design decision lessens the complexity of user selection, and numerical results in Section 3.7 shows that it performs just as well as the best feasible solution computed offline. This indicates that for massive MIMO, reducing the number of selected users is critical for optimizing performance, and which users are selected into the group does not matter much as long as each has an equal chance of being selected.

After user grouping, the base-station computes precoding according to (3.12) and the power distribution based on Algorithm 2. After this preliminary stage, we examine the MCS levels assigned to each user, and if any user is given a zero rate, it is deselected. In

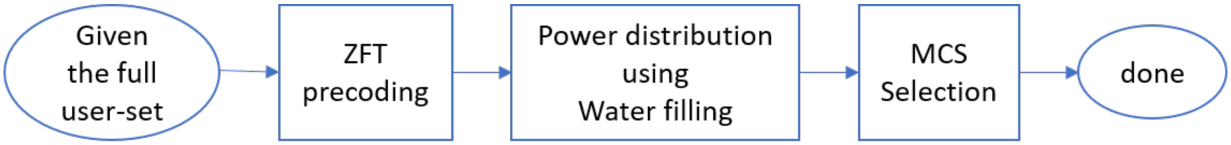

Figure 3.13: The sequence of RRM processes performed by the benchmark solution.

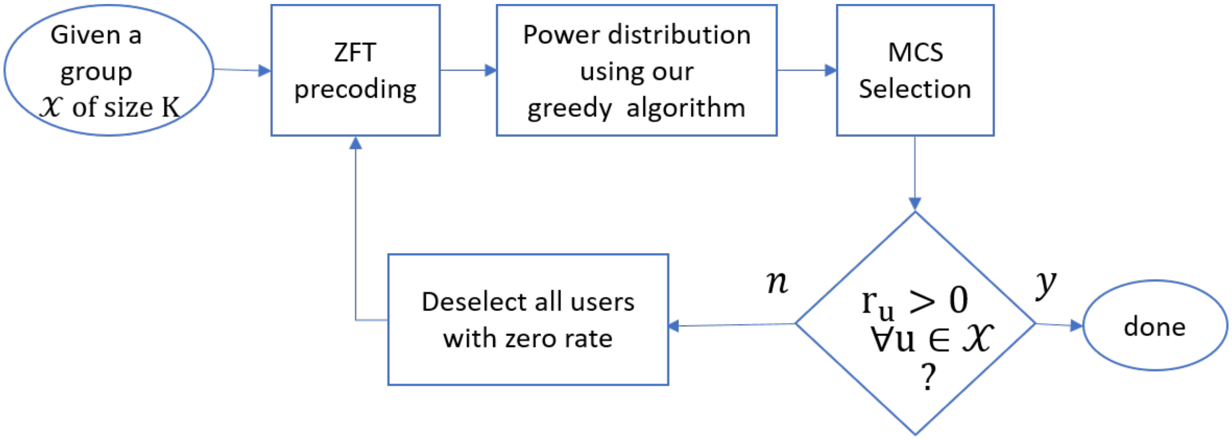
Given a group $\mathcal{X}$ of size K
ZFT precoding
Power distribution using our greedy algorithm
MCS Selection
$r_u > 0$ $\forall u \in \mathcal{X}$?
n
y
done
Deselect all users with zero rate

Figure 3.14: The sequence of RRM processes performed by the proposed solution.

other words, we fully drop all users receiving a zero rate. Then, the base-station recomputes precoding and power distribution resulting in potentially better rates for the remaining users since the equivalent SISO channel gains increase with the decrease in the number of selected users.

The overall complexity of operating a ZFT network using the proposed method is proportional to the number of selected users in the group, i.e., K_{RR}. Specifically, with the proposed network operation solution the worst case computational complexity is of the order $\mathcal{O}(M^2K_{RR} + K_{RR}^2)$. Since by definition, in massive MIMO systems the number of antennas is larger than the number of users, then the overall computational grows at a rate of the order $\mathcal{O}(M^2K_{RR})$.

As a summary in this section, we compare our proposed RRM solution with the benchmark RRM solution. As illustrated in Fig. 3.13, the benchmark starts by selecting all users, then performs power distribution using water-filling, then MCS selection is performed. Alternatively, our proposed method, illustrated in Fig. 3.14, starts by selecting a group of K_{RR} users randomly. Precoding is performed on the group followed by power distribution which is computed using the proposed greedy algorithm. Finally, if all selected users are assigned a non-zero rate, then the result is returned else we recompute precoding and power distribution for the set of users assigned a non-zero rate, i.e., we full-drop all users with zero-rate.

Next, we study the process of choosing the group size K_{RR} and show the importance of the 3 main parts of our proposed solution; namely, user-grouping, full-drop and the proposed greedy power distribution algorithm.

3.7 Simulation Results

3.7.1 Simulation Setup

The main goal is to compare the different practical solutions in terms of the geometric mean of the users' throughput (geometric mean throughput) over a large horizon and determine K_{RR}. The geometric mean throughput is selected as the performance metric since it is the objective function that is maximized by proportional fairness and it measures both fairness and performance simultaneously [50]. We also present runtime results. All computations are done using a MATLAB implementation executed on a Windows machine equipped with an Intel(R) Xeon(R) E5-2660 v3 with a total of 40 cores clocked at 2.6 GHz and 64 Gigabytes of RAM.

We assess the performance of the proposed solutions in a single cell network, assuming equal power allocation per sub-channel. We measure the average performance over 100 realizations where a realization is characterized by a channel matrix, per PRB. We assume stationary full-buffer users scattered randomly in a given area of radius R. The channel gains for each user are generated in accordance with the method given in [55], i.e., by multiplying the square root of the log-normal random path-loss component by a circularly symmetric random vector representing the small scale fading per antenna. Each realization corresponds to 1000 PRBs. The weights per user are updated on a per PRB basis as:

$$\frac{1}{\theta_u} := (1 - \frac{1}{W})\frac{1}{\theta_u} + \frac{1}{W}r_u, \tag{3.21}$$

where W is the fairness window. This approximates the exact mean rate R_u over a window W, derived in the Appendix, with an exponential moving average to remove the need for storing W rate values.

We use the rate function, i.e., the SNR to MCS mapping given in [1]. The other simulation parameters are summarized in Table 3.2 and used as the default setting unless stated otherwise.

Table 3.2: Single cell neutral host model parameters.

Simulation Parameters	**Symbols**	**Values**
Carrier frequency	f_c	2.5 GHz
Total available power	P^{total}	1 W
Number of antennas	M	64 or 100
Number of channels	C	100
Cell radius	R	200 m
Time-slot duration	τ	1 ms
PRB bandwidth	B_c	168 KHz
Fairness window	W	100 PRBs

3.7.2 Choosing the group size K_{RR}

First, we explore the performance of the proposed user grouping, i.e., picking, on a per PRB basis, a group of K_{RR} users randomly. We try grouping with and without full-drop (FD). Also we try waterfilling vs the proposed MCS-aware greedy power distribution. In Fig. 3.15, the geometric mean throughput performances of the following schemes are compared as a function of K_{RR} for $U = 60$ users:

1. **Shannon + ND:** In this scheme, the power distribution is computed assuming the rate function to be modelled using Shannon's capacity formula, i.e., power distribution is done using the waterfilling algorithm. Additionally, even if a user in the selected set is assigned a zero rate following MCS selection, they are not dropped (ND).

2. **Shannon + FD:** As opposed to Shannon + ND, all users assigned a zero rate are dropped at once and precoding, power distribution and MCS selection steps are recomputed.

3. **Greedy + ND:** Power distribution is computed using the proposed MCS-aware greedy algorithm. If some users are allocated a zero rate no re-computation is performed.

4. **Greedy + FD:** Power distribution is computed using the proposed MCS-aware greedy algorithm, however, users with zero rate are fully dropped.

In addition to the above schemes, we add dashed lines corresponding to the benchmark solution as well as the target performance set by the offline study and computed using GDAW + FD and ILP power distribution. First, it is observed that the benchmark is

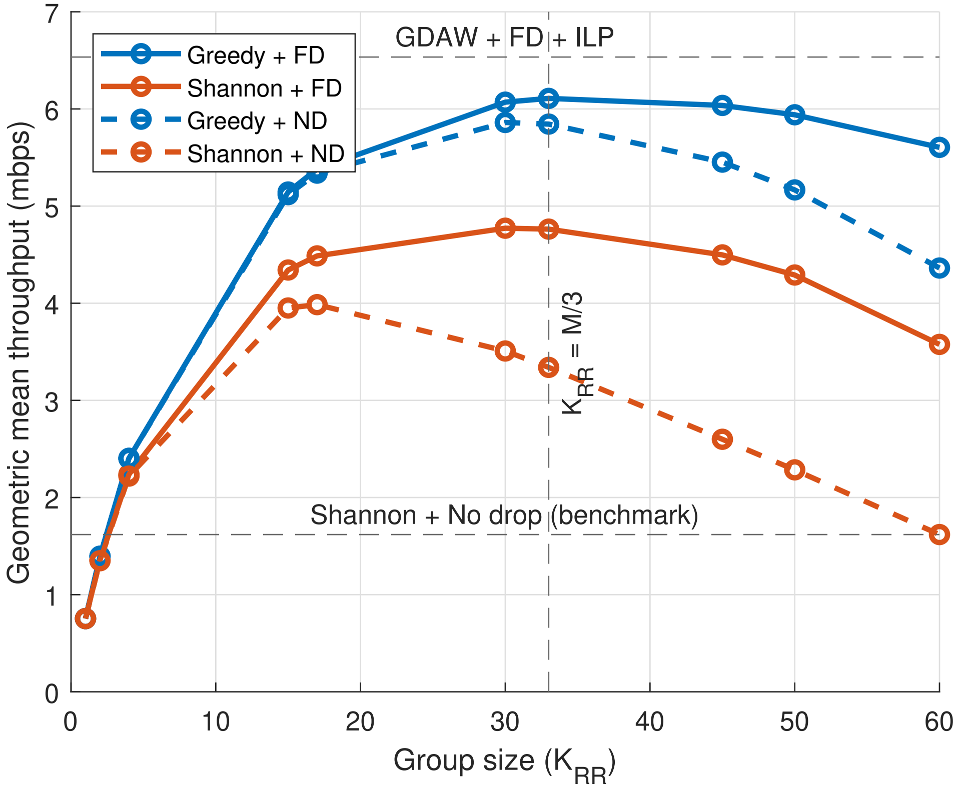

Figure 3.15: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 100$ antennas, $U = 60$ users and cell radius $R = 200$ meters.

far away from the target performance set by the offline study and thus, there is a need for a real-time solution that can perform similar to the target while having comparable runtime to the benchmark. Second, focusing on Shannon + ND, we can see that by just performing grouping a gain of 128% could be achieved by limiting the number of selected users. Furthermore, by comparing Shannon + FD and Shannon + ND, we can observe that an additional 21% gain can be obtained, however, a gap of 37.5% to the offline target performance still exists. Finally, by replacing water-filling with our MCS-aware greedy algorithm for power distribution, i.e., Greedy + FD with the best K_{RR} choice, we can achieve 96% of the performance achieved by the offline GDAW + FD + ILP. This solution brings a gain of 238% with respect to the benchmark for $U = 60$.

The choice of the best group size K_{RR} is found to be a function of static system

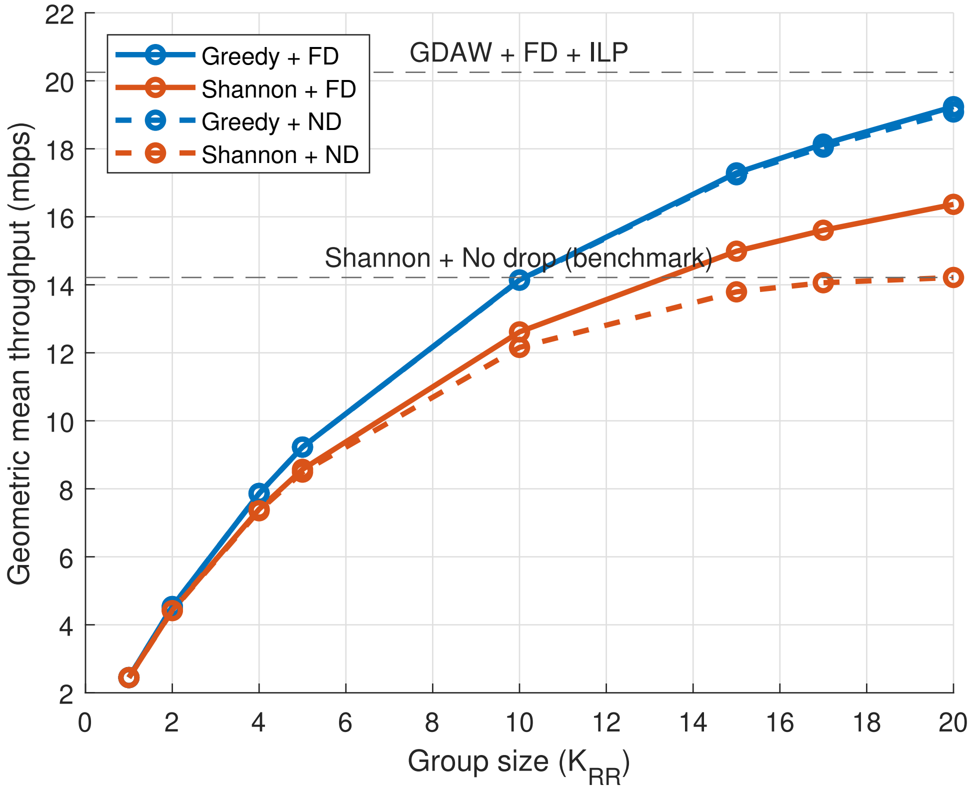

Figure 3.16: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 100$ antennas, $U = 20$ users and cell radius $R = 200$ meters.

parameters such as the number of antennas M and the cell radius R and independent of the number of users and thus it can be determined offline during the network planning phase. This is expected since with massive MIMO the effect of small scale fading vanishes due to channel hardening. This is supported by the an empirical study where a wide range of values for K_{RR} were tried and it was found that $\min(U, M/3)$ was the best setting. The results of this study on the group size K_{RR} for $M = 100$ and $U = 20, 40, 60$ and 80 are shown in Figs. 3.16, 3.17, 3.15 and 3.18 respectively.

In all the previous examples, $K_{RR} = \min(U, M/3)$ was found to be the best performing group size when M was set to 100 and the cell radius R was set to 200m. In Figures 3.19 and 3.20, we study the impact of changing the number of antennas from $M = 100$ to $M = 64$ and $M = 144$, respectively. It is still observed that the choice of $K_{RR} = \min(U, M/3)$

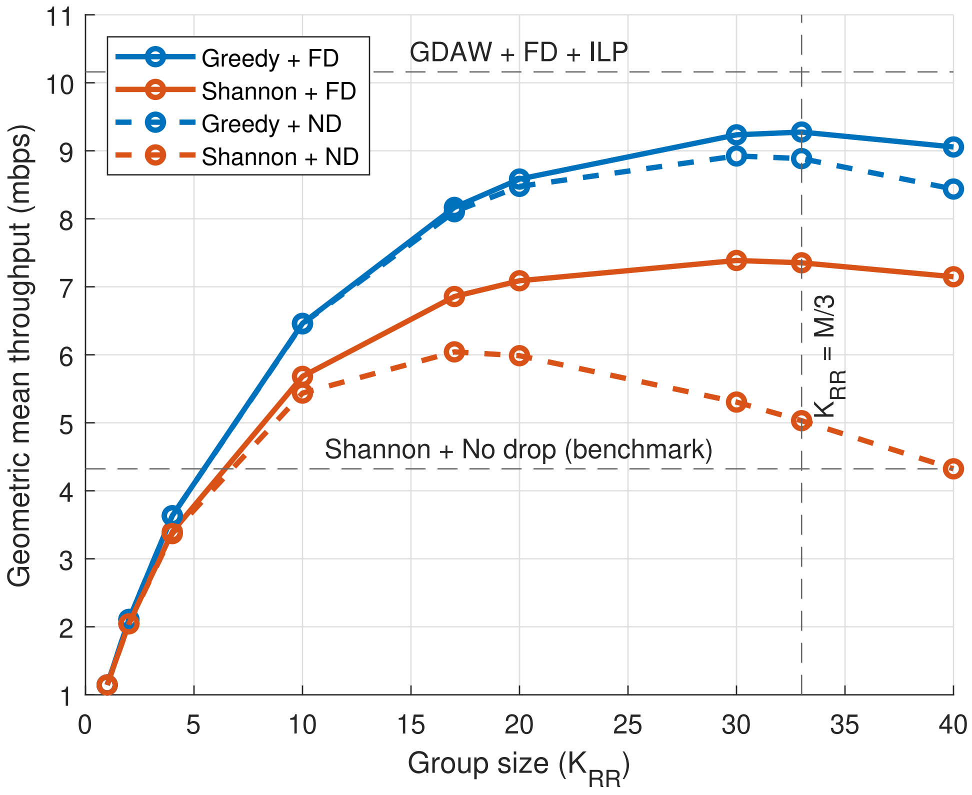

Figure 3.17: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 100$ antennas, $U = 40$ users and cell radius $R = 200$ meters.

performs best and remains the best selection. However, if instead of changing the number of antennas, the planned cell radius R was to change as shown in Fig. 3.21 where $R = 300$ meters, then the choice of $K_{RR} = \min(U, M/3)$ is no longer the best and the best choice becomes $K_{RR} = \min(U, M/6)$. In the following, we will set K_{RR} accordingly based on the value of R.

3.7.3 Performance and Runtime Results

In this section, our proposed solution is compared to both the benchmark and the offline target performance obtained by using GDAW + FD. In Fig. 3.22, the geometric mean throughput achieved by different schemes is plotted against the number of users in the cell

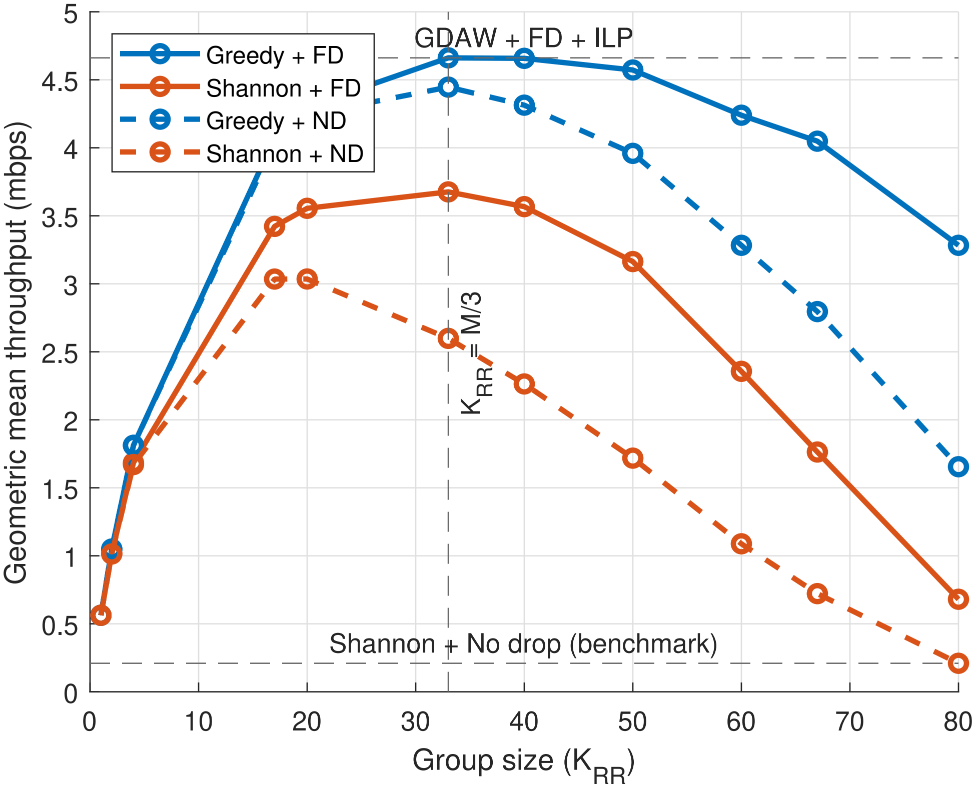

Figure 3.18: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 100$ antennas, $U = 80$ users and cell radius $R = 200$ meters.

assuming $M = 100$ antennas and $R = 200$ meters. The different schemes compared are:

1. **Bench:** The benchmark scheme that starts by selecting all users and computes the optimal ZFT from (3.12), then determines the power distribution by using waterfilling. No iteration are performed if some users see a zero rate.

2. **Bench + FD:** This is the benchmark approach to which an additional full-drop iteration is added if there are zero-rate users.

3. **Bench + Grouping + FD:** This is a scheme that uses grouping, waterfilling for power distribution and full-drop

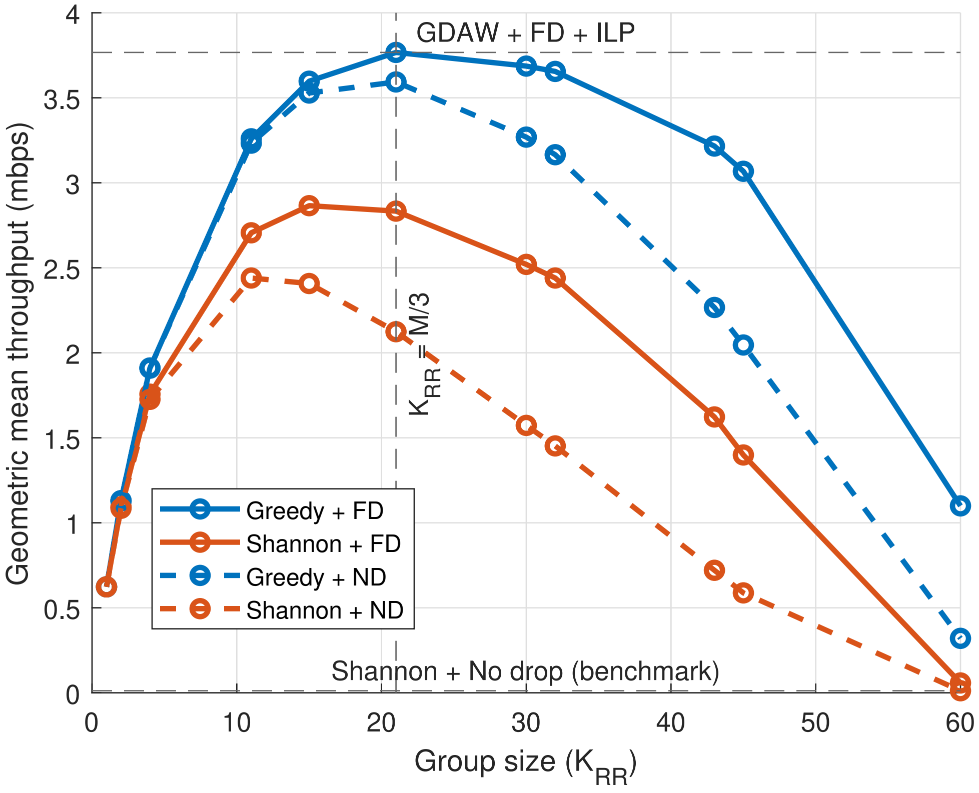

Figure 3.19: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 64$ antennas, $U = 60$ users and cell radius $R = 200$ meters.

4. **Proposed:** The proposed solution comprised of grouping, full-drop and the proposed greedy algorithm for power distribution.

5. **GDAW + FD:** The offline performance target, i.e., greedy search for finding a good user-set coupled with the optimal ILP-based power distribution and MCS selection computed using a commercial ILP solver.

From Fig. 3.22, Bench + FD sees a 64.5% improvement with respect to the benchmark for $U = 40$. This is due to the additional an extra full-drop iteration. However, this gain diminishes as the number of users increase. By doing grouping as well, i.e., with Bench + Grouping + FD, a consistent gain can be maintained as the number of users increases.

Focusing now on Fig. 3.23 that shows the percentage of the offline target performance

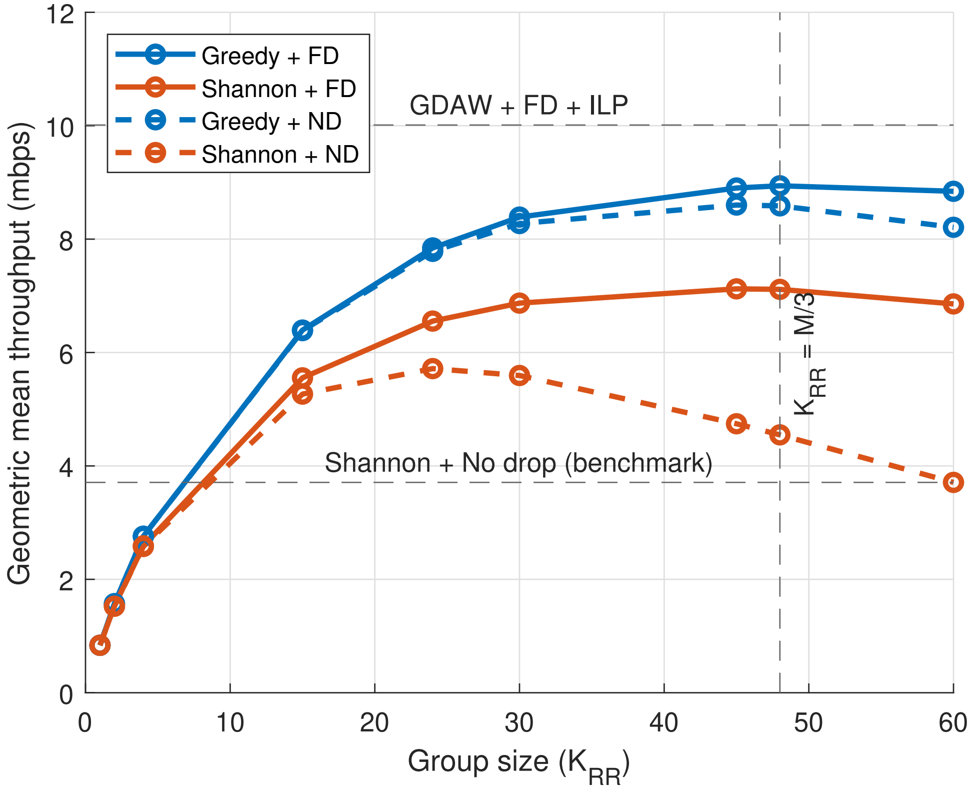

Figure 3.20: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 144$ antennas, $U = 60$ users and cell radius $R = 200$ meters.

achieved by the different online methods as a function of the number of users U. Bench + Grouping + FD achieves around 20% less compared to the offline performance target. Finally, the performance of our proposed solution is about 5% less than that obtained by the GDAW + FD.

In Figs. 3.24 and 3.25, we compare the proposed solution against the number of antennas M and the cell radius R, respectively for $U = 60$. It is clear that similar conclusions hold across the range of system parameters tested as the proposed solution continues to achieve similar performance to that obtained by the offline target performance.

In Fig. 3.26, on the runtime of the different schemes is presented for the case with $M = 100$ antennas and $R = 200$ meters. First, we notice that the runtime of the benchmark grows with the number of users since it always selects all users. Second, doing a second full-

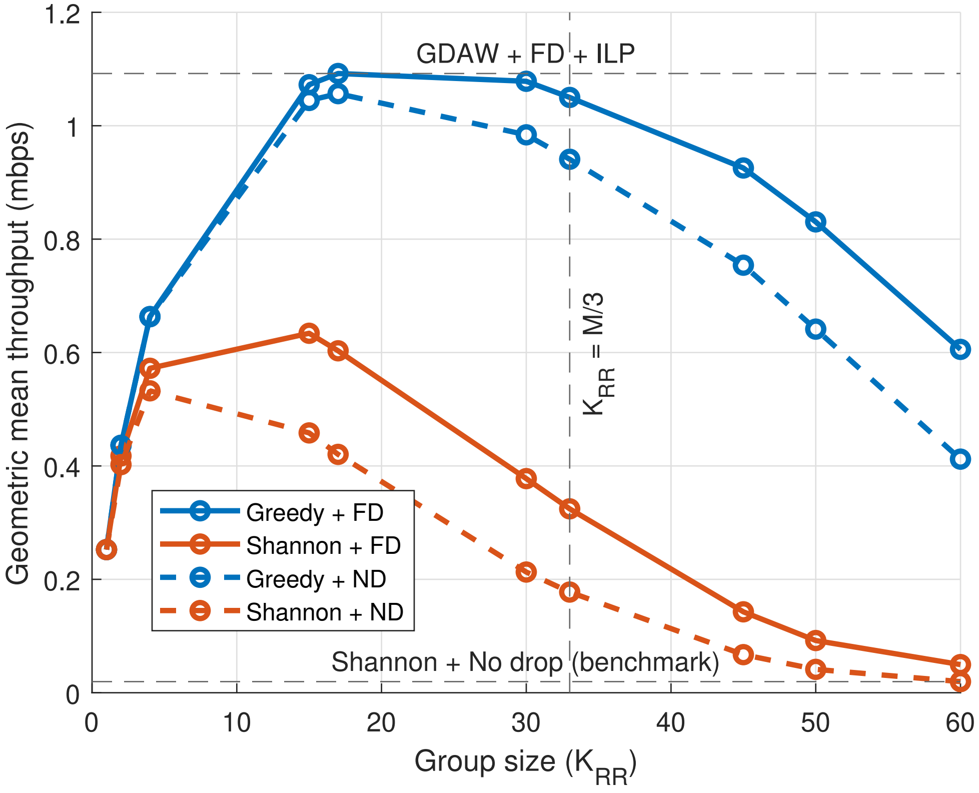

Figure 3.21: Comparison of the geometric mean throughput vs. the group size K_{RR} assuming $M = 100$ antennas, $U = 60$ users and cell radius $R = 300$ meters.

drop iteration increases runtime by 35.9% on average. Third, grouping caps the runtime since the number of users is limited to a maximum of K_{RR}. Finally, we see that our proposed MCS-aware greedy power distribution algorithim is comparable to water-filling power distribution and it is generally better in runtime than the benchmark when the number of users $U > 40$.

3.8 Conclusions

Operating a massive MIMO OFDMA network requires far more effort than a traditional SISO one. In this chapter, we have investigated three of the steps involved in practical resource allocation in massive MIMO networks, i.e., the user selection, power distribution

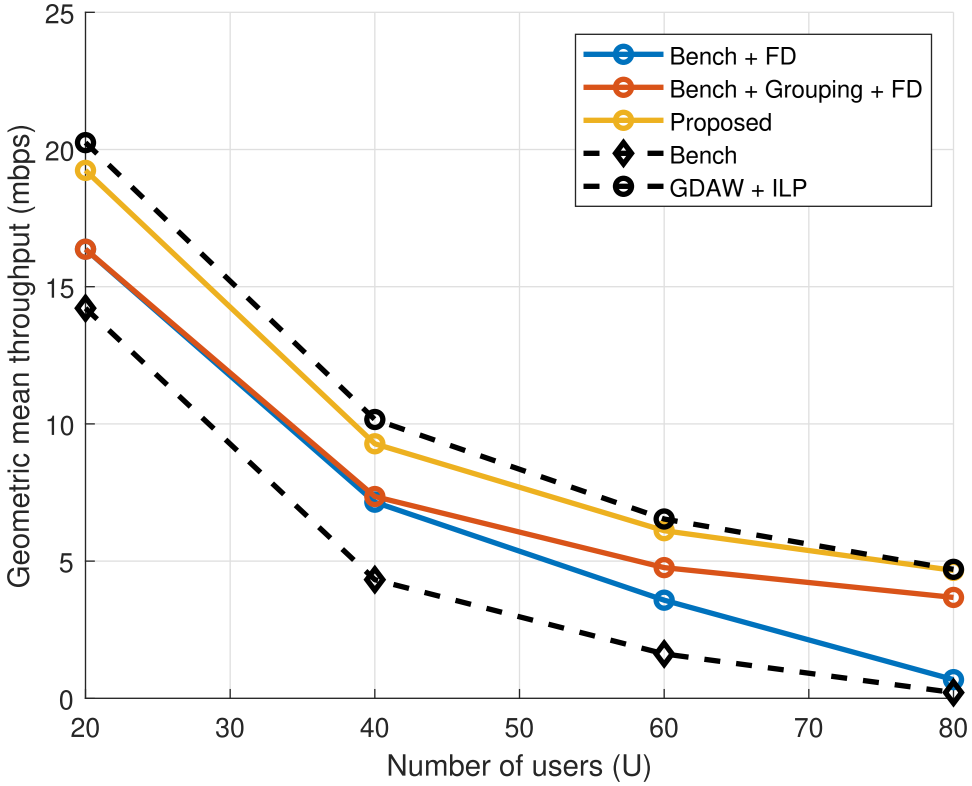

Figure 3.22: The geometric mean throughput achieved by adding different steps of the proposed network operation solution compared to the benchmark assuming $M = 100$ and $R = 200$. The offline target obtained via GDAW and ILP is included as a reference.

and MCS selection steps. The MCS selection step is often overlooked due to its supposed simplicity; yet, it has a significant influence on performance since it can trigger potential user de-selection and possible iterations. We have shown that iterations can yield significant performance gains but the impact on runtime is not negligible.

Careful distribution of power among the selected users is critical. The popular belief that water-filling power distribution is optimal, is only true if the rate and SNR relationship is governed by Shannon's capacity formula. However, practical limitations introduced by the usage of a finite set of MCSs makes this approach suboptimal. A significant drop in performance can be avoided by using the proposed greedy power distribution which jointly optimizes power distribution and MCS selection.

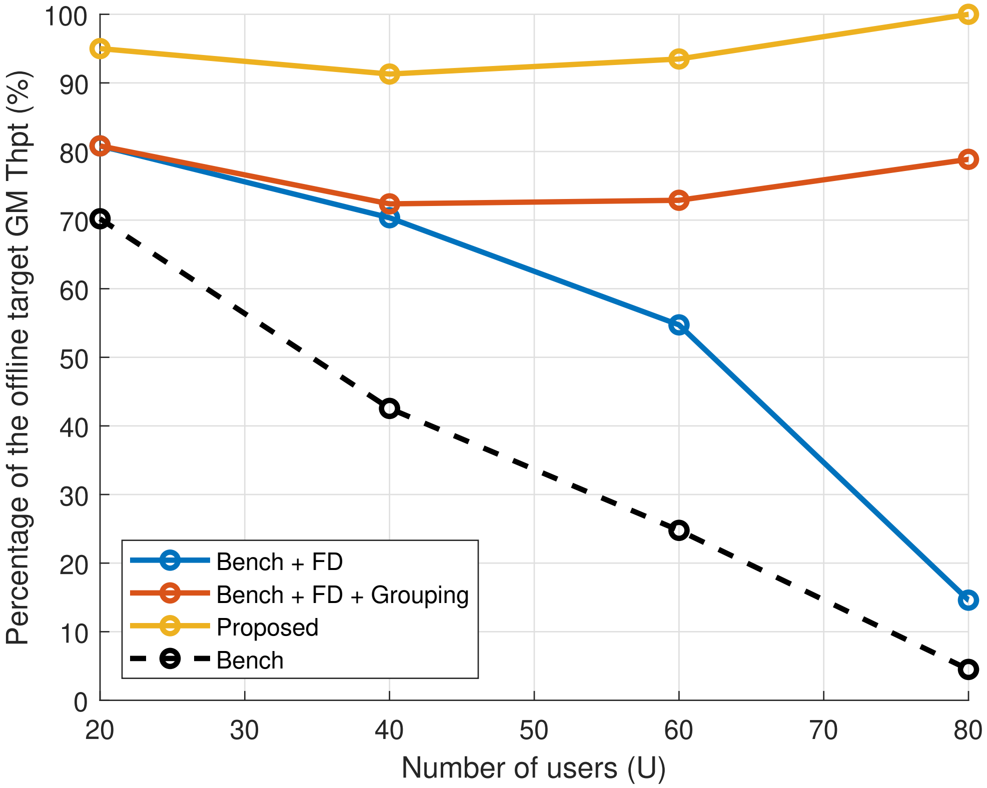

Figure 3.23: The percentage of the offline target geometric mean throughput performance achieved by adding different steps of the proposed network operation solution assuming $M = 100$ and $R = 200$.

What our results show is that the common myth that since in massive MIMO there are significantly more antennas than users, user selection is no longer needed, is not correct. Furthermore, while the most obvious user deselection strategy that removes all users with a rate of zero at once, full-drop, provides great gains with respect to the common approach, it is still far from the best solution in terms of geometric mean throughput performance if applied after initially selecting all users. Interestingly, random user grouping does not only provide a reduction in complexity, but also boosts performance. Coupling both full-drop and grouping can achieve in real-time 80% of the offline performance target, and when combined with the greedy power distribution algorithm, 95% of that target.

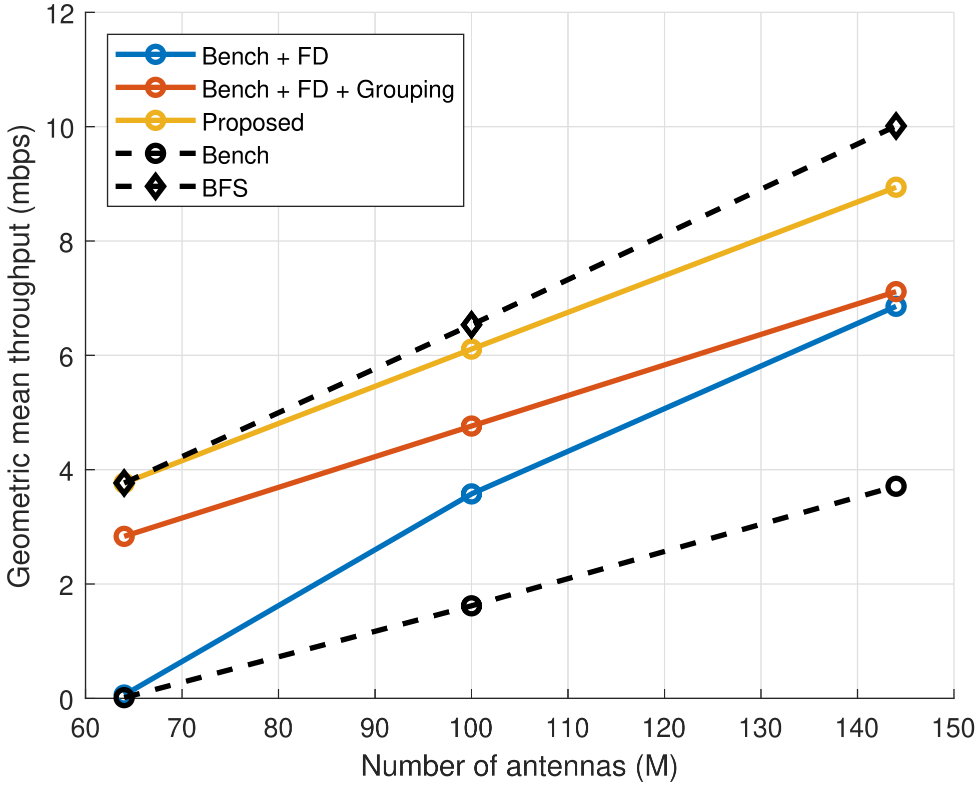

Figure 3.24: Geometric mean throughput achieved by adding different steps of the proposed network operation solution with respect to the best feasible solution obtained via GDAW and ILP assuming $U = 60$ and $R = 200$.

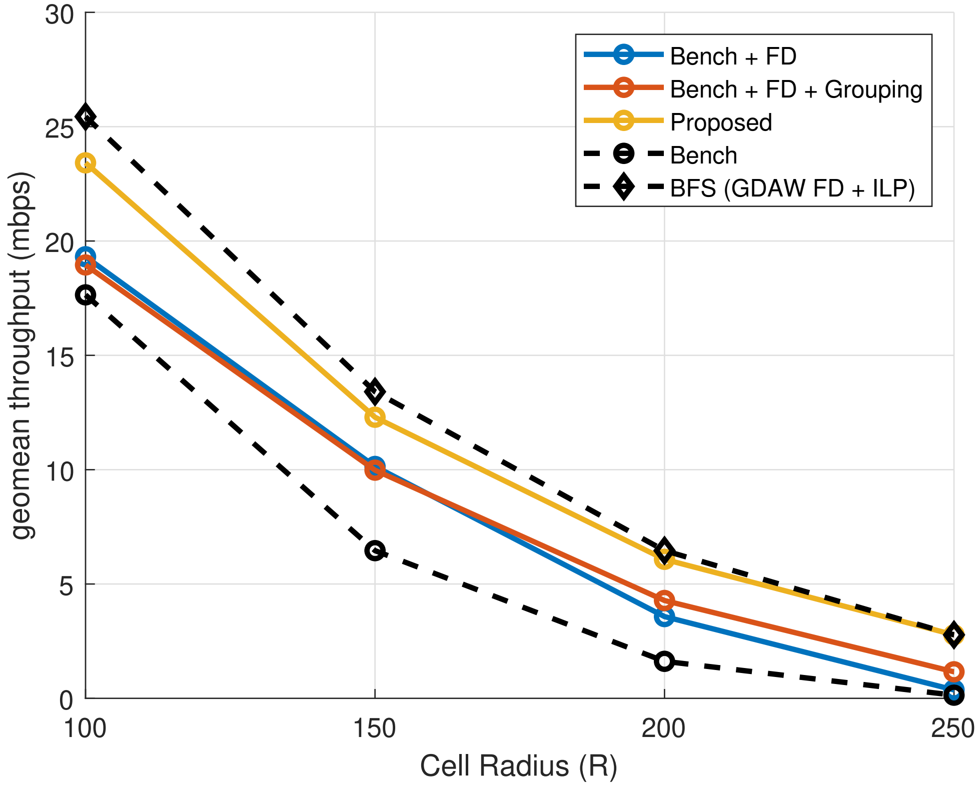

Figure 3.25: Geometric mean throughput achieved by adding different steps of the proposed network operation solution with respect to the best feasible solution obtained via GDAW and ILP assuming $U = 60$ and $M = 100$.

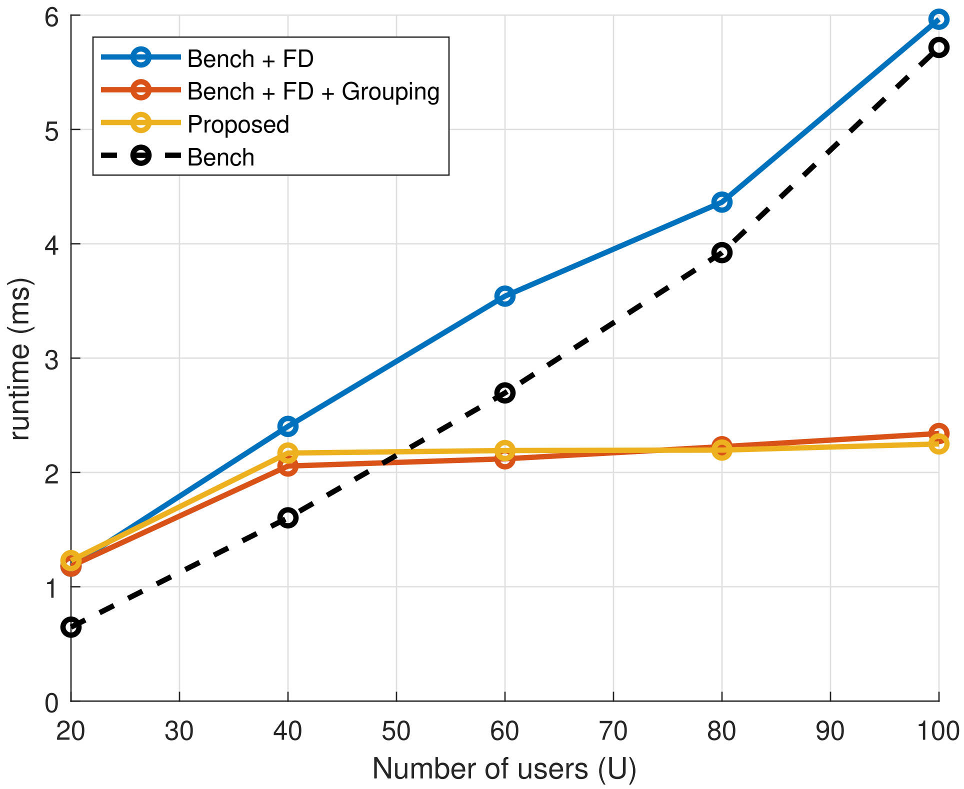

Figure 3.26: The runtime consumed by the different schemes assuming $M = 100$ and $R = 200$.

Chapter 4

Conclusion

4.1 Contributions

In this work, we have studied two multi-user transmission (MUT) techniques; namely, NOMA and massive MU-MIMO, proposed for 5G and beyond cellular networks. In both of the studies, we started by searching for the optimal or quasi-optimal performance achievable on the downlink, assuming an offline setting, i.e., no runtime limitations. The offline search allowed us to evaluate the usefulness of using such techniques in practice and to set a performance target which can be computed offline for the system under consideration.

We then proposed practical methods that can be used for real-time network operation. For NOMA, we first proposed a static power-map for power allocation. Then we proposed a family of schedulers with varying complexity and performance tradeoffs that dynamically perform both power distribution and user-selection.

The results indicate that both OMA and NOMA systems can benefit significantly from proper power allocation. Additionally, the results show that power allocation brings more gain than NOMA and that the percentage of SIC-capable devices affects NOMA gains. Furthermore, equipping users with NOMA-N for $N > 2$ results in significantly diminishing returns.

We also began our study of massive MU-MIMO by an offline study of the joint RRM problem. Even in an offline setting, the joint RRM problem, where we jointly optimize user-selection, power distribution and MCS selection, is too difficult to solve, despite the fact that it can be cast and solved using BRB (Branch-Reduce-and-Bound). BRB does, however, help to show the quasi-optimality of greedy search algorithms for use-cases with up to 30 users. Given a user set, the joint problem can be solved and then the problem become one of a search over user sets. The various search methods that we considered, including greedy search with different initial sets, have yielded nearly identical performance results and thus we chose the least complex method, GDAW + FD, to compute our performance target.

We then proposed a simple iterative per PRB procedure for operating massive MIMO networks. The procedure is based on grouping the users in random groups of size K_{RR}, ZFT precoding, a greedy algorithm for joint power distribution and MCS selection and if necessary, an additional iteration after performing full-drop of all users who saw an SNR less than Γ_1, and hence a zero-rate. Although the solution seems quite simple, numerical experiments shows that the system achieves, in real-time, performance very close to that achieved by the best-feasible solution found for the offline problem.

Showing that the proposed solutions are near-optimal was a challenging undertaking. In order to demonstrate that the proposed solutions were near-optimal, we had to employ complex optimization algorithms. One such algorithm was signomial programming, which is a type of non-convex optimization that was used to establish an offline performance target for NOMA systems. By using the proposed power maps as well as the simplified local scheduling (SLS) heuristic, near optimal performance was found to be achievable in a reasonable runtime.

For MU-MIMO, the joint RRM problem was solved using a BRB (Branch-Reduce-and-Bound) algorithm. This complex algorithm allowed us to efficiently search the large combinatorial space of user-selection, power distribution and MCS selection for ZFT-based MU-MIMO systems. For cases where BRB was infeasible due to the very large search space, we turned to alternative approaches using greedy search methods to find offline performance targets. Despite the computational complexity involved in our approach, we were able to demonstrate the effectiveness of our proposed solutions for user selection and power distribution in MU-MIMO systems.

4.2 Messages

We summarize the main messages below.

4.2.1 Opportunities for significant performance gains in RRM for NOMA and MU-MIMO systems

Clearly the first message is that by performing RRM carefully, huge gains can be obtained with respect to the existing benchmarks. More surprisingly maybe is that we were able to develop practical (i.e., fast) heuristics for MU-MIMO that perform quasi optimally.

4.2.2 The power of power allocation

Another important message is that in a multi-cell NOMA system, power allocation yields a much higher gain than that achieved by NOMA. With respect to the benchmark OMA with equal power allocation, we would gain 89.1% with static power allocation, while optimum NOMA-2, with equal power allocation, would only give a 38.5% increase. This NOMA-based improvement would necessitate UEs having access to SIC hardware and the base-station computing the optimal scheduling in realtime. A static power map, on the other hand, only needs the base-station to use an offline calibrated power map and no additional hardware at the UEs. Furthermore, adopting MU-MIMO in the downlink would be better to NOMA because it needs only the base-station to build more antennas rather than needing UEs to have additional hardware.

4.2.3 Navigating the Limitations of Shannon Capacity Formula in Practice: Challenges and Solutions

Shannon capacity formula quantifies the theoretical maximum capacity of a communication channel. The formula provides a mathematical expression for the maximum amount of data that could be transmitted reliably over a given communication channel as a function of the available bandwidth and the SINR. One of the advantages of the Shannon capacity formula is that it enables the abstraction of the key characteristics of the communication process as well as the physical properties of the channel. Furthermore, the formula is a concave and monotonically increasing function, which facilitates the study of system performance via optimization problems and the design of practical solutions.

For example, under the assumption that the rate function is defined using Shannon's capacity formula, the power distribution for both NOMA and ZFT-based MU-MIMO can be solved to optimality using the methods in [58] and [63], respectively.

Despite its popularity and usefulness, It is important to understand its limitations and the assumptions that Shannon capacity makes. Specifically, it assumes that the system uses an idealized channel code that ensures communications reliability, i.e., that the receiver can decode the transmitter's message successfully without errors. Practical systems, on the other hand, are limited by the set of predefined MCSs available in the system. This results in a piece-wise constant rate function that depends on the SINR of the channel. Solving the RRM problem with Shannon's capacity formula might lead to incorrect insights and sub-optimal design.

Shannon formula is misleading since it hides several key characteristics of the piece-wise constant rate function defined by the set of available MCSs. First, it neglects the critical fact that there is a minimum required SINR, Γ_1, for receiving a non-zero rate. Second, it indicates that the rate increase monotonically with any increase in power or bandwidth which is not the case with the practical piece-wise constant rate function. Finally, it ignores the fact that there is a maximum rate, $R_{\max}$, supported by the system as a result of using a finite set of MCSs. However, this last limitation can easily be addressed by adding a constraint in the optimization problems. It is worrying that this is not done more often in the literature.

With Shannon, we are more likely to distribute power to a large number of users in MU-MIMO systems than concentrate power on fewer users. This is a misleading insight if used to develop heuristics. As a result, solutions based on the Shannon capacity formula, such as waterfilling, often advocate distributing power among multiple users to maximize the overall system performance.

In practice, it is often more effective to focus the transmission power on a smaller number of users that can receive a strong signal, rather than distributing the power thinly over multiple users. Our results show that replacing Shannon-based solutions with MCS-aware solutions for power distribution can bring performance gains of up to 22%. This is shown for the proposed search method suggested for NOMA that searches for the best performing set of MCS in the nearest k_2 options. Also, it is shown for the greedy search method proposed as an alternative to waterfilling power distribution in ZFT MU-MIMO systems. Both of these solutions perform similarly in terms of runtime to Shannon-based solution while being superior in terms of performance.

Furthermore, Shannon's formula indicates that the rate increases as the available bandwidth increases. However, the UE's ability to correctly receive the signal is also dependent on the minimum SINR required, which is a function of the received signal power. If the minimum SINR is not achieved, even with a large bandwidth, the rate will not increase, and communication may not be even possible. As the bandwidth increases and the base-station is required to spread power thin, there comes a point where any further increase in bandwidth is no longer beneficial.

Handling the minimum SINR constraint is more challenging than handling the maximum rate constraint, as it makes the optimization problem non-convex and more difficult to solve. Incorporating the constraint that each user must have an SINR of at least Γ_1 into the optimization problem can result in a non-linear, non-convex problem that is difficult to solve. A practical approach for handling this challenging fact is discussed in the following.

4.2.4 The iterative approach to RRM in MUT systems is critical to balance performance and complexity

The first message here is that ignoring the minimum SINR constraint when distributing power is a very bad idea in terms of performance and insights.

We proposed an approach to handle the minimum SINR constraint after the fact, by revisiting the RRM decision, in the given PRB, if necessary. This provides a more practical solution than trying to handle the constraint up front. In this approach, power distribution is first solved without considering the minimum SINR constraint, and the resulting SINR for each user is checked to determine if any users have an SINR lower than Γ_1. If any users have an SINR lower than Γ_1, the RRM decision is revisited, as discussed next.

The minimum SINR constraint is typically not as critical in SUT-based systems as in MUT-based systems such as NOMA and MU-MIMO. The minimum SINR constraint is more likely to be violated with NOMA and MU-MIMO than SISO since multiple users sharing the same PRB (Physical Resource Block), means that the power needs to be distributed and hence "thinned" among those users while, at the same time, more interference is generated compared to SISO.

In NOMA systems, the users assigned a zero rate in a PRB can lead to significant power wastage. This is because these users are allocated power without contributing to the overall data rate. Hence, to conserve power and make a more efficient use of the available resources, it is important to revisit the RRM decision, i.e., remove some of these users by forcing their power to be zero in the PRB under consideration and redo the power distribution over the remaining users.

As a result, while Shannon capacity indicates that NOMA should always be employed (i.e., we should always select N users with NOMA-N). With real MCS, there is a high probability that using less than N users in a PRB is optimal. This is supported by our results in Chapter 2, which show that with practical MCS, it is better to use OMA instead of NOMA-2 20% of the time. Furthermore, with NOMA-3 with practical MCS, it is better to use NOMA-2 and OMA 30% and 20% of the time, respectively.

In MU-MIMO systems, the selection of the precoding matrix plays a crucial role in determining the SINR and the rate for each selected user. ZFT is a popular choice since it removes inter-user interference. However, when some of the users are assigned a zero rate, this results in power wastage as well as in inefficiencies in the ZF precoding matrix as it is designed for a larger group of users than necessary. Hence it is important to revisit the RRM decision and remove some of these zero rate users and recompute the ZF precoding matrix. Our solution based on a single additional iteration after removing all users that see a zero-rate results in up to 80% improvement in performance. Clearly, we could remove users one at a time but this would increase the number of iterations and the increase in performance with respect to our solution is not large.

4.2.5 Beyond necessity: The Role of Grouping in improving performance and Reducing Complexity in massive MIMO systems

In massive MIMO systems, the number of antennas is typically much larger than the number of available users and thus, the base-station can ignore user-selection and serve all users simultaneously. However, focusing on a smaller user set not only improves performance but also reduces complexity.

Our results show that by limiting the group size K_{RR} to a maximum of $M/3$, both performance and complexity improve compared to the benchmark that select all users. The performance improvement is a result of the base-station being able to focus and optimize the use of available resources on a smaller user set and maintain high system performance.

Furthermore, the complexity of the precoding matrix computation is a function of the total number of users selected. By selecting, a limited number of users this ensures that the complexity of computing the ZFT precoder is kept at a reasonable level.

4.2.6 Balancing Performance and Complexity: The Power of Search in Practical Wireless Systems

Although extensive search iterations are required for finding good solutions for RRM for offline studies, a well designed search strategy for user selection can make the process more cost-effective by reducing the amount of unnecessary iterations. By using a search strategy that focuses on the most promising solutions, we can avoid over-searching and still obtain good results. This approach is shown to reduce the computational complexity of the RRM process and make it more practical for real-world applications.

For NOMA, we proposed using a Simplified Local Scheduling (SLS) algorithm that can provide tailored performance and runtime requirements by tuning its parameters (k_1, k_2). The proposed SLS solution starts by sorting the users into sets. It computes the achievable performance for the first k_1 user sets, using the proposed power distribution method with parameter k_2 as discussed earlier in Chapter 2. Finally, it returns the best performing user set out of all searched sets. SLS provides a continuum of trade-offs depending on the values of k_1 and k_2, enabling tailored support for different performance and runtime requirement. Moreover, our results show that SLS can achieve 95% of the optimum performance found by exhaustive search in a fraction of the required runtime.

For MU-MIMO, we proposed a simple RRM algorithm, which involves a sequence of steps that efficiently compute an RRM decision of ZFT MU-MIMO networks. On a per PRB basis, the algorithm starts by selecting a group of users at random, for which we compute a ZFT precoding matrix and perform power distribution, with our proposed MCS-aware greedy algorithm, and MCS selection. Then we perform a single extra (full-drop) iteration where all users with zero-rate are dropped, and the precoding, power distribution and MCS selection are recomputed. Although the method limits the number of iterations, it has been shown to achieve approximately 95% of the performance obtained by more extensive search methods, which are only suitable for offline benchmarking. In addition, the proposed method can achieve significant performance gains compared to the benchmark while improving o the benchmark's runtime when the number of users exceeds 40 and performs comparably otherwise.

4.3 Future research directions

This thesis has considered two MUT (Multi-User Transmission) techniques and their usage for downlink cellular systems. While there is a lot of theoretical work on multi-user transmission techniques, practical network operation algorithms are required to harness the capacity gains in real systems with limitations due to finite MCS and run time.

First, the study on massive MIMO can be extended to general precoding to see how far is ZFT from optimal linear precoding with discrete MCS. Although, there are results showing that the relative gap with Shannon capacity diminishes quickly as the number of antennas increase, these are done for the case where the network operation is the sum-rate and the assumption that the rate function is modelled using Shannon's capacity formula.

Checking the case with practical MCS and a proportional fair optimization objective is interesting and it is left for future work.

Second, in the massive MIMO study, we assumed that all UEs are equipped with single antenna receivers. However, most UEs have access to a few antennas that can receive multiple streams at the same time. Although a simple solution would be to treat each UE antenna as an independent user or to allocate only one stream per UE, this is clearly sub-optimal. A further investigation is required to determine the best way to operate a network with multiple antenna UEs.

Third, in the massive MIMO study, we assumed a single cell system. The extension to multi-cell systems would require a revisit of the proposed equal power allocation and a solution using a power-map is likely to provide significant performance gain as shown in the case of SISO OMA [8] and NOMA [7].

Fourth, although we have studied both NOMA and massive MIMO independently, the combination of both is something potentially interesting and we leave it for future investigation. NOMA often needs users that have highly correlated channels to function efficiently, whereas MIMO typically demands that users have distinct channels. Therefore, both methods can complement each other.

Fifth, this work has focused on proportional fairness while different fairness objectives, such as max-min fairness or α-fairness, can potentially lead to different conclusions. Finally, all studied techniques have focused on the downlink and, the uplink is left for future work.

References

[1] J. Ghimire and C. Rosenberg, “Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, 2013.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[3] T. Cover, “Broadcast channels,” *IEEE Trans. Inform. Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[4] G. Caire and S. Shamai, “On the achievable throughput of a multiantenna gaussian broadcast channel,” *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[5] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, “A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends,” *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, 2017.

[6] F. P. Kelly, A. K. Maulloo, and D. K. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[7] A. Hussein, C. Rosenberg, and P. Mitran, “Hybrid NOMA in Multi-Cell Networks: From a Centralized Analysis to Practical Schemes,” *IEEE/ACM Transactions on Networking*, 2021.

[8] Y. Özcan, J. Oueis, C. Rosenberg, R. Stanica, and F. Valois, “Robust Planning and Operation of Multi-Cell Homogeneous and Heterogeneous Networks,” *IEEE Trans. Netw. Service Manag.*, pp. 1–1, 2020.

[9] J. Krause, "Study on scenarios and requirements for next generation access technology," *3GPP TR38*, vol. 913, 2016.

[10] J. H. Winters, J. Salz, and R. D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *IEEE transactions on Communications*, vol. 42, no. 234, pp. 1740–1751, 1994.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, 2005.

[12] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE transactions on wireless communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[13] T. Marzetta, *Fundamentals of massive MIMO.* Cambridge University Press, 2016.

[14] "Cisco annual internet report," Apr 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html

[15] 3GPP, " Study on Non-Orthogonal Multiple Access (NOMA) for NR," 3rd Generation Partnership Project (3GPP), Technical report (TR) 38.812, December 2018, version 16.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3236

[16] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 197–207, 1973.

[17] J. Ding, J. Cai, and C. Yi, "An improved coalition game approach for mimo-noma clustering integrating beamforming and power allocation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1672–1687, 2018.

[18] J. Wang, H. Xu, L. Fan, B. Zhu, and A. Zhou, "Energy-efficient joint power and bandwidth allocation for NOMA systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 780–783, 2018.

[19] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation in uplink mmwave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3000–3004, 2019.

[20] W. Hao, Z. Chu, F. Zhou, S. Yang, G. Sun, and K.-K. Wong, "Green communication for NOMA-based CRAN," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 666–678, 2018.

[21] J. Luo, J. Tang, D. K. So, G. Chen, K. Cumanan, and J. A. Chambers, “A Deep Learning-Based Approach to Power Minimization in Multi-Carrier NOMA With SWIPT,” *IEEE Access*, vol. 7, pp. 17 450–17 460, 2019.

[22] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, “Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, 2018.

[23] M. A. Sedaghat and R. R. Müller, “On user pairing in uplink NOMA,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3474–3486, 2018.

[24] Y. Zhang, H. Cao, M. Zhou, and L. Yang, “Spectral efficiency maximization for uplink cell-free massive MIMO-NOMA networks,” in *2019 IEEE International Conference on Communications Workshops (ICC Workshops).* IEEE, 2019, pp. 1–6.

[25] H. Wang, Y. Fu, Z. Shi, and R. Song, “Fractional power control for small cell uplinks with opportunistic NOMA transmissions,” in *ICC 2019-2019 IEEE International Conference on Communications (ICC).* IEEE, 2019, pp. 1–7.

[26] Ö. F. Gemici, İ. Hökelek, and H. A. Çırpan, “Noma power allocation for minimizing system outage under rayleigh fading channel,” in *IEEE 40th Sarnoff Symposium.* IEEE, 2019, pp. 1–6.

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

[28] A. Celik, F. S. Al-Qahtani, R. M. Radaydeh, and M.-S. Alouini, “Cluster formation and joint power-bandwidth allocation for imperfect noma in dl-hetnets,” in *GLOBECOM 2017-2017 IEEE Global Communications Conference.* IEEE, 2017, pp. 1–6.

[29] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, “Energy-Efficient Joint User-RB Association and Power Allocation for Uplink Hybrid NOMA-OMA,” *IEEE Internet Things J.*, 2019.

[30] J. A. Oviedo and H. R. Sadjadpour, “A fair power allocation approach to NOMA in multiuser siso systems,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7974–7985, 2017.

[31] W. Cai, C. Chen, L. Bai, Y. Jin, and J. Choi, “Subcarrier and power allocation scheme for downlink OFDM-NOMA systems,” *IET Signal Processing*, vol. 11, no. 1, pp. 51–58, 2016.

[32] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, 2016.

[33] Q.-V. Pham and W.-J. Hwang, "α-fair resource allocation in non-orthogonal multiple access systems," *IET Communications*, vol. 12, no. 2, pp. 179–183, 2018.

[34] A. Celik, M.-C. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M.-S. Alouini, "Distributed cluster formation and power-bandwidth allocation for imperfect noma in dl-hetnets," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1677–1692, 2018.

[35] P. Sindhu, A. H. KM *et al.*, "A novel low complexity power allocation algorithm for downlink NOMA networks," in *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2018, pp. 36–40.

[36] F. Liu and M. Petrova, "Performance of Proportional Fair Scheduling for Downlink PD-NOMA Networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 7027–7039, 2018.

[37] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, 2016.

[38] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, 2016.

[39] Z. Sheng, X. Su, and X. Zhang, "A Novel Power Allocation Method for Non-orthogonal Multiple Access in Cellular Uplink Network," in *IEEE International Conference on Intelligent Environments (IE)*, 2017, pp. 157–159.

[40] D. P. Palomar and Mung Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Select. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, 2006.

[41] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2015.

[42] J.-M. Kang and I.-M. Kim, "Optimal user grouping for downlink noma," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 724–727, 2018.

[43] P. Parida and S. S. Das, “Power allocation in OFDM based NOMA systems: A DC programming approach,” in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2014, pp. 1026–1031.

[44] A. Sayed-Ahmed, M. Elsabrouty, A. H. A. El-Malek, and M. Abo-Zahhad, “Energy efficient framework for multiuser downlink MIMO-NOMA systems,” in *2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2018, pp. 48–54.

[45] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, “Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster,” *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, 2017.

[46] J. Zhang, W. Xu, W. Chen, H. Gao, and J. Lin, “Joint subcarrier assignment and downlink-uplink time-power allocation for wireless powered ofdm-noma systems,” in *10th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2018, pp. 1–7.

[47] K. S. Ali, H. Elsawy, A. Chaaban, and M. Alouini, “Non-orthogonal multiple access for large-scale 5g networks: Interference aware design,” *IEEE Access*, vol. 5, pp. 21 204–21 216, 2017.

[48] C.-H. Lee, M. Kobayashi, H.-Y. Wei, S. Saruwatari, and T. Watanabe, “Adaptive resource allocation for icic in downlink NOMA systems,” in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. IEEE, 2019, pp. 1–6.

[49] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[50] R. Mazumdar, L. G. Mason, and C. Douligeris, “Fairness in network optimal flow control: Optimality of product forms,” *IEEE Transactions on communications*, vol. 39, no. 5, pp. 775–782, 1991.

[51] F. Kelly, “Charging and rate control for elastic traffic,” *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[52] H. Yaïche, R. R. Mazumdar, and C. Rosenberg, “A game theoretic framework for bandwidth allocation and pricing in broadband networks,” *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 667–678, 2000.

[53] G. Xu, "Global optimization of signomial geometric programming problems," *European Journal of Operational Research*, vol. 233, no. 3, pp. 500–510, 2014.

[54] M. Chiang, C. W. Tan, D. P. Palomar, D. O'neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, 2007.

[55] 3GPP, "Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz (Release 15)," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.901, 06 2018, version 15.0.0.

[56] Y. Ozcan, "Resource management in next generation cellular networks," *UWSpace*, 2019. [Online]. Available: http://hdl.handle.net/10012/14787

[57] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *IEEE Symp. on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2013, pp. 611–615.

[58] L. Salaün, M. Coupechoux, and C. S. Chen, "Joint subcarrier and power allocation in noma: Optimal and approximate algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2215–2230, 2020.

[59] M. Series, "Guidelines for evaluation of radio interface technologies for imt-advanced," *Report ITU*, vol. 638, pp. 1–72, 2009.

[60] E. Castaneda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 239–284, 2016.

[61] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 142–148, 2014.

[62] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4409–4418, 2008.

[63] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European transactions on telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[64] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor, "New viewpoint and algorithms for water-filling solutions in wireless communications," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1618–1634, 2020.

[65] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[66] A. F. Molisch, *Wireless Communications: From Fundamentals to Beyond 5G*. John Wiley & Sons, 2020.

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

[68] Y. Özcan, J. Oueis, C. Rosenberg, R. Stanica, and F. Valois, "Robust planning and operation of multi-cell homogeneous and heterogeneous networks," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1805–1821, 2020.

[69] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, 2015.

[70] H. Yang, "User Scheduling in Massive MIMO," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 1–5.

[71] H. Yang and T. L. Marzetta, "Massive mimo with max-min power control in line-of-sight propagation environment," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4685–4693, 2017.

[72] R. C. Elliott and W. A. Krzymien, "Downlink scheduling via genetic algorithms for multiuser single-carrier and multicarrier MIMO systems with dirty paper coding," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 7, pp. 3247–3262, 2008.

[73] Y. Hei, X. Li, K. Yi, and H. Yang, "Novel scheduling strategy for downlink multiuser MIMO system: Particle swarm optimization," *Science in China Series F: Information Sciences*, vol. 52, no. 12, pp. 2279–2289, 2009.

[74] G. Dimic and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Transactions on Signal processing*, vol. 53, no. 10, pp. 3857–3868, 2005.

[75] J. Wang, D. J. Love, and M. D. Zoltowski, "User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3713–3726, 2008.

[76] S. Huang, H. Yin, J. Wu, and V. C. Leung, "User selection for multiuser MIMO downlink with zero-forcing beamforming," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3084–3097, 2013.

[77] S. Huang, H. Yin, H. Li, and V. C. Leung, "Decremental user selection for large-scale multi-user MIMO downlink with zero-forcing beamforming," *IEEE Wireless Communications Letters*, vol. 1, no. 5, pp. 480–483, 2012.

[78] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian mimo broadcast channel," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.* IEEE, 2004, p. 174.

[79] H. Tuy, F. Al-Khayyal, and P. T. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization.* Springer, 2005, pp. 39–78.

[80] K. Shen and W. Yu, "Fractional programming for communication systems—part i: Power control and beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[81] A. A. Khan, R. S. Adve, and W. Yu, "Optimizing downlink resource allocation in multiuser MIMO networks via fractional programming and the hungarian algorithm," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5162–5175, 2020.

[82] Y. Zhang, P. Mitran, and C. Rosenberg, "Joint Resource Allocation for Linear Precoding in Downlink Massive MIMO Systems," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3039–3053, 2021.

[83] E. Björnson and E. Jorswieck, *Optimal resource allocation in coordinated multi-cell systems.* Now Publishers Inc, 2013.

[84] M. Chiang, *Geometric programming for communication systems.* Now Publishers Inc, 2005.

[85] L. B. Le, E. Hossain, and D. Niyato, *Radio resource management in multi-tier cellular wireless networks.* John Wiley & Sons, 2013.

[86] I. I. Cplex, "V12. 1: User's manual for cplex," *International Business Machines Corporation*, vol. 46, no. 53, p. 157, 2009.

[87] "MATLAB Optimization Toolbox," 2022b.

[88] H. Tuy, T. Hoang, T. Hoang, V.-n. Mathématicien, T. Hoang, and V. Mathematician, *Convex analysis and global optimization.* Springer, 1998.

[89] Y. Özcan and C. Rosenberg, "Revisiting Downlink Scheduling in a Multi-Cell OFDMA Network: From Full Base Station Coordination to Practical Schemes," in *Wireless Days (WD)*, 2019, pp. 1–8.