Design and Performance Analysis of Dynamic Resource Allocation in OMA and NOMA Networks

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines Doktors der
Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Fei Liu, M.Sc.
aus Henan, China

Berichter: Prof. Dr.-Ing. Marina Petrova
Prof. Dr.-Ing. Eduard A. Jorswieck

Tag der mündlichen Prüfung: 6. November 2019

Abstract

During the last decades, wireless cellular networks have been developing rapidly, moving from supporting only analog voice services to supporting a plethora of digital mobile applications. In this continuous and intensive evolutionary process, radio access technology has always played an important role in enabling efficient multi-user access to the limited available spectrum. Moreover, dynamic and efficient allocation of radio resources is necessary for supporting reliable and high-speed multi-user connectivity under fading and time-varying wireless channels. This thesis focuses on the dynamic resource allocation problems in both of the orthogonal and non-orthogonal multiple access networks. The major objectives of this work are to design highly efficient dynamic resource allocation schemes, in particular, user scheduling, channel and power allocation for optimized system performance, and develop analytical solutions to assess their performance in different communication scenarios.

In the first part of the thesis, we focus on the performance analysis of dynamic resource allocation in the orthogonal multiple access networks. First, we build stochastic channel models for depicting the probability distributions of instantaneous channel states in multi-cell networks. Based on our channel models, we develop a framework for performance analysis of various resource allocation schemes, such as the max-sum rate, max-min rate, and proportional fair scheduling. We utilize the analytical performance to estimate ergodic user data rates. The estimation results are verified to be very accurate even with limited channel state information. To present a practical application of our analytical performance, we then use the estimated data rates to assist user association and design an inter-cell handover scheme for traffic load balancing in heterogeneous cellular networks. Furthermore, we extend our performance analysis of dynamic resource allocation to the case of on-off bursty traffic flows, which are used for modeling the increasingly popular streaming services at the session level.

As one promising candidate radio access technology for the upcoming 5G networks, non-orthogonal multiple access (NOMA) introduces a new dimension of user multiplexing in the power domain. Thus, in the second part of the thesis, we address the power allocation problem and design dynamic channel and power allocation (DCPA) schemes for NOMA networks. Our contributions in this part are twofold. First, we design novel low-complexity DCPA schemes for single-channel and multi-channel NOMA systems, respectively. Compared to the other existing schemes in the literature, our proposed solutions significantly improve the transmission performance of NOMA with extremely reduced computational complexity, and are more favorable to practical applications. Then, we develop, as far we are aware of, the first analytical solution to the performance of DCPA for NOMA systems and apply it to user data rate estimation. With different system configurations, we study the impacts of multi-user and multi-channel diversities on the performance of

NOMA, serving as a guideline for its optimization and implementation in the future networks.

Kurzfassung

In den letzten Jahrzehnten haben sich die drahtlosen Mobilfunknetzwerke, die erst nur analoge Sprachdienstleistungen bereitstellten, rasant entwickelt, sodass heute eine Fülle von digitalen mobilen Anwendungen unterstützt wird. In diesem kontinuierlichen und intensiven evolutionären Prozess hat die Funkzugangstechnologie seit jeher eine wichtige Rolle bei der Ermöglichung einer effizienten Verteilung von Multi-Nutzer-Zugriffen auf das beschränkt verfügbare Spektrum gespielt. Außerdem ist eine dynamische und effiziente Zuweisung von Funkressourcen notwendig, um eine zuverlässige und schnelle Multi-Nutzer-Konnektivität bei schwindenden und zeitvariierenden Funkkanälen zu unterstützen. Diese Dissertation konzentriert sich auf die Zuweisungsprobleme von dynamischen Ressourcen sowohl in orthogonalen als auch in nicht-orthogonalen Vielfachzugriff-Netzwerken. Es wird in dieser Dissertation darauf abgezielt, hocheffiziente dynamische Algorithmen zur Ressourcenzuweisung, insbesondere zur Benutzerzeitplanung und zur Kanal- und Leistungszuweisung für eine optimierte Systemleistung, zu entwerfen und die analytischen Lösungen zur Näherung ihrer Leistung in verschiedenen Kommunikationsszenarien zu entwickeln.

Im ersten Teil der Dissertation wird auf die Leistungsanalyse der dynamischen Ressourcenzuweisung in orthogonalen Vielfachzugriff-Netzwerken eingegangen. Zunächst werden stochastische Kanalmodelle entwickelt, um die Wahrscheinlichkeitsverteilungen von Zuständen des Momentankanals in Mehrzellen-Netzen darzustellen. Auf Basis dieser Kanalmodelle wird ein Rahmen für die Leistungsanalyse verschiedener Zuweisungsalgorithmen der Ressourcen entwickelt, wie zum Beispiel die Max-Sum-Rate, die Max-Min-Rate und die proportionale faire Zeitplanung. Die analytische Leistung wird verwendet, um ergodische Nutzerdatenraten zu nähern. Diese Näherungen werden auch bei begrenzten Kanalzustandsinformationen als sehr genau verifiziert. Um eine praktische Anwendung der analytischen Leistung zu präsentieren, werden dann die genäherten Datenraten genutzt, um die Nutzerzuordnung zu unterstützen und einen Inter-Zellen-Verbindungsübergabealgorithmus für den Verkehrslastausgleich in heterogenen Mobilfunknetzen zu entwerfen. Darüber hinaus wird die Leistungsanalyse der dynamischen Ressourcenzuweisung auf den Fall von on-off stoßweisen Verkehrsströmen erweitert, die zur Modellierung der immer beliebter werdenden Streaming-Dienste auf dem Sitzungsniveau eingesetzt werden.

Als ein vielversprechender Kandidat der Funkzugangstechnologie für die kommenden 5G-Netze führt nicht-orthogonaler Vielfachzugriff (NOMA) eine neue Dimension des Benutzermultiplexings im Leistungsbereich ein. Daher wird im zweiten Teil der Dissertation das Problem der Leistungszuweisung behandelt und es werden dynamische Kanal- und Leistungszuweisungsalgorithmen (DCPA) für NOMA-Netze entwickelt. Dabei werden zwei Neuheiten vorgestellt. Zunächst werden neuartige DCPA-Schemata mit geringer Komplexität für Ein- bzw.

Mehrkanal-NOMA-Systeme entwickelt. Im Vergleich zu den anderen in der Literatur vorgestellten Ansätzen haben diese Lösungsvorschläge die Übertragungsleistung von NOMA mit extrem reduzierter Rechenkomplexität deutlich verbessert und sind damit für die praktische Anwendungen besser geeignet. Außerdem wird die nach bestem Wissen erste analytische Lösung für die Leistung von DCPA für NOMA-Systeme entwickelt und bei der Näherung der Nutzerdatenrate angewendet. Mit unterschiedlichen Systemkonfigurationen werden die Auswirkungen von Multi-Nutzer- und Multi-Kanal-Diversitäten auf die Leistung von NOMA untersucht, die als Richtlinie für deren Optimierung und Einführung in zukünftigen Netzwerken dienen kann.

Acknowledgements

This six-year journey towards Ph.D. has been a life-changing experience for me. I would like to express my appreciation towards all those who have helped and supported me along this journey.

Foremost, my deep gratitude goes to my doctoral supervisor, Professor Marina Petrova, who has given me continued support and guidance of my Ph.D. study and research. Her immense knowledge, strong motivation, and constructive suggestions have greatly promoted my research and contributed to my thesis work. Without her persistent encouragement and help, this thesis would never have been completed.

I would like to express my sincere thanks to Professor Petri Mähönen for the fruitful discussion and his helpful feedback on my research. His profound insight in both academic and industrial fields has inspired me deeply.

I would also like to thank Professor Eduard Jorswieck for his efforts in reviewing my thesis and valuable comments on it.

It was my fortune to work with many nice and intelligent colleagues in the iNETS group, who have given me countless help. For this, I want to thank Dr.-Ing. Ljiljana Simić, Dr.-Ing. Andrea Munari, Andra Voicu, Nikos Perpinias, Ping Ren, and Lars Kuger. I am deeply grateful to Dr.-Ing. Janne Riihijärvi, who gave me many valuable suggestions for my research and contributed to my modeling and analytical work. My special gratitude goes to Dr.-Ing. Peng Wang for generously sharing his knowledge and experience with me, and the unforgettable friendship we have built.

Finally, I want to express my deepest thanks to my parents for supporting me spiritually throughout this long journey and my life in general. I would also like to thank my cousin, Xiaomeng, who has always enlightened me whenever I met difficulties and challenges.

Contents

1 Introduction

Over the past few decades, mobile communication systems have been evolved through countless innovations, resulting in significantly improved performance and largely reduced cost in every generation. To meet the rapidly increasing demands of high-speed mobile services nowadays, the limited radio spectrum must be efficiently utilized for enhancing the data rates of wireless transmissions. In a multi-user access system, the served user links have different and time-varying radio channels. To improve the spectrum efficiency, it is necessary to allocate the radio resources to the user links according to their instantaneous channel states. This operation is referred to as the dynamic resource allocation and plays an important role in the transmission performance of multi-user access networks.

This thesis focuses on the dynamic resource allocation problems in both of the orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) networks. We investigate fundamental characteristics of dynamic resource allocation for performance optimization, address the optimization problems by designing low-complexity algorithms, and build analytical models for their performance analysis. In Section 1.1, we discuss the motivation of this work. In Section 1.2, the main contributions of the thesis are elaborated. The outline of the subsequent chapters is summarized in Section 1.3.

1.1 Motivation

As an underlying and essential part in wireless communication systems, radio access technology (RAT) supports the mobile devices to access the communication network through radio channels. In the past few decades, every generation of wireless communication systems has its emblematic innovation of RAT. The first generation (1G) has fulfilled the basic mobile voice services with frequency division multiple access (FDMA). The second generation (2G) has improved network capacity and coverage by time division multiple access (TDMA). This is followed by the third generation (3G), which adopts code division multiple access (CDMA) and supports high-speed data services. The fourth generation (4G) provides access to a wide range of mobile services with orthogonal frequency division multiple access (OFDMA). In the upcoming fifth generation (5G) networks, NOMA is one promising RAT candidate for the performance enhancement, which introduces a new degree of freedom in the power domain [1]. While using different RATs, the radio resources are allocated to the users in different forms, namely, frequency bands, time slots, spreading codes, time-frequency blocks, and transmit power levels. The radio channels have the natural feature of variance and fluctuation for different user links in wireless networks. To cope with this problem and improve the system performance, it is necessary to implement dynamic resource allocation according to

the instantaneous channel states. This is also referred to as opportunistic scheduling since the users are normally scheduled on the channels when their channel qualities are relatively good.

The RATs employed in the 1G to 4G mobile networks are based on OMA. This means that user signals are separable by keeping orthogonality in time, frequency, or code domain, so that the inter-user interference can be effectively avoided. Therefore, the major problem to be addressed in the dynamic resource allocation schemes for OMA systems is how to assign the divisible radio resources to multiple users. There have been a large number of schemes designed for OMA systems to meet the requirements of various mobile services [2]. Besides the designing work of dynamic resource allocation schemes, their performance analysis is also significant for their comparison and practical applications. The analytical performance can be used to estimate or predict the obtainable user data rates with a given resource allocation scheme and assists system operations, such as user association, cell handover, radio resource management, etc.

To analyze the performance of dynamic resource allocation, it is necessary to consider three major factors that determine its behavior, i.e., the channel states, the user data traffic flows, and the scheduling strategy. In the existing research works, however, some of the factors have been studied with only simplified models for the sake of making the performance analysis tractable [3–7]. For instance, the accurate analytical models of user channel states have not been developed for the multi-interference scenario which yet can be foreseen in the future cellular networks. Therefore, we are motivated to build accurate stochastic channel models and to analyze the performance of various dynamic resource allocation schemes, such as the max-sum rate, max-min rate, proportional fair scheduling, and potential schemes in the 5G mobile networks.

Compared to the OMA technique, NOMA offers a set of desirable benefits, including higher spectrum efficiency, dense connections, and better user fairness [8]. Thus, it is considered to be one promising RAT candidate for the next generation of wireless networks. It realizes power-domain multiplexing by using successive interference cancellation (SIC) for multi-user detection [9]. Hence, NOMA allows multiple user signals to be superposed within the same physical channel. However, this brings new challenges to the design of dynamic resource allocation since the power allocation problem needs to be addressed together with the channel allocation problem.

In order to improve system performance, it is necessary to jointly design the dynamic channel and power allocation (DCPA) for NOMA systems. This problem has attracted much recent research interests. There have been several DCPA schemes proposed in the existing research which are based on searching or iterative approaches [10–13]. However, these schemes suffer from such high computational complexity that they can hardly be applied in practice. Some efforts have been made to reduce the complexity of DCPA but the system performance is far from the optimal one [14,15]. Therefore, we embark on the challenge to design low-complexity DCPA schemes with the optimal or close-to-optimal performance for practical applications in NOMA systems. In addition, since the DCPA schemes are complex

and normally intractable in NOMA systems, there is a lack of performance analysis in the literature so far. The research work in this direction provides guidelines for the optimization and implementation of DCPA schemes. Thus, we are motivated to develop analytical solutions to the performance of DCPA for NOMA systems.

1.2 Contributions

The main contributions of this thesis cover two main aspects: we build an analytical framework for data rate performance analysis of dynamic resource allocation schemes based on stochastic modeling of time-varying wireless channels in multi-cell networks, and we address the DCPA problems for the power-domain NOMA, including the design and performance analysis of DCPA schemes in single-channel and multi-channel NOMA systems. More precisely, the contributions of the thesis are elaborated as follows.

1. We build a stochastic model of the instantaneous user channel state under Rayleigh fast fading, depicting the impact of inter-cell interference in the multi-cell networks. It is derived into a concise and closed-form expression and provides access to accurate estimation of the probability distributions of instantaneous signal-to-interference-plus-noise ratios (SINRs) based on the channel state information (CSI) feedback. Considering the limited available CSI in practice, we further formulate two distribution models of SINR that are proved to be the upper and lower bounds, respectively. Based on them, we propose a weighted-sum approximation model in order to improve the accuracy of our models with partial CSI feedback. To the best of our knowledge, this is the first analytical work on the stochastic channel modeling with partial CSI in multi-cell networks.

   Then, we apply our stochastic channel models to the data rate performance analysis of dynamic resource allocation. By modeling the features of user channel states, we analyze the ergodic user data rates under the scheduling schemes with various targets, including the max-sum rate (MSR), max-min rate (MMR), and proportional fairness (PF). The analytical performance can be utilized to estimate the obtainable data rate per user in the long run even though they have not yet been actually scheduled. We evaluate and compare the estimation results by simulations with different configurations and scenarios to assess the impacts of various factors on the estimation accuracy, including the radio propagation environments, base station deployment, user distributions, limited and imperfect CSI feedback, etc. We also compare our analytical model with the ones presented in the existing research to verify its superiority in the multi-cell networks. The results of this analytical work have been partially published in two conference papers [16, 17].

2. We apply our stochastic channel models and data rate estimation results to assisting user association and inter-cell handover in heterogeneous cellular

networks (HCNs). The cellular networks have a trend to be heterogeneous and densely deployed with the aim of high spatial reuse. This leads to potential unbalanced traffic loads in different types of cells. To alleviate this problem, we design user association schemes oriented towards traffic load balance. By simulations, we evaluate their performance in HCNs with various deployments. Compared to the conventional user access strategy, which is based only on the mean user channel qualities, our proposed schemes can effectively reduce the traffic load imbalance and consequently enhance throughput and user fairness. This part of our work has been published in two conference papers [17, 18].

3. We extend our performance analysis of dynamic resource allocation by taking into account the impact of bursty traffic flows. We consider high-speed streaming media which is an increasingly popular service type nowadays. It can be modeled as the bursty on-off traffic at the session level. With this traffic model, we derive the analytical performance of various dynamic resource allocation schemes, including round-robin (RR), MSR, and MMR. We utilize the results of system-level simulations to verify our performance analysis. The scheduling behavior of the proportional fair scheduling (PFS) scheme depends on both of the instantaneous channel qualities and historical scheduling results. Thus, the uncertainty of the user data flows makes its performance analysis intractable in the wireless environment. To address this problem, we design a hybrid approximation method to estimate its performance under bursty on-off traffic, combining our analysis in the case of saturated traffic and the Gaussian approximation (GA) method proposed in [14]. By simulation results, it is confirmed to be highly accurate and is insensitive to the changes in the system parameters. Our research paper on the hybrid approximation has been published in IEEE Communications Letters [19].

4. In the context of upcoming 5G networks, we address the DCPA problems in the power-domain NOMA systems. Compared to the OMA system, the multiple users in the NOMA system share not only the orthogonal radio resources but also the transmit power budget on them. Therefore, it is necessary to solve the dynamic inter-user power allocation and channel allocation problems jointly. Consistent with the goal of NOMA, i.e., making improvement in terms of both system efficiency and user fairness, we adopt the proportional fairness as the optimization objective in our design of DCPA schemes. We first consider the DCPA problem in the single-channel NOMA (SC-NOMA) system and assume that the SIC receiver supports an arbitrary number of multiplexed users. We derive a closed-form solution to the optimal power allocation for this ideal SC-NOMA system. Based on it, we design a low-complexity algorithm for joint power allocation and user set selection (USS). This DCPA scheme serves as a benchmark and provides a reference upper bound performance. Then, we derive its data rate performance based on our analytical framework and stochastic channel models built in the first part. To the best of our knowledge, this is the first analytical solution to the performance of DCPA in NOMA systems.

Due to the limited decoding capability of SIC receivers in practice, the number of multiplexed user on a channel is finite. Taking this limitation into account, we design two additional USS schemes based on our optimal solution of power allocation, including an optimal tree-searching-based USS (TSU) scheme and a suboptimal preselection-based USS (PSU) scheme. Compared to the existing schemes in the literature, the TSU scheme reduces the computation complexity for obtaining the optimal solution, while the PSU scheme achieves close-to-optimal performance with the lowest complexity. We utilize the analytical results of the upper bound performance to estimate the ergodic user data rates obtained in the practical NOMA systems. The estimation accuracy is evaluated by simulation results under various influencing factors. The research related to the DCPA problems in the SC-NOMA systems has been published in three conference papers [20–22] and a journal article in IEEE Transactions on Wireless Communications [23].

5. Finally, we extend our proposed DCPA schemes for the SC-NOMA systems to the multi-channel (MC) cases. In an MC-NOMA system, there are multiple subchannels instead of a single channel in the SC-NOMA case. In addition to the inter-user power allocation problem within each subchannel, it is necessary to implement inter-channel power allocation according to the time-varying channel states of the multiple subchannels. This makes the design of DCPA schemes more complex for MC-NOMA since the intra- and inter-channel power allocation problems need to be addressed jointly. We first propose a low-complexity optimal DCPA scheme for the ideal MC-NOMA systems based on our solution for SC-NOMA. Similarly, it obtains the upper bound performance for the practical cases in which the number of multiplexed users per subchannel is limited.

   Then, we design a novel user-preselection (UP)-based DCPA scheme, taking into account the limitation of SIC for practical MC-NOMA systems. The optimal solution developed for the ideal MC-NOMA is utilized as a preselection of the candidate user sets in order to reduce the computational complexity for their comparison. The simulation results indicate that it achieves close-to-optimal performance with extremely low computational complexity. In addition, we analyze the upper bound performance based our derivation in the study of SC-NOMA systems. This analytical work fills the research gap of the performance analysis for DCPA in MC-NOMA systems. We utilize the analytical results for data rate estimation in practical MC-NOMA systems and investigate the estimation accuracy under different system configurations. Furthermore, the impacts of multi-user and multi-channel diversities on the performance of MC-NOMA are analyzed and discussed for guiding the optimization and implementation of MC-NOMA systems in practice. The research results of our proposed DCPA schemes were published in the conference of IEEE PIMRC 2018 and won the best student paper award [24]. The analytical work on the performance of MC-NOMA has been published in IEEE Communications Letters [25].

1.3 Outline

The rest of the thesis is organized as follows. In Chapter 2, we provide a comprehensive background to the problems of dynamic resource allocation. We also describe the fundamentals of NOMA and the challenges in the design of DCPA schemes for it. The relevant state of the art in this field is reviewed and discussed. Chapter 3 presents our stochastic channel modeling for multi-interference scenarios and the analytical framework for data rate analysis and estimation. The simulation results are presented and compared to the analytical ones in order to evaluate the estimation accuracy under different configurations and scenarios. In Chapter 4, we adopt the analytical performance to predict user data rates and propose user association and handover schemes for traffic load balancing in HCNs. In Chapter 5, we study the performance of various resource allocation schemes under bursty on-off traffic flows, including RR, MSR, and MMR, and propose a hybrid approximation method for the performance analysis of the PFS scheme. Chapter 6 presents our study on the DCPA schemes in the SC-NOMA systems, including their designing process, computational complexity, optimality, performance analysis and evaluation, etc. Then, the extended study for the MC-NOMA case is presented in Chapter 7. Finally, we draw conclusions and discuss some potential directions for future research in Chapter 8.

2 Background

In this chapter, we introduce the relevant background knowledge. Section 2.1 describes the dynamic resource allocation problems and basic system models in downlink cellular networks. In Section 2.2, we review the research on the performance analysis of different resource allocation schemes and outline their pros and cons. The dynamic channel and power allocation (DCPA) problems in NOMA systems and the DCPA schemes proposed in the existing research works are reviewed and discussed in Section 2.3.

2.1 Dynamic Resource Allocation

In wireless cellular networks, user devices are associated with base stations (BSs) for accessing the core network and mobile services. BSs provide centralized control of user access and data transmission via wireless links, acting as the radio access network of mobile communications. To serve the associated users and exchange data with them, radio resources are allocated to the user links for carrying the transmitted data. From another perspective, this operation is also called user scheduling since users are mapped to the limited radio resources.

In comparison to the wired connections, one of the most notable features of the wireless channels lies in the time variation. Therefore, the dynamic resource allocation based on the instantaneous channel state information (CSI) is necessary for improving the transmission reliability and efficiency of wireless systems. In this thesis, we focus on the data transmissions in the downlink direction, i.e., traffic flows from BSs to users, since they account for the majority of the data flows generated

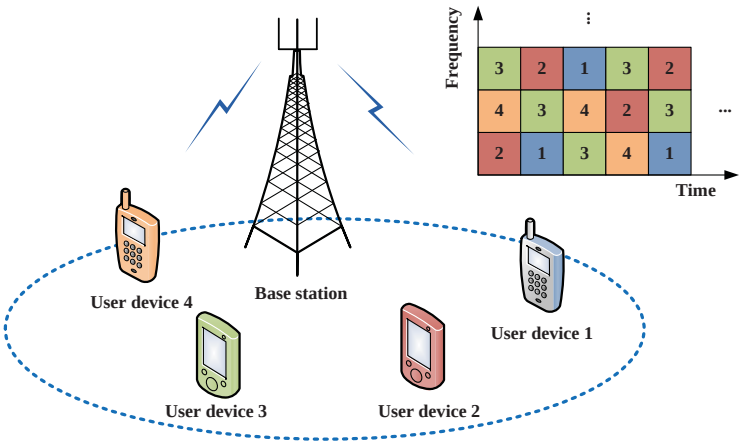


Figure 2.1: Illustration of resource allocation in a downlink OFDMA cellular network.

by the mobile applications. One example of the dynamic resource allocation in the downlink cellular network with OFDMA is presented in Figure 2.1. The orthogonal time-frequency blocks, i.e., the resource blocks (RBs), are assigned to different user links according to their instantaneous CSI.

In order to enhance the aggregate throughput, the limited radio resources should be allocated to the user links with better channel qualities. However, this leads to an unfair situation in which the cell-edge users with poor channel conditions are starved of service. Therefore, a common approach is to schedule the users when their channel qualities are relatively better than their average levels over time. This is referred to as opportunistic scheduling [2]. The scheduling behavior of a dynamic resource allocation scheme is generally determined by three major factors: instantaneous user channel states, user traffic flows, and the scheduling strategy. They are introduced and discussed specifically in the following parts.

2.1.1 User Channel States

To implement dynamic resource allocation, CSI is essential for the BS. It is normally modeled with the block fading in the resource allocation problems. This means that the fast fading factor is approximately constant within a given resource block in both of the time and frequency domains. Therefore, the BS can estimate user channel states in a resources block and determines the scheduled user and the data rate to be assigned. The data rate control mechanism is realized by the adaptive modulation and coding [26]. For ease of description, we first consider a downlink transmission system serving a set of users with one narrow-band channel in the frequency domain. The user set is denoted as

$$\mathbf{U} = \{u \,|u = 1, \ldots, U\}\,. \tag{2.1}$$

where U is the number of associated users in the cell. In the time domain, the BS allocates each scheduling frame to only one user $u \in \mathbf{U}$. The obtainable data rate of user u in the t-th frame, denoted as $r_u(t)$, can be approximated with the CSI feedback and serves as one essential factor for user scheduling.

As a major factor that determines the instantaneous data rate, channel fading is modeled with a few different variables that can be classified into two categories: slow fading and fast fading. Slow fading arises when the coherence time of the channel is large relative to the delay requirement of the application [26]. It is caused by propagation attenuation and shadowing. For instance, the users have better channel qualities on average if they are close to the BS, while the cell-edge users suffer from weaker received signals and stronger inter-cell interference. Fast fading occurs when the coherence time of the channel is smaller than the period of transmission. It leads to time-varying link capacities and consequently fluctuation of $r_u(t)$ in different scheduling frames.

2.1.2 User Traffic Flows

The various rapidly emerging mobile applications bring many different types of user data traffic flows. They can be generally classified into two groups: the saturated traffic flow and the bursty traffic flow. The first group means that a user always has data to be transmitted. Thus, it is also referred to as the full buffer traffic flow because the data buffer of each user is nonempty during a given scheduling period. This traffic flow model is applied to the scenario where a user is using an application that continuously generates data to transmit. Normally, the generated data rate is larger than the transmission capacity that the radio access network offers. Thus, the wireless system is considered to be saturated. In this case, the task of the dynamic resource allocation scheme is to take advantage of the opportunistic scheduling for throughput enhancement. In a multi-user access system, user fairness needs to be taken into account for keeping a balance among the multiple user traffic flows. The saturated traffic flow model can be utilized to evaluate the maximum system performance under the full load condition.

In contrast to the saturated one, the bursty traffic flow means that users may not always have data to be transmitted and the transmission status is dynamically changing due to the user application behaviors. One typical bursty traffic model is the on-off bursty traffic flow which is developed for elaborating the streaming services at the session level. A user has data to be transmitted only during the random *on* periods. In this traffic model, it is desired to improve the data rates for the user links when they are active. Another typical class of the bursty traffic flow is modeled with the packet queuing system. In the BS, there is a packet scheduler and multiple packet queues, each of which buffers the randomly arriving user data packets. The wireless links act as servers for the queuing system and their processing speeds are determined by the link capacities and the scheduling scheme. The packet queuing model can be applied to describe the packet-level processes of some application types, which are closely related to user behaviors, such as web browsing and instant message services. In this model, the main performance index is the average transmission delay of the user data packets.

2.1.3 Scheduling Objectives and Strategies

In the multi-user access systems, the scheduling strategy plays a key role in the transmission performance, including instant and long-term throughput, end-to-end delay, user fairness, etc. We introduce several widely applied scheduling objectives and strategies in this subsection.

2.1.3.1 Round-Robin and Random Allocation

The simplest scheduling method is round-robin (RR), i.e., the radio resources are equally allocated to the users in turn. Therefore, it is fair for every user from the perspective of the obtained resource amount. One similar approach is random

allocation (RA), which allocates radio resources randomly to the users. In the long run, it has the equivalent scheduling result as RR if all users have equal probabilities to be scheduled. Both of the RR and RA schemes are easy to be implemented but offers rare improvement in the system performance since the real-time channel states are not taken into account.

2.1.3.2 Max-Sum Rate

As we discussed, the time-varying feature of the wireless channels can be exploited by the opportunistic scheduling for performance enhancement. Several common single-objective scheduling schemes are introduced as follows. One greedy strategy for the aggregate throughput improvement is max-sum rate (MSR). In every scheduling frame, the BS estimates the link capacity of each user according to the reported CSI. Then, the radio resource is allocated to the user that is able to obtain the maximum volume of transmitted data. In this way, the overall throughput of the system can be maximized. However, the cell-edge users are less likely to be scheduled due to their relatively poor channel qualities than the users located near the BS. Thus, MSR scheduling leads to serious unfairness among users since the users with poor channel conditions are starved of services.

2.1.3.3 Max-Min Rate

In order to guarantee strict fairness among users, max-min rate (MMR) can be adopted as the scheduling objective. It utilizes the historical records of the obtained user data rates for scheduling. Specifically, the user with the lowest long-term averaged data rate is scheduled in each scheduling frame in order to raise its obtained data rate to the average level in the network. Therefore, every user in the system obtains approximately equal transmission performance. However, the price of maximizing user fairness is that more radio resources are allocated to the low-SINR users for improving their throughput without considering the time-varying channel states. Hence, the MMR scheme reduces the overall transmission efficiency. In summary, it is always a tradeoff between the aggregate throughput and user fairness in the design of scheduling schemes.

2.1.3.4 Weighted-Sum Rate

To provide flexible control of the tradeoff between the overall efficiency and user fairness, the weighted-sum rate (WSR) can be utilized as the scheduling metric. Each user has a weight factor that is denoted as ω_u. In each scheduling frame, the user with the maximum weighted data rate is scheduled. To guarantee fairness, the users with low mean SINRs are normally given higher priorities, i.e., larger weight factors. Thus, a balance between the overall throughput and user fairness is achievable with appropriate control of the weight factors assigned to different users.

Similar to the MSR strategy, the basic WSR with fixed weight factors does not consider the historical data rate obtained per user.

An improved WSR variant is given as follows. The scheduling factor of each user is designed as

$$\rho_u(t) = r_u(t) R_u^{-\alpha}(t), \quad u \in \mathbf{U}. \tag{2.2}$$

where $R_u(t)$ is the long-term averaged data rate of user u. It is adopted as the weight factor with a predefined coefficient $\alpha \geq 0$. In particular, the scheduling factor includes only the instantaneous obtainable data rate when $\alpha = 0$, which is identical with the MSR target. When $\alpha \gg 1$, the obtained average throughput dominates the scheduling factor and makes it the same as the MMR target.

2.1.3.5 Proportional Fair Scheduling

To maintain a good balance between the above two contradictory objectives, proportional fairness (PF) has been proposed and become a widely accepted scheduling metric [27]. It is a special case of the WSR variant given in (2.2) with $\alpha = 1$. The long-term averaged rate can be calculated by the exponential-moving-average (EMA) method as follows,

$$R_u(t+1) = \left(1 - \frac{1}{\tau}\right) R_u(t) + \frac{r_u(t)}{\tau}, \quad u \in \mathbf{U}. \tag{2.3}$$

where τ is the averaging coefficient and normally a large positive integer. The EMA approach gives higher weights to the recently obtained data rates. Thus, the recently scheduled users have larger EMA data rates and lower weight factors than their average levels. In this way, no user can monopolize the radio resources for a long time and consequently user fairness can be guaranteed. Proportional fair scheduling (PFS) is proved to be capable of maximizing the logarithmic sum of the EMA data rates, equally, their geometric mean [28].

2.1.3.6 Other Scheduling Objectives

To provide stable and high-quality mobile services, some advanced scheduling schemes have been proposed by cross-layer designing, in which the specific upper layer performance requirements are taken into account [29]. For instance, the throughput outage metric is used for the streaming applications that demand a certain level of transmission data rate. In the real-time services, such as the voice over IP (VoIP) service and monitoring systems, it is necessary to maintain the end-to-end delay under a given threshold. In the mobile applications with multiple performance indices, e.g., mobile games, multi-objective optimization needs to be implemented for improving the quality of user experience [30].

2.2 Data Rate Analysis and Estimation

Corresponding to our discussion in Section 2.1, we introduce research works on the modeling and analysis of dynamic resource allocation schemes as in the following three aspects.

2.2.1 Stochastic Channel Modeling

The stochastic channel modeling is essential for performance analysis of various scheduling schemes since the random process of channel states determines the link capacities for data transmission. The stochastic channel models are developed according to the communication scenarios. In downlink cellular networks, they can be broadly classified into two categories: the single-cell and multi-cell scenarios. In the single-cell case, only one BS is deployed for serving multiple associated users. Nevertheless, the single-cell model is also applicable to the case where the neighbor BSs use orthogonal radio resources. Therefore, no inter-cell interference is involved and the user channels are noise-limited. In this case, the channel qualities are measured in terms of signal-to-noise ratio (SNR). In general, the power of the received noise is modeled with the additive white Gaussian noise and is a time-invariant value. Thus, the SNRs of the received signals fluctuate due to the time-varying channel gains, which can be modeled with the Rayleigh fading in the non-line-of-sight (NLOS) propagation environment, and the Rician or Nakagami-m fading in the line-of-sight (LOS) propagation environment [31].

Due to the densification of BS deployment in the future wireless networks, the inter-cell interference is inevitable for the sake of a high special reuse factor. Therefore, it is significant to build stochastic channel models for multi-cell scenarios. The user channels are interference-limited due to the received noise is normally much weaker than the inter-cell interference signals. Particularly, the power of the interference signals received from the neighbor cells is comparable to the useful signal power for the users located near the cell edge. Thus, the channel models in the single-cell scenarios are not applicable to the multi-cell (or multi-interference) scenarios. In contrast to the single-cell case, the fluctuation of the channel states is attributed also to the time-varying channel gains of interference signals. Some simplified models have been developed for the multi-interference channels, such as the interference as noise (IaN) model in [3, 4], the single-interference model in [5, 6], and the symmetric interference model in [7]. However, the precise stochastic channel model for the general multi-cell scenario is desired but has not been developed so far.

Based on the stochastic channel models, the spectrum efficiency can be derived for user link capacity modeling. One simple way is to use the mapping function of the Shannon capacity [31]. This approach provides the upper bound of the instantaneous obtainable data rate based on the user SINR. Nevertheless, it is a close approximation model while the finite modulation and coding scheme (MCS) set has fine granularity [32]. Besides the SINR-based link capacity modeling, some works

rely on the Gaussian approximation (GA) to formulate the random distribution of the instantaneous user data rate under Rayleigh fading [33, 34]. This model makes the subsequent work on performance analysis easier by formulating the multiple user channel capacities with the same form, i.e., the Gaussian distribution. It has been verified to be accurate but only for the single-cell scenario, i.e., the noise-limited case. In [35], the finite-state Markov channel (FSMC) model has been proposed to formulate the discrete user data rate set. It provides more details for describing the time-correlated channel states but requires much more information of the channel state transition.

2.2.2 Ergodic Data Rate

The ergodic data rate per user can be computed based on the stochastic channel model under some dynamic resource allocation schemes. It is the expectation of the actually obtained user data rate and equals to the mean data rate during a long enough running period. It is worth noting that the ergodic data rate is different from the long-termed averaged data rate used in the scheduling factor. In this thesis, the latter one is dynamically calculated and gives higher weights to the recently obtained data rate values by the EMA method, as shown in (2.3). However, when the averaging coefficient τ is set to a very large number, the EMA data rate approximately equals to the ergodic one and fluctuates around it under steady states [36].

We consider the saturated data traffic model at first for throughput performance analysis under different scheduling schemes. Since the users always have data to be transmitted, the set of active users in the system is fixed. The performance analysis methods of some classic scheduling schemes are briefly introduced in the following part.

2.2.2.1 RR and RA

The RR and RA schemes use neither the channel state nor historical scheduling information. As we discussed in Section 2.1.3, they offer an equal amount of radio resources to each user in the long run. In other words, the users are scheduled with the same chance. Thus, every user in the system obtains $1/U$ of the whole system bandwidth. It is easy to calculate the ergodic data rate per user with their stochastic channel models and the channel capacity mapping function.

2.2.2.2 MMR

The MMR scheme always allocates the radio resources to the user with the lowest long-term averaged data rate. This strategy results in an approximately equal averaged data rate for every user in the system. Therefore, the amount of the radio resources that a user obtains, equivalently, its scheduling probability, is in inverse

proportion to its mean channel capacity. Considering the limited system bandwidth, the scheduled probability per user can be calculated based on their channel models. Thus, the overall throughput is the sum of U approximately equal ergodic user data rates.

2.2.2.3 MSR and WSR

In the MSR scheme, only the user with the largest instantaneous channel capacity is scheduled in each frame. Thus, the conditional probability of a user to be scheduled with a given channel state is the probability that its channel state is better than all of the other users in the system. If the discrete channel states are utilized, e.g., in the FSMC model, the ergodic data rate is the sum of link capacities weighted by the corresponding scheduling probabilities. While using the continuous capacity functions, such as the Shannon capacity, the expectation of user data rate can be calculated by integration. In the WSR scheduling scheme, the users are assigned with different weight factors for their channel capacities. Similar to the MSR case, the scheduled probability of a user can be calculated based on the event that its instantaneous weighted link capacity is the largest.

2.2.2.4 PFS

In contrast to the above schemes, the performance analysis of the PFS scheme is more complicated. This is due to the correlation of its scheduling behavior and results. Specifically, PFS considers the EMA data rates as the weight factors for scheduling. On the other hand, the EMA data rates are calculated according to the scheduling results. In the literature, there have been some efforts to analyze the throughput performance of PFS with simplified models. For instance, the noise-limited systems are considered for single-cell scenarios in [37–39]. In [36, 40–43], the GA method is used for symmetric modeling of user link capacities. In [44], the IaN model is utilized to simplify the stochastic channel model in the multi-interference scenario. In this way, the instantaneous user SINR obeys the exponential distribution under Rayleigh fading as in the noise-limited case. All of the above analytical works have simplified the stochastic channel models by introducing symmetry for the user channels. This simplification makes the analytical performance of PFS tractable, however, leads to inaccurate results in multi-cell networks.

2.2.3 Traffic Flow Models

Under the saturated traffic flows, the active user set is fixed. In contrast, the users can be scheduled only when they have data to be transmitted under the bursty traffic flows. The dynamic change of the active user set brings new challenges to the performance analysis.

Under the bursty on-off traffic flows, the state switching of a user data flow is independent and irrelevant to the channel fluctuation. Thus, the random process of

the active user set can be derived by combining the multiple independent on-off processes of user data flows. For one certain active user set, the throughput performance can be calculated approximately by the analytical result in the saturated traffic case if the change rate of the active user set is much lower than the scheduling rate. Then, by computing the data rates in all possible active user sets and their corresponding probabilities, we can obtain the ergodic results. However, the number of user combinations increases exponentially with the user amount, resulting in high computational complexity. Therefore, the symmetry of user channel features is assumed for simplifying the computation, e.g., the GA model. Specifically, the impacts of different active user sets on transmission performance lie only in the user set sizes. In Chapter 4, this simplified model is utilized for the performance analysis of PFS under on-off bursty traffic flows.

Due to the more complex scheduling behavior in the packet queuing system, it is more difficult to analyze its throughput or delay performance. In contrast to the on-off bursty traffic, the time-varying channels influence not only user scheduling but also the queue status. For instance, a user with a better channel state may have a higher data rate and thus gets its packet queue cleared faster. Then, it is inactive and is not scheduled until the arrival of new packets in its queue. Therefore, the transmission performance is closely correlated to the queuing process.

When simple resource allocation schemes are used, such as RR and RA, the active users share the radio resources equally regardless of their channel qualities. In this case, the processor-sharing model can be used for performance analysis [45]. However, the analytical performance is intractable for the opportunistic scheduling schemes, such as MSR or PFS. Again, by introducing symmetry of user channels, the throughput performance can be approximated. In [46], a linear mapping function is assumed for calculating user link capacities under the low-SNR and noise-limited conditions. Based on this assumption, the analytical performance of the MSR scheduling scheme is derived with the multi-class processor-sharing model. In [47], the GA method is utilized for simplifying the analytical model of PFS in queuing systems. A more sophisticated analytical work has been presented in [35] where the stochastic network calculus is utilized for deriving the upper bound performance of MSR. In summary, accurate analytical models are unobtainable for advanced scheduling strategies under packet queuing traffic flows. The research in this direction is not involved in this thesis work.

2.3 Non-Orthogonal Multiple Access

In order to meet the increasing demand for high-speed data services and massive connectivity, the upcoming 5G networks have to deploy novel and highly efficient technologies [48]. A number of candidates have been proposed to address the challenges in 5G, such as network densification, massive multiple-input-multiple-output (MIMO), full duplex transmissions, millimeter wave (mmWave) communications, device-to-device (D2D) networks, NOMA, etc [49, 50]. Among these technologies, NOMA provides a new feature for multi-user access

and transmission. It supports multi-user superposition transmission (MUST) that has been taken into account in 5G by 3GPP [51, 52]. In contrast to the OMA systems, NOMA allows multi-user signal superposition and uses the multi-user detection technique to eliminate the inter-user interference [53, 54]. So far, there have been various NOMA schemes proposed in the literature, among which two major categories are code-domain NOMA (CD-NOMA) and power-domain NOMA (PD-NOMA).

Similar to CDMA, CD-NOMA utilizes user-specific spreading sequences for the code-domain multiplexing. Depending on the coding schemes, CD-NOMA can be further divided into several different classes, e.g., low-density spreading CDMA (LDS-CDMA), low-density spreading based OFDM (LDS-OFDM), sparse code multiple access (SCMA), multi-user shared access (MUSA), etc [55]. In LDS-CDMA, low-density (or sparse) spreading sequences are used instead of the dense ones in the conventional CDMA with the aim of limiting the impact of interference on each chip [56,57]. Thus, interference can be efficiently reduced among the multiplexed users by appropriately designed spreading sequences. At the receiver side, the message passing algorithm can be used for multi-user detection [58]. In LDS-OFDM, the data symbols are first spread across LDS sequences and then are transmitted on multiple OFDM subchannels [59, 60]. Thus, LDS-OFDM combines the features of LDS-CDMA and OFDM. To improve the spectrum efficiency, the number of carried symbols is allowed to be larger than the number of subchannels so that overloading is achievable. SCMA is an enhanced version of LDS-CDMA. In contrast to LDS-CDMA, the information bits are directly mapped into sparse codewords in SCMA [61, 62]. It provides a low-complexity receiving process and offers improved performance in comparison to LDS-CMDA.

On the other hand, PD-NOMA realizes the power-domain multiplexing by using successive interference cancellation (SIC). Due to the near-far effect, user channel states may differ greatly. Multiple user signals can be overlapped by superposition coding at the transmitter side. According to their instantaneous channel states, they are assigned different power levels. At the receiver side, SIC is used to decode the superposed information for multi-user detection. In this way, PD-NOMA provides access to the diversity gain in a new degree of freedom, namely, the power domain.

Apart from the above two major categories, a few other multiple access schemes are also closely-related to NOMA, such as pattern-division multiple access (PDMA) [63], bit-division multiplexing (BDM) [64], spatial division multiple access (SDMA) [65, 66], trellis-coded multiple access (TCMA) [67], and interleave division multiple access (IDMA) [68, 69].

Due to its simple mechanism and high feasibility in practice, PD-NOMA has attracted considerable research interests recently. Therefore, in this thesis, we focus on the PD-NOMA system and use the abbreviation NOMA hereinafter to refer to it. In the following part, the basic theory of SIC decoding in NOMA systems is described. Then, we briefly review the current research status of the DCPA schemes for NOMA systems.

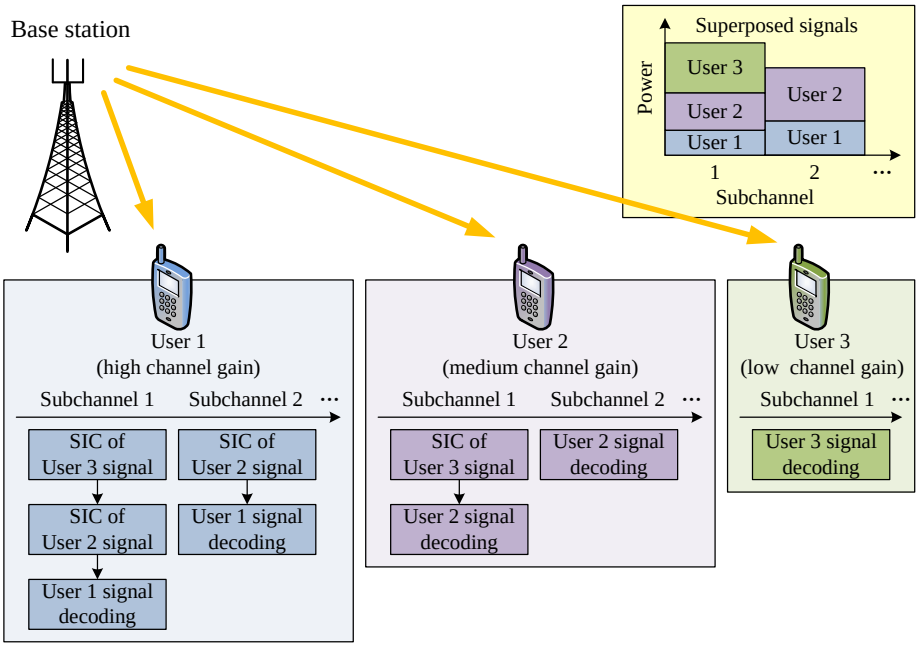


Figure 2.2: Illustration of the SIC decoding process in a downlink MC-NOMA system.

2.3.1 Fundamentals of SIC and NOMA

NOMA allows multiple users to be multiplexed within one channel by employing SIC receivers [70, 71]. In a cellular network, users have different distances from the associated BS. This means that their channel gains can be significantly different from each other. In general, users near the BS have better channel states, namely, higher channel gains, than the cell-edge users. This near-far effect can be utilized by NOMA for the multiplexing users that have different channel gains.

In a downlink NOMA system, the signals of the multiplexed users are superposed and broadcasted by the BS. A near user decodes the data of the users with worse channel states and then performs SIC to eliminate the inter-user interference. In order to achieve the highest decoding performance, SIC should be carried out in ascending order of user channel gains [31]. A far user regards the signals of the near users as noise during its decoding process. In an uplink NOMA system, the SIC decoding process is carried out by the receiver at the BS side in decreasing order of user channel gains.

Figure 2.2 illustrates the receiving and decoding process in a downlink multi-channel NOMA (MC-NOMA) system. On subchannel 1, three users are multiplexed. User 1 has the best channel state thus it firstly decodes and cancels the interference signals of user 3 and user 2 in turn. Then, it decodes the desired signal of its own. User 2 only needs to carry out SIC to eliminate the interference from the signal of user 3. No SIC is necessary for user 3 which has the lowest channel gain. The interference signals of user 1 and 2 are regarded as pure noise during its decoding. On subchannel 2, only two users are multiplexed, and user 1 with a higher channel gain needs to perform

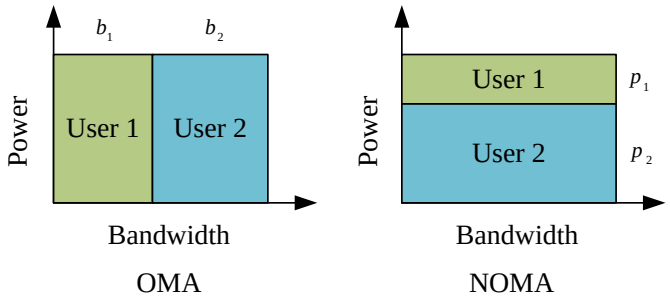


Figure 2.3: Bandwidth and power allocation in OMA and NOMA systems.

SIC.

To clarify the superiority of NOMA over OMA from the theoretic perspective, we present a 2-user SIC example as follows. Consider a BS transmitting to two user receivers with a single-antenna transmitter under the Gaussian channels. The two user have received SNRs as $\gamma_1 = 20$ dB and $\gamma_2 = 0$ dB with normalized transmit power $p = 1$.

We assume that FDMA is adopted in the OMA case as shown in Figure 2.3. The normalized bandwidths allocated to the two users are denoted as b_1 and b_2, which satisfy that $b_1 + b_2 = 1$. Therefore, the obtainable link capacities of the two users are calculated as follows.

$$r_1 = b_1 \log_2 (1 + \gamma_1) , \tag{2.4}$$

$$r_2 = b_2 \log_2 (1 + \gamma_2) , \tag{2.5}$$

where the link capacities of the two users are controlled by the bandwidths allocated to them.

As shown in Figure 2.3, the two users in the NOMA case are assigned power p_1 and p_2, respectively. By using SIC, the post-processing SNR of user 1 is $p_1\gamma$ and its link capacity is

$$r_1 = \log_2 (1 + p_1\gamma_1) . \tag{2.6}$$

User 2 regards the signal of user 1 as noise during its decoding process. Thus, its link capacity is calculated as

$$r_2 = \log_2 \left(1 + \frac{p_2\gamma_2}{p_1\gamma_2 + 1} \right) . \tag{2.7}$$

We present the capacity regions of the OMA and NOMA systems in Figure 2.4. The performance of NOMA is strictly better than that of OMA except for the cases where only one user occupies the bandwidth and transmit power. This means that for any rate pair obtained by the OMA scheme, there exists a power split resulting in a larger rate pair. This performance improvement increases with the gap between the user channel gains [31].

To implement NOMA, assistant information is required for SIC decoding at

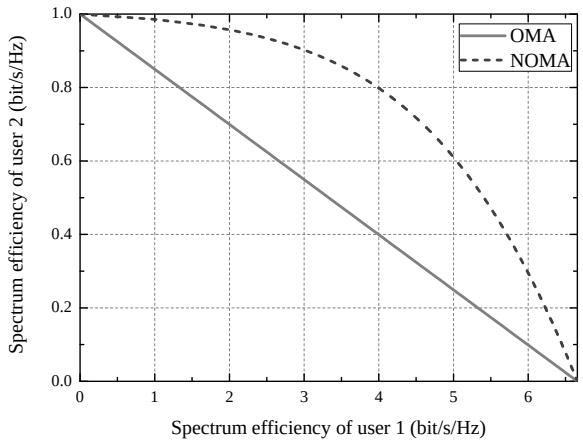


Figure 2.4: The link capacity regions of OMA and NOMA systems.

user receivers. Besides the blind detection, additional signalling provides more reliable information thus enhances the decoding success ratio with a price of increased overhead. Several crucial parameters need to be included in the assistant information. Firstly, it is necessary to inform the SIC receiver about the number of inter-user interference signals that it is required to process and cancel. In particular, this can be indicated by only one bit in a 2-user NOMA system while more bits are used if the multiplexed user number is larger. Besides, according to the SIC order, the corresponding MCS and power allocation information needs to be attached as well.

In addition, nonlinear detection is favorable to SIC, e.g., the maximum-likelihood receiver, for reliable decoding and avoiding error propagation. However, this leads to extra processing complexity and latency. When there are more users multiplexed by NOMA, the number of SIC stages required for decoding increases. Therefore, considering the processing complexity for SIC and signalling overhead, the number of multiplexed users per channel is limited by a predefined integer parameter, which is denoted as S. On the other hand, the multiplexed user with the lowest channel gain needs no SIC. Thus, when its signal has a large power ratio or a low-rate MCS, linear detection can be utilized, such as the minimum-mean-square-error receiver.

2.3.2 Channel and Power Allocation Schemes

As shown in Section 2.3.1, the user data rates are controlled by their allocated power in the NOMA system. Therefore, besides the channel allocation problem, multi-user power allocation plays another key role in the performance of NOMA systems and consequently attracts increasing research interests. In the following part, we introduce the current research progress of the DCPA schemes in various NOMA systems and application scenarios.

2.3.2.1 DCPA Schemes for Downlink Single-Channel NOMA

In downlink single-channel NOMA (SC-NOMA) systems, the DCPA problem can be decoupled into two stages, namely, power allocation and channel allocation. Firstly, the power allocation problem is solved with a given strategy for each candidate multiplexed user set on the channel. Then, the optimal user set is selected and the channel is assigned to the corresponding users with the optimal power allocation. Hence, the channel allocation problem in the second stage is also referred to as user set selection (USS) or user clustering in the literature. Note that the optimality can be guaranteed for a given objective with this decoupled method in SC-NOMA systems. For instance, the full searching-based power allocation (FSPA) for each candidate user set and the full user set comparison (FUSC) for USS can be adopted as straightforward solutions in the two stages [10, 70, 72, 73]. In the following part, these two stages are discussed respectively.

The simplest approach to assigning power to the multiplexed users is the fixed power allocation (FPA) scheme, in which the power ratio between the far and near users is a constant [70, 72, 74]. The advantage of this scheme lies in the minimum computational complexity for power allocation. However, since the diverse and fluctuating user channels are not fully exploited, the throughput performance and user fairness are far from the optimal. Therefore, dynamic power allocation is desired and has been investigated in recent research.

The fractional transmit power allocation (FTPA) scheme is a straightforward improvement of FPA by utilizing the real-time CSI [70, 72, 74]. The power ratio assigned to a user is in inverse proportion to its channel gain. Thus, the cell-edge users obtain more power so that user fairness is improved. The FTPA scheme is more flexible than FPA and consequently achieves better performance with the extra requirement of CSI feedback.

Furthermore, several specific optimization objectives have been considered in the design of dynamic power allocation, such as MSR, MMR, WSR, and PF. Notably, with the normal MSR objective, power and channel sources are all allocated to the user with the best channel state, leading to the identical results with the MSR scheduling scheme in the OMA system. Therefore, the cell-edge users are starved of service [75]. To avoid this, additional conditions are required to guarantee user fairness in the MSR-oriented power allocation schemes, for instance, the maximum assigned power or the minimum obtained data rate per user [75–77]. In contrast to MSR, the power allocation schemes with the MMR target aim at maximizing user fairness [78]. In order to achieve a good balance between efficiency and fairness, WSR and PF have been widely adopted as the optimization target in the latest research [10, 11, 15, 79, 80].

In the first stage of the decoupled DCPA problem, the power allocation is optimized for each candidate user set. Due to the finite MCS selection space in practical systems, the mapping function between the SINR and the obtainable data rate is non-smooth. However, with fine granularities for both power allocation and the MCS set, this constraint can be relaxed in the optimization problem. In other words,

continuous power levels and Shannon capacity mapping function can be used as a close approximation. In the literature, there have been various approaches proposed to solve the power allocation problem. We classify them into the following three categories:

- ▷ Searching algorithms
  In practice, the assigned power value can be quantized according to a predefined granularity. Thus, the optimal allocated power is solved by searching the discretized power levels. A straightforward method is to use FSPA [70,72]. However, this method has an extremely high computational cost that is unacceptable for practical applications. In order to reduce searching complexity, a tree-search based transmission power allocation (TTPA) algorithm was designed with the PF objective in [11].

- ▷ Iterative algorithms
  With continuous power variables, the optimal power allocation is achievable with the iterative algorithms, such as the iterative linear programming proposed in [78] for MMR, the alternating maximization algorithm proposed in [76] for MSR, and the iterative water-filling algorithm proposed in [10] for PF. However, the defect of iterative approaches is that the convergence process depends on the actual parameters and can be highly time-consuming in practice.

- ▷ Closed-form solutions
  Although both of the searching and iterative algorithms provide access to the optimized solutions, they still suffer from such high computational complexity that they can hardly be applied to dynamic power allocation. Thus, the solutions in concise forms are more favorable in practice. To this end, we have developed the closed-form solution for PFS-oriented power allocation in [20,21]. In [75,81], the optimal allocated power for MSR has been solved in a closed form with pre-defined user clusters.

In the first stage, the power allocation problem is solved for each candidate multiplexed user set. Then, the second stage in the decoupled DCPA problem is selecting the optimal multiplexed user set according to the optimized power results. Note that the number of multiplexed users in each set is no more than S due to the limitation on SIC decoding as we discussed in Section 2.3.1. A straightforward way is to compare all possible user sets, i.e., FUSC [70, 72]. However, this USS method results in a rapidly increasing complexity with the number of users.

With the aim of reducing USS complexity, some candidate user sets can be omitted for comparison. For example, some user sets may not satisfy the scheduling conditions and cannot be selected with a given optimization objective. Hence, it is unnecessary to carry out the optimization of power allocation in the first stage for these user sets, and the computational complexity can be reduced greatly for the subsequent USS. In [79] and [15], several predefined conditions are utilized for PFS to exclude some invalid candidate user sets. A greedy algorithm is proposed to reduce

the USS complexity in [80]. However, these USS schemes yield only suboptimal solutions.

2.3.2.2 DCPA Schemes for Downlink Multi-Channel NOMA

Due to time-varying channel states in the frequency domain, dynamic inter-channel power allocation is necessary to guarantee transmission stability and to improve spectrum efficiency in MC-NOMA systems. Therefore, besides the multi-user power allocation problem, it is necessary to control the power assigned to different subchannels. The former problem is referred to as the intra-channel power allocation while the latter one is referred to as the inter-channel power allocation in this thesis. Because of the changeable power on each subchannel, the channel and power allocation problems are strongly correlated to each other [13]. Hence, the DCPA problem in MC-NOMA systems is more complex than in the SC-NOMA case.

There are many works making efforts to improve the performance of MC-NOMA. Various optimization algorithms have been adopted for power allocation and can be classified into two main types: iteration-based algorithms [12, 82–86] and searching-based algorithms [13,87]. In [88], the optimal solutions to the intra-channel power allocation were derived with the MMR, MSR and WSR objectives in closed-forms. To solve the channel allocation problem for MC-NOMA systems, matching algorithms were utilized in [12, 88, 89]. However, they provide only suboptimal results and cost very high computational complexity. In [82] and [14], greedy user selection schemes were proposed with the aim of reducing the complexity for channel allocation. Besides the DCPA schemes that focus on improving user data rates, energy efficiency (EE) has been adopted as another optimization objective for MC-NOMA systems in some latest research [88, 90, 91]. Again, the iteration-based and matching algorithms were used for improving the EE performance.

2.3.2.3 DCPA Schemes for Uplink NOMA

In the uplink NOMA systems, SIC is carried out at the BS side. The conventional uplink transmit power control in the OMA systems intends to equalize the received power from all users. On the contrary, NOMA systems require distinctness among the superposed user signals. In comparison to the downlink scenario, the DCPA problem in the uplink NOMA systems has an additional constraint on the transmit power for every user due to the limited capacities of their batteries. Specifically, the sum power transmitted by a user terminal on one channel or multiple subchannels is limited.

Besides the consideration of limited uplink transmit power, the synchronization problem is more challenging for implementation of the uplink NOMA systems. Due to the near-far effect, the multiplexed users need to transmit their signals with different time advance parameters in order to guarantee the aligned signal superposition at the BS receiver. However, even slight offsets among the received signals can lead to large distortion of the superposed signals and consequent failure

Table 2.1: Comparison of Various DCPA Schemes in NOMA Systems (1/2)

Ref.	**System**		**CPA scheme**			
	DL/ UL	**SC/ MC**	**Obj.**	**PA scheme**	**CA scheme**	**Res.**
[78]	DL	SC	MMR	Iteration based on linear programming	All users are multiplexed	Opt.
[75]	DL	SC	MSR	Closed-form solution	Fixed user pair	Sub.
[77]	DL	SC	MSR	Iterative minorization-maximization algorithm (MMA)		Sub.
[76]	DL	SC	MSR	Alternating maximization (AM) algorithm	Fixed user pair	Opt.
[70, 72, 74]	DL	SC	PF	FSPA	FUSC	Opt.
				Fixed PA		Sub.
				FTPA		Sub.
[11, 71]	DL	SC	PF	Tree-search based transmission PA (TTPA)	FUSC	Opt.
[73]	DL	SC	PF	Fixed PA	FUSC	Sub.
[10]	DL	SC	PF	Iterative water-filling	FUSC	Opt.
[15]	DL	SC	PF	Closed-form solution	Partial user set comparison	Sub.
[79]	DL	SC	PF	Closed-form solution	Conditional user set comparison	Sub.
[80]	DL	SC	PF	Closed-form solution	Greedy algorithm	Sub.
[92, 93]	DL	SC	Alpha fairness	Iterative algorithm	All users are multiplexed	Opt.
[87]	DL	MC	MSR	Lagrangian duality and dynamic programming (LDDP)		Sub.
[13]	DL	MC	WSR	LDDP	Iterative algorithm	Sub.

Abbreviation - **Ref.**: reference; **DL**: downlink; **UL**: uplink; **Obj.**: objective; **Res.**: result; Opt.: optimal; Sub.: suboptimal.

Table 2.2: Comparison of Various DCPA Schemes in NOMA Systems (2/2)

Ref.	System		CPA scheme			
	DL/ UL	**SC/ MC**	**Obj.**	**PA scheme**	**CA scheme**	**Res.**
[12]	DL	MC	WSR	Geometric programming	Matching algorithm	Sub.
[82]	DL	MC	WSR	Difference of convex (DC) programming	Greedy user selection	Sub.
[83, 84]	DL	MC	WSR	Outer polyblock approximation algorithm		Opt.
[85]	DL	MC	PF	FTPA and water-filling	FUSC	Sub.
[14]	DL	MC	PF	FTPA	Greedy algorithm	Sub.
[86]	DL	MC	Alpha fairness	Sequential convex programming	Matching algorithm	Sub.
[88, 89]	DL	MC	MMR	Closed-form solution	Matching algorithm	Sub.
			MSR	Closed-form solution		Sub.
			WSR	Closed-form solution		Sub.
			EE	Iterative algorithm		Sub.
[90]	DL	MC	EE	DC programming	Matching algorithm	Sub.
[91]	DL	MC	EE	Convex programming	Iterative algorithm	Sub.
				Branch-and-bound approach		Opt.
[94]	UL	SC	PF	FSPA	FUSC	Sub.
[95]	UL	MC	MSR	Sequential convex programming	Matching algorithm	Sub.
[59]	UL	MC	MSR	Water-filling algorithm	Iterative algorithm	Sub.
[81]	DL UL	SC	MSR	Closed-form solution	Pre-defined user clustering	Sub.

Abbreviation - **Ref.**: reference; **DL**: downlink; **UL**: uplink; **Obj.**: objective; **Res.**: result; Opt.: optimal; Sub.: suboptimal.

in SIC. Therefore, the DCPA scheme for uplink NOMA transmission has been less studied as for the downlink case. A few research works on the uplink DCPA scheme mainly focus on improving the throughput performance but obtain only suboptimal results with high-complexity algorithms [59, 81, 94, 95].

2.3.2.4 Summary of DCPA Schemes

For a simple and straightforward comparison of various DCPA schemes that have been proposed for NOMA systems, we survey the existing research works and list their features and solutions as in Table 2.1 and Table 2.2. It is clear that the optimal DCPA solutions can be obtained only by high-complexity approaches, such as iterative or searching algorithms. Moreover, low-complexity schemes result only in suboptimal solutions. To the best of our knowledge, the optimal solution for DCPA in NOMA systems has not been developed in a concise form so far. Thus, it is still an open research question to be studied.

2.3.3 Performance Analysis of DCPA Schemes

In the literature, the performance of many DCPA schemes for NOMA has been evaluated by simulation approaches. In order to further investigate and better utilize the DCPA schemes, the analytical solutions to their performance are desired. In the downlink single-cell network, the outage probability and ergodic sum-rate have been analyzed based on the FPA scheme in [96–99]. In [100], the fixed power ratio between a pair of multiplexed users is optimized in order to improve the expectation of their ergodic sum rate. To better adapt to the time-varying channels, a series of dynamic fractional power allocation schemes have been proposed and analyzed for the 2-user NOMA systems with given quality of service (QoS) requirements [101–104].

Besides the above analytical works that are based on the Rayleigh fading channel and downlink single-cell networks, some other researches study the performance of NOMA systems in different scenarios. For instance, the relaying networks have been investigated in [105–107] and the Nakagami fading channel has been considered in [106, 108, 109] for throughput analysis of NOMA systems. In particular, the spectrum efficiency of the uplink NOMA transmission was analyzed in [108].

Although there have been some analytical outcomes, the performance of DCPA schemes has not been studied for NOMA sufficiently. For example, the performance of the DCPA schemes with the WSR or PF targets has not been analyzed although they have been widely accepted as the optimization objective. The main difficulty in the performance analysis lies in the intractability of the optimal DCPA scheme and the strong correlation between the dynamic scheduling behaviors and time-varying channel states. The performance analysis of DCPA in NOMA systems is important to provide guidelines for its optimization and application [35]. In particular, the analytical results can be applied to performance prediction and assisting user association, traffic load balancing, radio resource management, etc. Therefore, it is

desired for more research attention on the analytical performance of DCPA schemes in NOMA systems.

3 Data Rate Analysis Based on Stochastic Channel Modeling

In order to meet the demand of enjoying mobile communication services anywhere and anytime, wireless cellular networks are trending towards increasing density, making inter-cell cooperation more significant than ever before. Inter-cell related operations, such as user association, user handover, load balancing, and interference coordination, require an accurate estimation of the transmission performance to make system decisions and actions dynamically. Among others, the inter-cell interference (ICI) is the most crucial impact factor on the system performance in multi-cell wireless networks.

To estimate the transmission performance, precise analytical models of the received signals and link capacities under given resource allocation schemes are prerequisites. In a multi-interference wireless environment, user devices can measure the reference signal received powers (RSRPs) of the serving BS and relatively strong inter-cell interference [110]. The statistical results of the measurements are reported to their associated BSs and can be used for user data rate estimation and assisting system operations. The estimation accuracy is influenced by multiple aspects, such as channel measurement precision, the amount of feedback channel state information (CSI), propagation environment, the analytical models of the resource allocation schemes, etc.

In the literature, there have been several stochastic channel models developed for the multi-cell networks. One simple model is called interference as noise (IaN) [3, 4]. It means that the channel gains of the interference signals are regarded as constant values. Thus, the received interference power is time-invariant as the noise power. However, this approximation is inaccurate, especially for the cell-edge users that are affected more by the time-varying interference signals. In [5] and [6], only one co-channel neighbor cell is considered and the probability distribution of SINR is derived under Rayleigh fading channels. In [7], the channel model is built based on the assumption that the interference signals received from different neighbor cells share the same mean power and the noise is neglected. In order to overcome these restrictions, we are motivated to build generalized and precise stochastic channel models for the multi-cell scenario.

In this chapter, we first describe the system model of a downlink multi-cell wireless network in Section 3.1. In Section 3.2, we develop a stochastic channel model and derive a closed-form probability distribution of the instantaneous user SINR. Considering the limited reported CSI, two additional channel models are formulated and proved to be upper and lower bounds of the actual SINR distribution, respectively. We further extend these two models to a weighted sum SINR model in order to improve the accuracy with partial and imperfect CSI. Based on the stochastic channel models, we analyze the ergodic user data rates under different

scheduling schemes, including MMR, MSR, and PFS, in Section 3.3. The analytical results are utilized for user data rate estimation. We analyze the major factors that influence the estimation results in Section 3.4. Then, we assess the estimation accuracy by simulations in Section 3.5. The relative errors of our estimated user data rates are compared to the existing works in the literature. Finally, this chapter is summarized in Section 3.6.

3.1 System Model

We consider a downlink cellular network containing multiple BSs. We denote the set of the BS indices by

$$\mathbf{B} = \{b \,|b = 1, \ldots, B\}\,. \tag{3.1}$$

Each BS uses a single-antenna transceiver and offers the coverage of one cell. The index set of the user terminals in the network is denoted by

$$\mathbf{U} = \{u \,|u = 1, \ldots, U\}\,. \tag{3.2}$$

Each of them also uses a single-antenna transceiver and is associated to one BS in $\mathbf{B}$.

The frequency band is reused by all cells in the network and is divided into K resource blocks (RBs) in each scheduling frame. The BSs distribute RBs to their connected users according to the scheduling scheme. All of the RBs within a given frequency band are assumed to have independent and identically distributed (i.i.d.) channel gains that are modeled with the Rayleigh fast fading for each user. Without loss of generality, we focus on the channel state and scheduling problem with only one of the RBs in the following part due to their symmetry.

The instantaneous received power of the reference signal (RS) at user u from BS b is modeled as

$$P_{u,b} = p_b L_{u,b} \|h_{u,b}\|^2, \tag{3.3}$$

where p_b is the transmit power of RS from BS b, $L_{u,b}$ is the slow-fading factor, including the path loss and shadow fading between user u and BS b, and $h_{u,b}$ is the instantaneous channel gain of the Rayleigh fading, which is modeled as a circularly symmetric complex Gaussian random variable. Its mean value is 0 and covariance is 1. Thus, the power gain of $\|h_{u,b}\|^2$ is exponentially distributed with a unit mean value. We assume that $L_{u,b}$ keeps steady during a long scheduling period while the user is not moving in a high speed. Therefore, $P_{u,b}$ is modeled as a random variable with the exponential distribution. Its mean value can be estimated by the detected RSRP and is expressed as

$$p_{u,b} = \mathbb{E}\left[P_{u,b}\right] = p_b L_{u,b}. \tag{3.4}$$

This is the average power of the symbols that carry cell-specific RSs over the entire bandwidth. A user reports this averaged value to its serving BS for reporting its channel state. In the long-term evolution (LTE) standard, the range of the reported

Table 3.1: Mapping Indices of the Reported RSRP

Reported index	Measured value (dBm)
RSRP_00	RSRP<-140
RSRP_01	-140≤RSRP<-139
RSRP_02	-139≤RSRP<-138
...	...
RSRP_95	-46≤RSRP<-45
RSRP_96	-45≤RSRP<-44
RSRP_97	-44≤RSRP

RSRP is defined from -140 dBm to -44 dBm with 1 dBm resolution [111]. The mapping indices of the RSRP values are listed in Table 3.1. In the following part, we use the RSRP in the linear form for our derivation.

The total instantaneous power of the RSs received by user u is expressed as

$$P_u = P_{u,b} + P_{u,\mathbf{I}_u} + \sigma_N, \tag{3.5}$$

where $\mathbf{I}_u$ is the interfering BS set of user u, including $I_u = |\mathbf{I}_u|$ independent inter-cell interferers, $P_{u,\mathbf{I}_u}$ is the sum power of the received RSs of the interferers in $\mathbf{I}_u$, and σ_N is the power of additive white Gaussian noise.

We denote the interference RSRP (IRSRP) as $p_{u,i} = \mathbb{E}\left[P_{u,i}\right], i \in \mathbf{I}_u$ and denote the mean value of the total received RS power as $p_u = \mathbb{E}\left[P_u\right]$, which can be calculated with the received signal strength indicator (RSSI) as follows.

$$p_u = \frac{\text{RSSI}}{N_{sc}K}, \tag{3.6}$$

where N_{sc} is the number of physical subcarriers per RB. The RSSI is pure wide-band power measurement, including useful power, interference, and noise over all used subcarriers [111]. We assume that all subcarriers within the frequency band are used by the system.

The instantaneous SINR is expressed as

$$\Phi_u = \frac{P_{u,b}}{P_{u,\mathbf{I}_u} + \sigma_N} = \frac{P_{u,b}}{\sum_{i\in\mathbf{I}_u} P_{u,i} + \sigma_N}. \tag{3.7}$$

User devices can measure and calculate the IRSRPs of the relatively strong interferers and report them to its serving BS via control channels. The maximum number of the reported IRSRPs per user is controlled by a parameter *maxReportCells* [110]. We denote it as an integer I_R. Hence, a user can report no more than I_R IRSRPs to its serving BS. The reported interfering BS set of user u is denoted as $\mathbf{I'}_u$ which includes the indices of the I_R BS that have the largest IRSRPs. The set of other interferers is denoted as $\mathbf{I''}_u = \mathbf{I}_u - \mathbf{I'}_u$, which have relatively lower IRSRPs than those in $\mathbf{I'}_u$.

3.2 Stochastic Channel Models

Based on the statistic channel parameters, we derive the probability distribution of the instantaneous user SINR in the multi-interference environment. In the following derivation, we consider one arbitrary user $u \in \mathbf{U}$ and its serving BS $b \in \mathbf{B}$. For ease of the derivation, we first normalize the reported parameters of user u and its received noise power by the mean power of the useful signal $p_{u,b}$ as follows.

$$\hat{p}_u = \frac{p_u}{p_{u,b}}, \tag{3.8}$$

$$\hat{p}_{u,i} = \frac{p_{u,i}}{p_{u,b}}, \tag{3.9}$$

$$\hat{\sigma}_N = \frac{\sigma_N}{p_{u,b}}, \tag{3.10}$$

$$\hat{p}_{u,b} = 1. \tag{3.11}$$

3.2.1 Probability Distribution of SINR

Due to the Rayleigh fast fading, the instantaneous received RS power from the serving BS obeys exponential distribution. Thus, its probability density function (PDF) is expressed as

$$f_{P_{u,b}}(x) = \frac{1}{\hat{p}_{u,b}} \exp\left(-\frac{x}{\hat{p}_{u,b}}\right) = \exp(-x), \quad x > 0. \tag{3.12}$$

The interference signal power received from each neighbor BS is also exponentially distributed. Their PDFs are given as

$$f_{P_{u,i}}(x) = \frac{1}{\hat{p}_{u,i}} \exp\left(-\frac{x}{\hat{p}_{u,i}}\right), \quad x > 0, \quad i \in \mathbf{I}_u. \tag{3.13}$$

The PDF of the sum power of the independent interference signals is the convolution of their PDFs [112]. It is calculated as

$$f_{P_{u,\mathbf{I}_u}}(y) = \sum_{i \in \mathbf{I}_u} \left[\exp\left(-\frac{y}{\hat{p}_{u,i}}\right) \prod_{j \in \mathbf{I}_u}^{j \neq i} \frac{\hat{p}_{u,i}}{\hat{p}_{u,i} - \hat{p}_{u,j}} \right], \quad y > 0. \tag{3.14}$$

According to (3.7), the PDF of the instantaneous user SINR is derived as follows.

$$\begin{aligned} f_{\Phi_u}(z) &= \int_{\sigma_N}^{\infty} y f_{P_{u,b}}(yz)\, f_{P_{u,\mathbf{I}_u}}(y-\sigma_N)\, dy \\ &= \int_{\sigma_N}^{\infty} \sum_{i\in\mathbf{I}_u} \left[\frac{y}{\hat{p}_{u,i}} \exp\left(-\frac{y-\sigma_N}{\hat{p}_{u,i}} - yz \right) \prod_{j\in\mathbf{I}_u}^{j\neq i} \frac{\hat{p}_{u,i}}{\hat{p}_{u,i}-\hat{p}_{u,j}} \right] dy \\ &= \sum_{i\in\mathbf{I}_u} \left\{ \exp\left(\frac{\sigma_N}{\hat{p}_{u,i}}\right) \prod_{j\in\mathbf{I}_u}^{j\neq i} \frac{\hat{p}_{u,i}}{\hat{p}_{u,i}-\hat{p}_{u,j}} \int_{\sigma_N}^{\infty} \left[\frac{y}{\hat{p}_{u,i}} \exp\left(-\frac{y+yz\hat{p}_{u,i}}{\hat{p}_{u,i}} \right) \right] dy \right\} \\ &= \sum_{i\in\mathbf{I}_u} \left\{ \left[\sigma_N + \frac{\hat{p}_{u,i}}{\hat{p}_{u,i}z+1} \right] \frac{\exp(-z\sigma_N)}{\hat{p}_{u,i}z+1} \prod_{j\in\mathbf{I}_u}^{j\neq i} \frac{\hat{p}_{u,i}}{\hat{p}_{u,i}-\hat{p}_{u,j}} \right\}, \quad z>0. \end{aligned} \tag{3.15}$$

Then, its cumulative distribution function (CDF) is derived as follows.

$$\begin{aligned} F_{\Phi_u}(z) &= \int_0^z f_{\Phi_u}(y)dy \\ &= \sum_{i\in\mathbf{I}_u} \left\{ \prod_{j\in\mathbf{I}_u}^{j\neq i} \frac{\hat{p}_{u,i}}{\hat{p}_{u,i}-\hat{p}_{u,j}} \right\} - \exp(-z\sigma_N) \sum_{i\in\mathbf{I}_u} \left\{ \frac{1}{\hat{p}_{u,i}z+1} \prod_{j\in\mathbf{I}_u}^{j\neq i} \frac{\hat{p}_{u,i}}{\hat{p}_{u,i}-\hat{p}_{u,j}} \right\}, \\ & z>0. \end{aligned} \tag{3.16}$$

To simplify the above PDF and CDF expressions, we utilize Theorem 3.1 and its corollary given as follows.

Theorem 3.1. *If $x \in \mathbb{R}^+$, and $a_i \in \mathbb{R}^+$, $i \in \{1,2,\ldots,N\}$ satisfy that $\forall i \neq j$, $a_i \neq a_j$, then*

$$\prod_{j=1}^{N} \frac{x}{x+a_j} = \sum_{i=1}^{N} \left(\frac{x}{x+a_i} \prod_{j=1,j\neq i}^{N} \frac{a_i}{a_i-a_j} \right). \tag{3.17}$$

Corollary 3.1. *If $a_i \in \mathbb{R}^+$, $i \in \{1,2,\ldots,N\}$ satisfy that $\forall i \neq j$, $a_i \neq a_j$, then*

$$\sum_{i=1}^{N} \left(\prod_{j=1,j\neq i}^{N} \frac{a_i}{a_i-a_j} \right) = 1. \tag{3.18}$$

Theorem 3.1 and Corollary 3.1 are proved as follows.

Proof. Theorem 3.1 is proved with induction:

▷ When $N=1$, it is true that (3.17) holds.

▷ Assuming (3.17) true when $N = M$, i.e.,

$$\prod_{j=1}^{M} \frac{x}{x+a_j} = \sum_{i=1}^{M} \left(\frac{x}{x+a_i} \prod_{j=1, j\neq i}^{M} \frac{a_i}{a_i - a_j} \right), \tag{3.19}$$

the result holds for $N = M + 1$ as shown by the following derivation.

$$\begin{aligned}
&\sum_{i=1}^{M+1} \left(\frac{x}{x+a_i} \prod_{j=1, j\neq i}^{M+1} \frac{a_i}{a_i - a_j} \right) \\
=&\sum_{i=1}^{M} \left(\frac{x}{x+a_i} \frac{a_i}{a_i - a_{M+1}} \prod_{j=1, j\neq i}^{M} \frac{a_i}{a_i - a_j} \right) + \frac{x}{x+a_{M+1}} \prod_{j=1}^{M} \frac{a_{M+1}}{a_{M+1} - a_j} \\
\overset{(3.19)}{=}&\sum_{i=1}^{M} \left[\left(\frac{x}{x+a_i} \frac{a_i}{a_i - a_{M+1}} + \frac{x}{x+a_{M+1}} \frac{a_{M+1}}{a_{M+1} - a_i} \right) \prod_{j=1, j\neq i}^{M} \frac{a_i}{a_i - a_j} \right] \\
=&\frac{x}{x+a_{M+1}} \sum_{i=1}^{M} \left[\frac{x}{(x+a_i)} \prod_{j=1, j\neq i}^{M} \frac{a_i}{a_i - a_j} \right] \\
\overset{(3.19)}{=}&\prod_{j=1}^{M+1} \frac{x}{x+a_j}
\end{aligned} \tag{3.20}$$

Thus, Theorem 3.1 is proved. When $x \to \infty$, it holds that

$$\sum_{i=1}^{N} \left(\prod_{j=1, j\neq i}^{N} \frac{a_i}{a_i - a_j} \right) = 1. \tag{3.21}$$

Therefore, Corollary 3.1 is proved. □

By Theorem 3.1 and Corollary 3.1, we simplify the CDF of SINR in (3.16) into the following form.

$$F_{\Phi_u}(z) = 1 - \exp(-z\hat{\sigma}_N) \prod_{i \in \mathbf{I}_u} (\hat{p}_{u,i} z + 1)^{-1}, \quad z > 0. \tag{3.22}$$

Based on the above CDF solution, the PDF of SINR can be further derived as

$$\begin{aligned}
f_{\Phi_u}(z) &= \exp(-z\hat{\sigma}_N) \left[\hat{\sigma}_N + \sum_{i \in \mathbf{I}_u} \frac{\hat{p}_{u,i}}{(\hat{p}_{u,i} z + 1)} \right] \prod_{j \in \mathbf{I}_u} (\hat{p}_{u,j} z + 1)^{-1} \\
&= [1 - F_{\Phi_u}(z)] \left[\hat{\sigma}_N + \sum_{i \in \mathbf{I}_u} \frac{\hat{p}_{u,i}}{(\hat{p}_{u,i} z + 1)} \right], \quad z > 0.
\end{aligned} \tag{3.23}$$

3.2.2 Upper and Lower Bounds

As we discussed in the system model, the number of reported IRSRPs per user is controlled by the parameter I_R. Therefore, we use partial CSI for estimating the probability distributions of SINRs in this part. We consider the remaining part of the received undesired power, excluding the reported IRSRPs in $\mathbf{I'_u}$, as a whole and denote it as

$$\begin{aligned}\hat{\delta}_u\left(\mathbf{I'}_u\right) &= \hat{\sigma}_N + \hat{p}_{u,\mathbf{I''}_u} \\ &= \hat{p}_u - \hat{p}_{u,b} - \hat{p}_{u,\mathbf{I'}_u}, \end{aligned} \tag{3.24}$$

$$\tag{3.25}$$

where

$$\hat{p}_{u,\mathbf{I'}_u} = \sum_{i\in\mathbf{I'}_u} \hat{p}_{u,i}, \tag{3.26}$$

$$\hat{p}_{u,\mathbf{I''}_u} = \sum_{i\in\mathbf{I''}_u} \hat{p}_{u,i}. \tag{3.27}$$

To obtain $\hat{\delta}_u\left(\mathbf{I'}_u\right)$ of a user, the RSSI is utilized for calculating $\hat{p}_u$ in (3.24), i.e., the mean total received power. We design two analytical SINR models based on the limited reported CSI as follows.

3.2.2.1 SINR Model 1 (S1)

We consider the reported IRSRPs separately. The other undesired power $\hat{\delta}_u\left(\mathbf{I'}_u\right)$ is regarded as pure noise. We denote the instant user SINR based on this assumption as $\Phi_u^{(1)}$. Its CDF can be calculated according to (3.22) as

$$F_{\Phi_u^{(1)}}\left(z,\mathbf{I'}_u\right) = 1 - \exp\left[-z\hat{\delta}_u\left(\mathbf{I'}_u\right)\right] \prod_{i\in\mathbf{I'}_u}\left(\hat{p}_{u,i}z+1\right)^{-1}, \quad z>0. \tag{3.28}$$

This SINR model results in a higher value of the SINR CDF than the actual one, i.e.,

$$F_{\Phi_u^{(1)}}\left(z,\mathbf{I'}_u\right) \geq F_{\Phi_u}\left(z\right). \tag{3.29}$$

Thus, it can be adopted as an upper bound of the actual SINR CDF. We prove (3.29) as follows.

Proof.

$$\begin{aligned} F_{\Phi_u^{(1)}}\left(z, \mathbf{I'}_u\right) &\geq 1 - \exp\left(-z\hat{\sigma}_N\right) \frac{\prod\limits_{i\in\mathbf{I'}_u} \left(\hat{p}_{u,i}z+1\right)^{-1}}{\prod\limits_{i\in\mathbf{I''}_u} \left(\hat{p}_{u,i}z+1\right)} \\ &= 1 - \exp\left(-z\hat{\sigma}_N\right) \prod_{i\in\mathbf{I}_u} \left(\hat{p}_{u,i}z+1\right)^{-1} \\ &= F_{\Phi_u}\left(z\right). \end{aligned}$$

□

3.2.2.2 SINR Model 2 (S2)

We regard $\hat{\delta}_u\left(\mathbf{I'}_u\right)$ as the mean received power of an imaginary interference RS, which also has an exponential distribution of the delivered power as the other interferers. We denote the instantaneous user SINR based on this assumption as $\Phi_u^{(2)}$. According to (3.22), its CDF is given as

$$F_{\Phi_u^{(2)}}\left(z, \mathbf{I'}_u\right) = 1 - \left[\hat{\delta}_u\left(\mathbf{I'}_u\right) z + 1\right]^{-1} \prod_{i\in\mathbf{I'}_u} \left(\hat{p}_{u,i}z+1\right)^{-1}, \quad z > 0. \tag{3.30}$$

This model results in a lower SINR CDF than the actual one, i.e.,

$$F_{\Phi_u^{(2)}}\left(z, \mathbf{I'}_u\right) \leq F_{\Phi_u}\left(z\right). \tag{3.31}$$

We present the proof of (3.31) as follows.

Proof.

$$\begin{aligned} F_{\Phi_u}\left(z\right) &\geq 1 - \frac{1}{\hat{\sigma}_N z+1} \prod_{i\in\mathbf{I}_u} \frac{1}{\hat{p}_{u,i}z+1} \\ &= 1 - \frac{\prod\limits_{i\in\mathbf{I'}_u} \left(\hat{p}_{u,i}z+1\right)^{-1}}{\left(\hat{\sigma}_N z+1\right) \prod\limits_{i\in\mathbf{I''}_u} \left(\hat{p}_{u,i}z+1\right)} \\ &\geq 1 - \frac{\prod\limits_{i\in\mathbf{I'}_u} \left(\hat{p}_{u,i}z+1\right)^{-1}}{\left(\hat{\sigma}_N z + \sum\limits_{i\in\mathbf{I''}_u} \hat{p}_{u,i}z+1\right)} \\ &= F_{\Phi_u^{(2)}}\left(z, \mathbf{I'}_u\right). \end{aligned}$$

□

3.2.3 Weighted Sum (WS) Model

In order to obtain an accurate estimation of the SINR CDF with the limited amount of reported CSI, we design a weighted sum (WS) model based on the S1 and S2 models.

The SINR expectations of the upper and lower bounds can be calculated with their probability distributions, respectively. We denote them as $\phi_u^{(1)}$ and $\phi_u^{(2)}$ and calculate them as

$$\phi_u^{(1)} = \int_0^\infty z \, dF_{\phi_u^{(1)}}\left(z, \mathbf{I'}_u\right), \tag{3.32}$$

$$\phi_u^{(2)} = \int_0^\infty z \, dF_{\phi_u^{(2)}}\left(z, \mathbf{I'}_u\right). \tag{3.33}$$

On the other hand, the reported reference signal receiving quality (RSRQ) indicates the mean ratio of the RSRP of the serving BS to the RSSI [110]. It is defined as

$$\text{RSRQ} = \frac{K \times \text{RSRP}}{\text{RSSI}}. \tag{3.34}$$

According to the reported RSRQ, we can calculate the measured mean SINR as

$$\phi_u = \left(\frac{1}{N_{sc} \times \text{RSRQ}} - 1\right)^{-1}. \tag{3.35}$$

We use a weighted-sum approach to combine the SINR CDFs obtained by the S1 and S2 models so that the actual one can be approximated. The SINR CDF of the WS model is designed as the mixture

$$F_{\Phi_u^{(w)}}\left(z, \mathbf{I'}_u\right) = w_1 F_{\Phi_u^{(1)}}\left(z, \mathbf{I'}_u\right) + w_2 F_{\Phi_u^{(2)}}\left(z, \mathbf{I'}_u\right), \quad z > 0, \tag{3.36}$$

where

$$\omega_1 = \begin{cases} \min\left\{\max\left\{\dfrac{\phi_u - \phi_u^{(2)}}{\phi_u^{(1)} - \phi_u^{(2)}}, 0\right\}, 1\right\}, & \phi_u^{(1)} \neq \phi_u^{(2)}, \\ 0.5, & \phi_u^{(1)} = \phi_u^{(2)}, \end{cases} \tag{3.37}$$

$$\omega_2 = 1 - \omega_1. \tag{3.38}$$

In (3.36), w_1 and w_2 are weight factors of the S1 and S2 models in the range of 0 to 1, respectively. The WS model is a straightforward linear combination of the upper and lower bounds of the SINR CDFs. Therefore, it is easy to be computed in practice.

3.3 Data Rate Analysis and Estimation

Based on the stochastic channel models developed in Section 3.2, we analyze the ergodic user data rates under different scheduling schemes in this section.

3.3.1 Max-Min Rate

We first calculate the obtainable data rate of an RB with the Shannon capacity as

$$r\left(\Phi_u\right) = \frac{N_{sc} S_e}{T_s} \log_2\left(1 + \Phi_u\right), \tag{3.39}$$

where N_{sc} is the number of physical subcarriers per RB, S_e is the number of effective symbols per RB, and T_s is the time duration of an RB. Then, the ergodic data rate of an RB while it is distributed to user u can be calculated by

$$\hat{r}_u = \int_0^\infty r\left(z\right) f_{\Phi_u}\left(z\right) dz. \tag{3.40}$$

Note that the result of the user data rate obtained by (3.40) depends on the CDF of the instantaneous user SINR, i.e, the stochastic channel model used for its calculation. While using the S1, S2 and WS SINR models that are based on partial CSI, it generates approximated results that are applicable for data rate estimation.

With the MMR scheduler, the ergodic data rate per user in a cell is approximately identical [113]. Therefore, it is the harmonic mean of the user data rates per RB, which is calculated as

$$\overline{r}_{u,mmr} = \left(\sum_{u \in \mathbf{U}_b} \hat{r}_u^{-1} \right)^{-1}. \tag{3.41}$$

3.3.2 Max-Sum Rate

The MSR scheduling scheme considers the instantaneous user channel states and allocates the radio resource to the user with the highest obtainable data rate in each frame. Thus, it is necessary to calculate the probability distribution of the scheduled SINR for each user at first. It is defined as the conditional PDF of a user's SINR while it is the highest in the associated user set and consequently scheduled by the MSR scheduler. Specifically, it is calculated as

$$f_{u,msr}\left(z\right) = f_{\Phi_u}\left(z\right) \prod_{v \in \{\mathbf{U}_b / u\}} F_{\Phi_v}\left(z\right), \quad z > 0. \tag{3.42}$$

Then, the ergodic data rate of a user is the expectation of its obtained data rate

under the condition that it is scheduled, i.e.,

$$\overline{r}_{u,msr} = \int_0^\infty r\left(z\right) f_{u,msr}\left(z\right) dz. \tag{3.43}$$

3.3.3 Proportional Fair Scheduling

To analyze the performance of PFS, we utilize the approximate relationship between the EMA data rate $R_u\left(t\right)$ and the ergodic user data rate $\overline{r}_{u,pfs}$ as described in Theorem 3.2. Since the EMA data rate acts as a constant parameter in the scheduling problem for a given scheduling frame, the frame index t is ignored in the theorem.

Theorem 3.2. *In PFS, when $\tau \gg 1$, there exists the approximation that*

$$\overline{r}_{u,pfs} \approx R_u. \tag{3.44}$$

Proof. According to the definition in (2.3), the expectation of the EMA data rate is expressed as

$$\mathbb{E}\left[R_u\left(t+1\right)\right] = \left(1-\frac{1}{\tau}\right)\mathbb{E}\left[R_u\left(t\right)\right] + \frac{1}{\tau}\mathbb{E}\left[r_u\left(t\right)\right]. \tag{3.45}$$

Assuming ergodicity of $R_u\left(t\right)$ for the stable PFS, it holds that

$$\mathbb{E}\left[R_u\left(t+1\right)\right] = \mathbb{E}\left[R_u\left(t\right)\right]. \tag{3.46}$$

Substituting (3.46) into (3.45), we have

$$\mathbb{E}\left[R_u\left(t\right)\right] = \mathbb{E}\left[r_u\left(t\right)\right] = \overline{r}_{u,pfs} \tag{3.47}$$

When the averaging coefficient $\tau \gg 1$, we have the approximation that

$$\frac{1}{\tau^T} \approx 0, \quad T \geq 2. \tag{3.48}$$

Considering again the ergodicity of $R_u\left(t\right)$, we assume that for a certain frame t there exists

$$R_u\left(t+T\right) = R_u\left(t\right), \tag{3.49}$$

which means that the status $R_u\left(t\right)$ repeats after a long enough period, i.e, T frames after frame t. Then, we can deduce $R_u\left(t\right)$ as follows,

$$\begin{aligned} R_u\left(t+T\right) &= \left(1-\frac{1}{\tau}\right)^T R_u\left(t\right) + \sum_{n=0}^{T-1}\frac{1}{\tau}\left(1-\frac{1}{\tau}\right)^{T-n-1} r_u\left(t+n\right) \\ &\overset{(3.48)}{\approx} \left(1-\frac{T}{\tau}\right) R_u\left(t\right) + \sum_{n=0}^{T-1}\frac{1}{\tau} r_u\left(t+n\right), \end{aligned} \tag{3.50}$$

By (3.49), when T is large enough, we have

$$R_u(t) \approx \frac{1}{T}\sum_{n=0}^{T-1} r_u(t+n) \approx \overline{r}_{u,pfs}. \tag{3.51}$$

Combining (3.47) and (3.51), it is proved that

$$R_u(t) \approx \mathbb{E}[R_u(t)] = \overline{r}_{u,pfs}, \quad \forall t. \tag{3.52}$$

□

Different from MSR, the PFS scheme selects the user with the highest weighted instantaneous data rate in each scheduling frame. Thus, the probability distribution of the scheduled SINR under PFS is calculated as

$$f_{u,pfs}(z) = f_{\Phi_u}(z) \prod_{v\in\{\mathbf{U}_b/u\}} F_{\Phi_v}\left(r^{-1}\left(\frac{r(z)R_v}{R_u}\right)\right), \quad z>0 \tag{3.53}$$

where $r^{-1}(\bullet)$ is the inverse function of the obtainable data rate given in (3.39). By Theorem 3.2, the PDF of scheduled SINR is approximately

$$f_{u,pfs}(z) \approx f_{\Phi_u}(z) \prod_{v\in\{\mathbf{U}_b/u\}} F_{\Phi_v}\left(r^{-1}\left(\frac{r(z)\overline{r}_{v,pfs}}{\overline{r}_{u,pfs}}\right)\right), \quad z>0. \tag{3.54}$$

Then, the ergodic user data rate under PFS is calculated as

$$\overline{r}_{u,pfs} = \int_0^\infty r(z) f_{u,pfs}(z)\, dz, \tag{3.55}$$

Hence, we obtain U equations of the ergodic user data rates with (3.55). Although their closed-form solutions are unobtainable, we can compute the results by numerical methods [114].

3.4 Influencing Factors of the Estimation Accuracy

The accuracy of data rate estimation is influenced by multiple factors as we mentioned. In this section, we analyze the major ones in detail and verify our analysis by the means of simulations in the next section.

3.4.1 Maximum Number of the Reported Cells

Recall that the number of the maximum reported neighbor cells per user is controlled by the system parameter I_R, namely *maxReportCells*, which is an integer in the range

of 1 to 8 [110]. A larger I_R allows more IRSRPs to be reported by every user. This increases the power proportion of knowable interference in the undesired signals, which is defined as

$$\psi_u\left(\mathbf{I'}_u\right) = \frac{\hat{p}_{u,\mathbf{I'}_u}}{\left(\hat{p}_{,\mathbf{I}_u} + \sigma_N\right)} = 1 - \frac{\hat{\delta}_u\left(\mathbf{I'}_u\right)}{\left(\hat{p}_{u,\mathbf{I}_u} + \sigma_N\right)}. \tag{3.56}$$

The analytical results of the SINR probability distributions are more accurate with less power in the unknown part of the undesired signals. However, more reported IRSRPs increase the signaling overhead on the control channel per user. On the contrary, the decrease of parameter I_R may result in larger errors of the analytical SINR models as well as the data rate estimation. This is because more IRSRPs are included in the indistinguishable power variable $\hat{\delta}_u\left(\mathbf{I'}_u\right)$ and their independence is ignored in the SINR models.

3.4.2 Increased Error by the Low-SINR Effect

The inaccurate SINR result can undoubtedly lead to an error in the estimated data rate. Due to the natural features of the relationship between the SINR and link capacity, the relative difference between the estimated user data rate and the actual one is related to the SINR level of a user. We calculate the relative differences of theoretically obtainable spectrum efficiencies with various SINR offsets according to the Shannon capacity as follows.

$$\varepsilon_{se}\left(\Phi_{\mathrm{dB}}, \Delta_{\mathrm{dB}}\right) = \left[\frac{\log_2\left(1 + 10^{\frac{\left(\Phi_{\mathrm{dB}} + \Delta_{\mathrm{dB}}\right)}{10}}\right)}{\log_2\left(1 + 10^{\frac{\Phi_{\mathrm{dB}}}{10}}\right)} - 1\right] \times 100\%, \tag{3.57}$$

where Φ_{dB} is the user SINR and Δ_{dB} is the offset of SINR in dB.

According to this formula, we plot $\varepsilon_{se}\left(\Phi_{\mathrm{dB}}, \Delta_{\mathrm{dB}}\right)$ as in Figure 3.1. The same SINR offset at high and low SINR levels results in different errors of the spectrum efficiencies. A user with a lower SINR is more sensitive to the deviation of its SINR. For the users with high SINRs, the same SINR offset leads to smaller relative differences of the obtained link capacities. For instance, to guarantee a relative difference between $\pm 5\%$, a user with the SINR of 20 dB can tolerate ± 1.0 dB errors. However, when the user SINR is -10 dB, only ± 0.2 dB and smaller offsets are allowable.

3.4.3 Impact of Propagation Environments

The path loss exponent varies in different communication environments. It is normally larger in the urban areas than in the suburban and rural areas due to its dense and tall buildings. This difference influences the accuracy of data rate estimation via two aspects analyzed as follows.

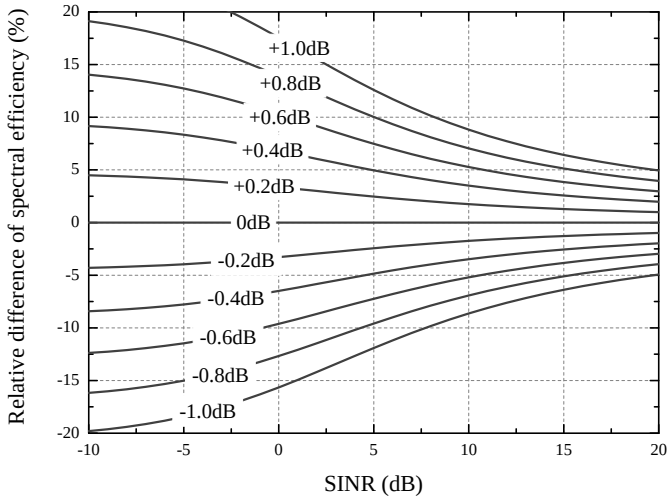


Figure 3.1: The relative differences of the theoretical spectrum efficiencies with different SINR values and offsets.

We denote the path loss exponent as α and rewrite the normalized IRSRP of an interferer as

$$\hat{p}_{u,i} = \frac{p_{u,i}}{p_{u,b}} = 10\alpha \lg \frac{d_{u,b}}{d_{u,i}} - X_{u,i} + X_{u,b}, \tag{3.58}$$

where $d_{u,i}$ and $d_{u,b}$ are the distances of user u from interferer i and its serving BS b, respectively. $X_{u,b}$ and $X_{u,i}$ are the corresponding loss factors of shadow fading in dB. A larger path loss exponent α increases the received power difference between the closer serving BS and farther interferers, resulting in a higher SINR. This law has been verified in [115] in detail. As we analyzed in Section 3.4.2, the data rate estimation errors caused by inaccurate channel modeling and measurement are lower while the users SINRs increase.

The same influence of the path loss exponent also exists in the relationship among the interferers. A higher α enlarges the power differences among interferers at different locations. Specifically, we calculate the power ratio between two IRSRPs as

$$\frac{p_{u,i}}{p_{u,j}} = 10\alpha \lg \frac{d_{u,j}}{d_{u,i}} - X_{u,i} + X_{u,j}. \tag{3.59}$$

Thus, the power of the reported closer and stronger interferers occupy a higher proportion of the total undesired signal power, resulting in a larger $\psi_u(\mathbf{I'}_u)$. As we explained in Section 3.4.1, this can improve the accuracy of the estimated SINR results and user data rates. According to the above two reasons, the data rate estimation errors are smaller in urban areas, where the path loss exponent is larger than that in suburbs.

Table 3.2: Simulation Parameters of the OFDMA Networks

Parameter	Value
BS Tx power	46 dBm
BS Tx antenna gain	18 dBi
Carrier frequency	1800 MHz
Urban path loss model	$138.47 + 38.22 \lg(d)$ dB
Suburban path loss model	$130.41 + 33.77 \lg(d)$ dB
Standard deviation of shadowing	8 dB
Noise power density	-174.5 dBm/Hz
Noise figure	7 dB
Number of RBs (K)	50 in 10 MHz
Number of subcarries per RB (N_{sc})	12
Number of effective symbols (S_e)	10 per frame
Frame duration (T_s)	1 ms
Minimum distance to BS	20 m
Average user density	20 per cell

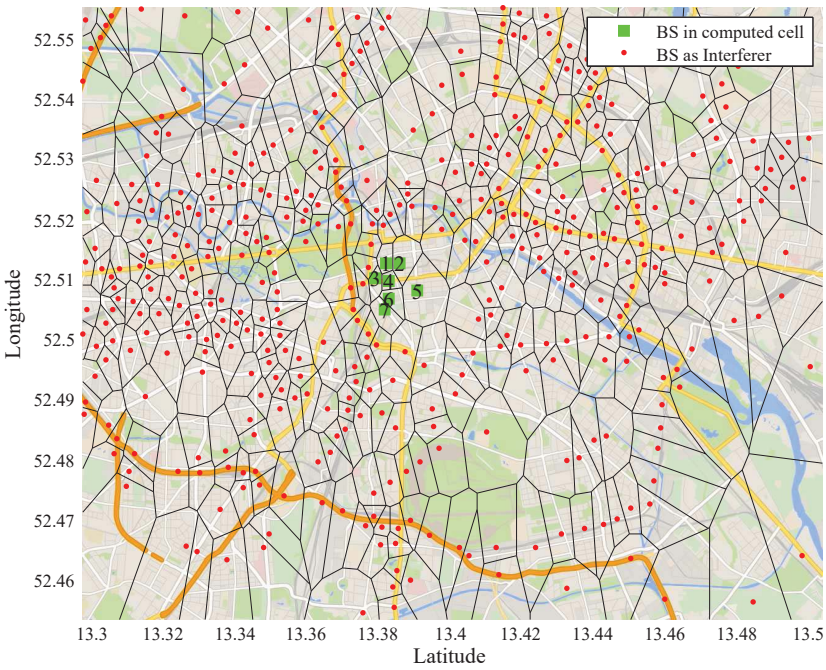
3.4.4 Accuracy of CSI Measurement

The RSRPs are the statistical results calculated according to the results of RS measurements. Since one RS exists only for one symbol at a time, the measurement is carried out on all of the RBs that contain the cell-specific RSs. The accuracy of the statistical RSRP results can be improved with more power samples of the received RSs utilized for calculation. In other words, the measurement of RSRP with a larger number of RBs can decrease the deviation of the reported CSI and consequently improve the accuracy of data rate estimation.

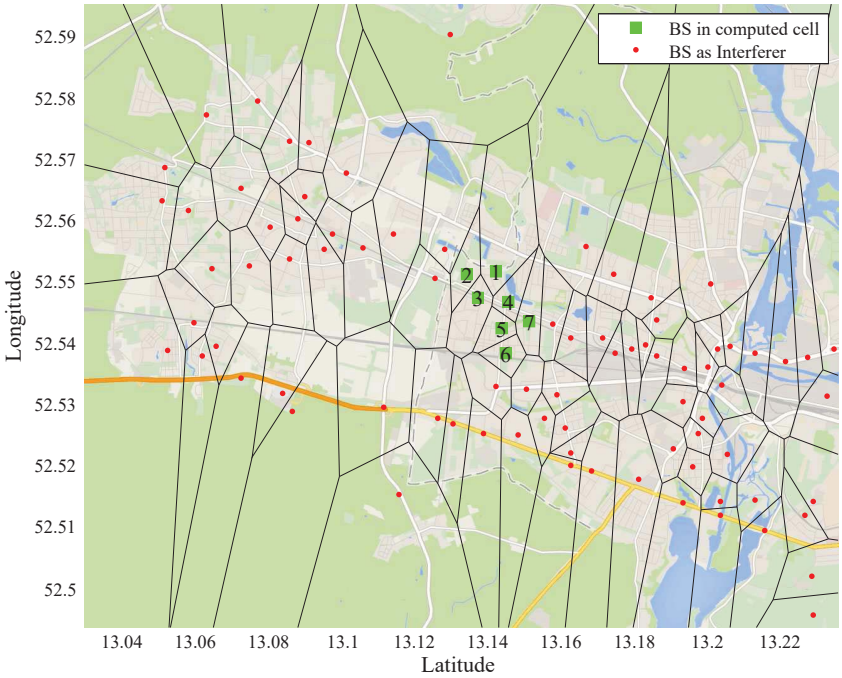
3.5 Simulations and Numerical Results

In this section, we evaluate the accuracy of the data rate estimation based on our stochastic channel models by simulations. We consider OFDMA-based downlink networks for the simulations with the system parameters listed in Table 3.2. The BS deployments in 13x12 km^2 rectangular areas in the urban area of Berlin and a nearby suburban area are adopted for the simulation scenarios, as shown in Figure 3.2 [116]. User terminals are uniformly randomly distributed in each scenario. To avoid edge effects, only the user data rates in the central 7 cells are computed and the BSs in other cells perform as pure interferers.

Figure 3.3 presents the SINR CDFs of three randomly selected users in the central no. 4 cell in the urban scenario, which are obtained by different stochastic channel models as well as the simulations. The results obtained by the S1 and S2 models result in the upper and lower bounds of the simulation results, respectively. This



(a) Urban scenario (Berlin, Germany).



(b) Suburban scenario (Spandau, Germany).

Figure 3.2: The BS deployments in the urban and suburban scenarios (Map source and copyright Google Maps, Google Inc.).

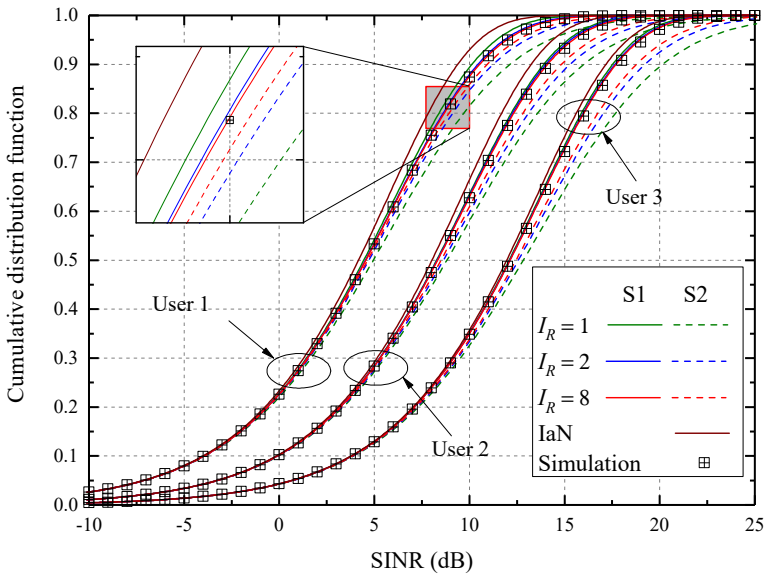
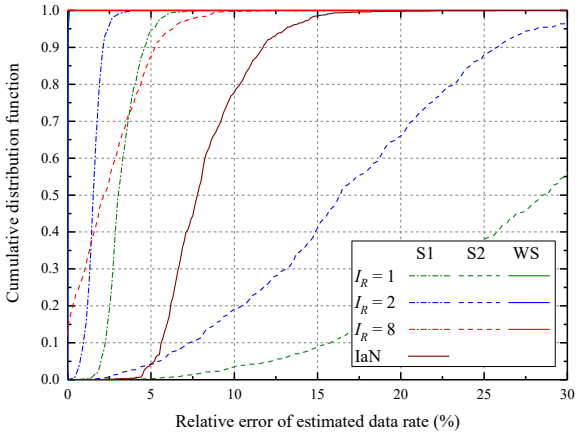


Figure 3.3: The CDFs of instantaneous user SINRs obtained by different stochastic channel models.

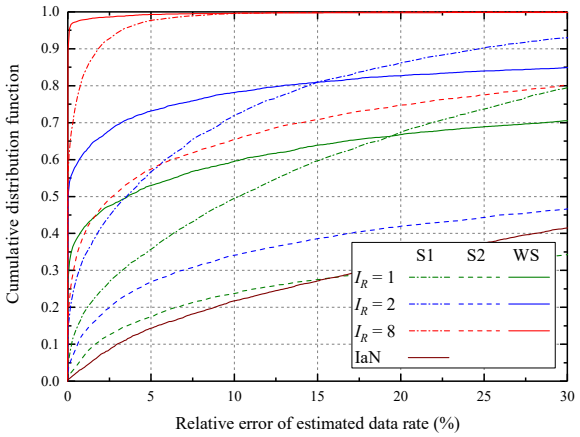
is consistent with our analysis in Section 3.2. As the parameter I_R decreases, the gaps between our analytical results and the actual ones increase, i.e., the relative differences become larger. This is because fewer IRSRPs reported from the user devices lead to a higher error of the analytical SINR model. The results based on the IaN model are also presented in the figure. It can be regarded as a special case of the S1 model while $I_R = 0$. Thus, it ignores the independence of interference signals and results in much higher CDF values and larger errors in comparison to the S1 model.

To evaluate the accuracy of our stochastic channel models while applying them to user data rate estimation, we compare the estimated data rates to the simulation results. The relative errors are calculated for the evaluation and their CDFs under different scheduling schemes are presented in Figure 3.4. The estimated data rates with the S1 model obtains lower errors than those with the S2 model. The difference between these two models is that the former one regards the unreported low IRSRPs as constant noise, which is closer to the actual case. However, the unreported useless signals are modeled with an exponentially distributed variable that has a larger variation in the S2 model. Therefore, S2 is less accurate than S1, especially when parameter I_R is low. As the number of reported IRSRPs increases, both of the S1 and S2 models result in lower estimation errors and are superior to the IaN model. In addition, the WS model yields more accurate data rate estimation even with very few reported IRSRPs.

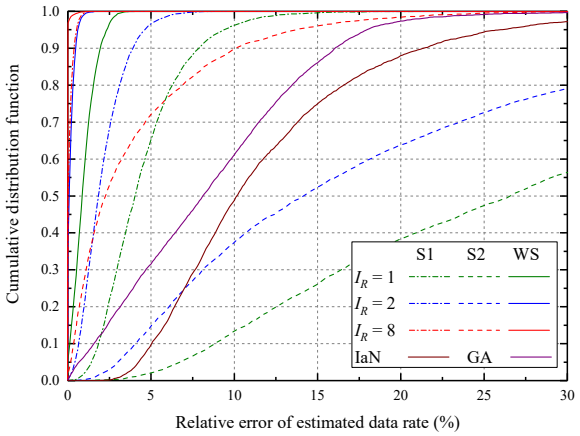
In particular, the WS model obtains extremely high accuracy in the data rate estimation under MMR scheduling. As shown in Figure 3.4(a), the estimated results have nearly no errors even with only one reported IRSRP per user. In contrast, the relative error is much larger in the data rate estimation under MSR scheduling,



(a) MMR.

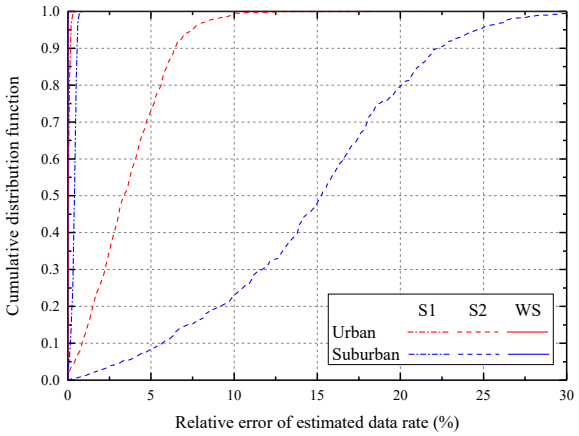


(b) MSR.

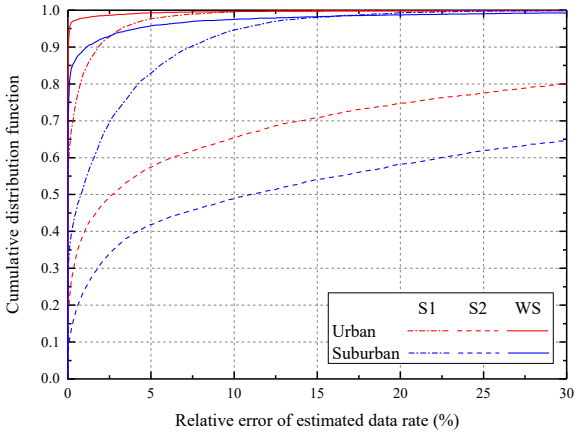


(c) PFS.

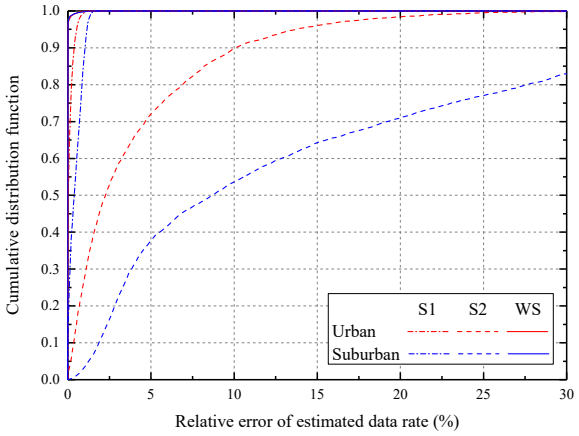
Figure 3.4: Relative errors of the estimated user data rates with different stochastic channel models (urban scenario).



(a) MMR.



(b) MSR.



(c) PFS.

Figure 3.5: Relative errors of the estimated user data rates in the urban and suburban scenarios ($I_R = 8$).

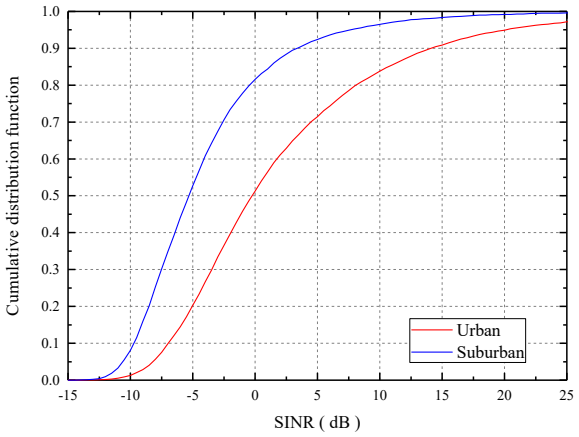


Figure 3.6: The CDFs of mean SINRs in the urban and suburban scenarios.

as shown in Figure 3.4(b). The results of the WS model are accurate only when parameter I_R is large enough, e.g., $I_R = 8$. The MMR and MSR schemes have different sensitivities to the change of parameter I_R due to their different scheduling strategies. In the MMR scheme, the instantaneous channel state is not taken into account for scheduling. Thus, the inaccuracy of the stochastic channel modeling has an insignificant impact on its scheduling behavior. The estimated data rate is mainly determined by the mean spectrum efficiency of each user link. As shown in (3.40), the impact of random channel characteristics is eliminated. However, the scheduling behavior of the MSR scheme entirely depends on the instantaneous channel states. Therefore, the inaccurate SINR models influence not only the calculation of link capacities but also the scheduled probabilities given in (3.42). Thus, even very small errors in the probability distributions of user SINRs can lead to considerably inaccurate estimated data rates in the case of MSR.

The accuracy of data rate estimation under PFS is in between the MMR and MSR schemes, as shown in Figure 3.4(c). Its scheduling behavior also depends on the instantaneous link capacities, but with weighted factors. Thus, its data rate estimation is influenced by inaccurate SINR modeling less than the MSR scheme. We also present the data rate estimation results obtained by the GA model which is a simplified symmetric user channel model [33,34]. It is better than the S2 model only when the parameter I_R is very low. However, our WS model is superior to both of the IaN and GA models even with only one reported IRSRP per user.

We compare the data rate estimation errors in the urban and suburban scenarios, as shown in Figure 3.5. The results obtained by all of the three SINR models in the urban scenario are more accurate than those in the suburb. As we discussed in Section 3.4.3, the main reason for this difference lies in the higher path loss exponent in the urban environment.

To further confirm the influence of propagation environments, we present the distributions of user mean SINRs in Figure 3.6. In the urban area, users have higher SINRs than those in the suburb. In addition, the power proportion of the reported IRSRPs per user is always higher with different I_R values in the urban

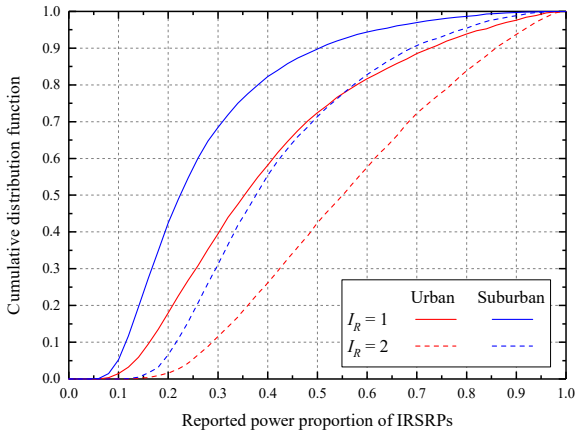


Figure 3.7: The CDFs of reported interference power proportion $\psi_u(\mathbf{I'}_u)$ in the urban and suburban scenarios.

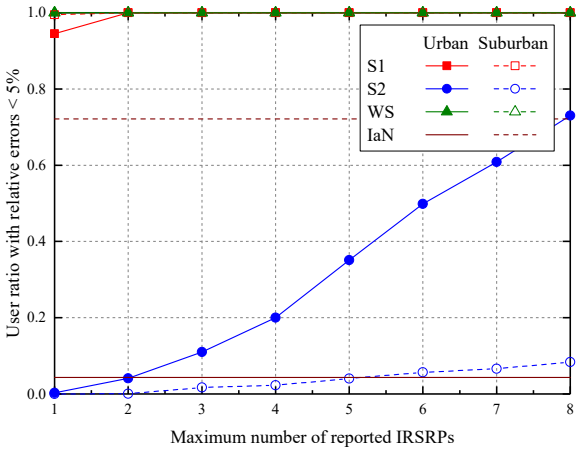
area, as shown in Figure 3.7. This means that the random channel characteristics of the signals with smaller power proportions are ignored. Therefore, the estimated probability distributions of SINRs are more accurate under this condition. Both of the above two reasons contribute to the lower data rate estimation errors in the urban scenario as we analyzed. In addition, the WS model overcomes this defect caused by the radio propagation environment in suburbs and achieves high estimation accuracy as in the urban area.

In practical applications, small errors of the estimated data rates are always inevitable and tolerable. We define the low-error users as the ones which have their data rate estimation errors lower than 5%. We focus on the low-error users and compute their population ratios with different numbers of reported IRSRPs, as shown in Figure 3.8.

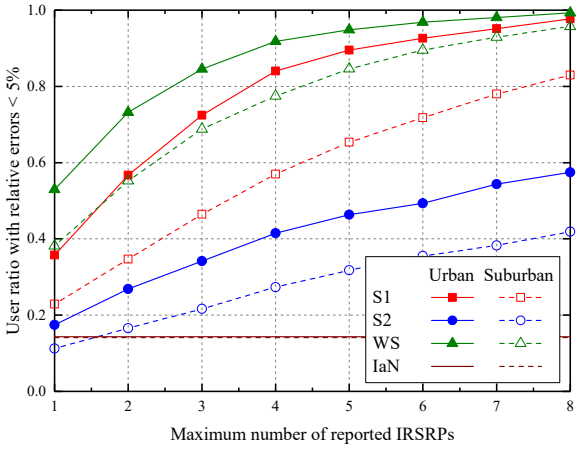
Under MMR scheduling, both of the S1 and WS models obtain very high low-error user ratios that are insensitive to the change of parameter I_R, as shown in Figure 3.8(a). This means that they are robust even with very few reported IRSRPs. Thus, the signalling overhead can be reduced for CSI feedback in practice. However, the S2 model is influenced by the parameter I_R largely and is worse than the IaN model in the suburban scenario. Since the IaN model uses no independent IRSRP information for its SINR modeling, I_R has no impact on the data rate estimation based on it.

The similar results are obtained under PFS, as shown in Figure. 3.8(c). However, the S1 model needs more reported IRSRPs, i.e., $I_R > 2$, to guarantee that all users are the low-error ones. On the other hand, the GA model results in more inaccurate estimated data rates since it is designed for the single-cell networks and is infeasible for the multi-interference cases.

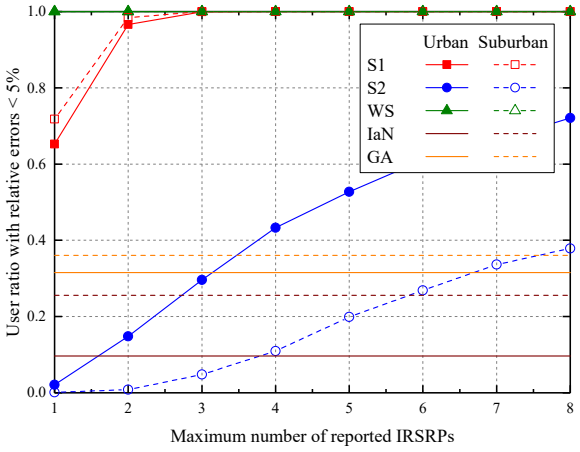
In contrast to MMR and PFS, there are much fewer low-error users under MSR scheduling, as shown in Figure 3.8(b). The WS model results in a high number of low-error users only when parameter I_R is large enough. For instance, to achieve 90%



(a) MMR.



(b) MSR.



(c) PFS.

Figure 3.8: The low-error user ratios with different numbers of reported IRSRPs, I_R.

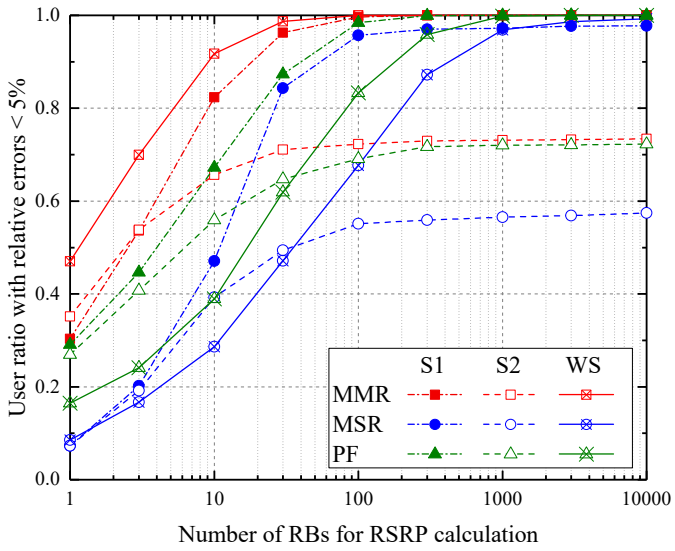


Figure 3.9: The low-error user ratios with imperfect measurements of RSRPs ($I_R = 8$).

low-error users, 4 and 6 reported IRSRPs are required in the urban and suburban areas, respectively. In addition, the low-error user ratios are lower in the suburban scenario with all of the three SINR models. Therefore, more detailed CSI is necessary for improving the accuracy of data rate estimation under MSR scheduling, especially in suburbs.

We also investigate the estimation error caused by the imperfect measurement of CSI. In practice, each reported RSRP value is computed by averaging the received RS power samples over multiple RBs. Under Rayleigh fading channels, the mean value of the power samples follows the Erlang distribution. Thus, the deviations of the statistical RSRP results are large if the RSs in very few RBs are measured. Figure 3.9 presents the low-error user ratios with different number of RBs used for calculating RSRPs. The results of the MMR scheme are less sensitive to the imperfect channel measurement than MSR and PFS due to their different scheduling strategies as we discussed. However, it is still necessary to use a certain amount of RBs for the RSRP calculation to ensure a high ratio of low-error users. As shown in the figure, to achieve a 90% low-error user ratio under MMR scheduling, the WS model needs at least 10 RBs to calculate each reported RSRP. The required RB amount increases to 1,000 in the MSR and PFS cases for obtaining a similar user ratio. In addition, with imperfect CSI, it is possible that the WS model obtains more inaccurate estimation results than the S1 and S2 models. This is because the weight factors in the WS model are also calculated according to the reported RSRPs. Thus, inaccurate CSI can lead to deviated weight factors, making the WS model failed. This effect is more significant for the MSR and PFS schemes due to the strong correlation between their scheduling behaviors and the instantaneous channel states. Therefore, accurate detection of the channel states is the prerequisite for robust data rate estimations.

3.6 Summary

In this chapter, we have focused on the problem of data rate estimation with stochastic channel models in multi-cell wireless networks. We derived the closed-form probability distribution of the instantaneous user SINR, based on which we further developed its upper and lower bounds while considering a limited number of reported interferers. Then, we extended these SINR models to a WS model with the purpose of more accurate estimation of user data rate. It is applied to data rate performance analysis of three classic scheduling schemes, namely, MMR, MSR, and PFS. The simulation results have verified that it can obtain very accurate results for user data rate estimations in both of the urban and suburban scenarios and is superior to the IaN and GA models even with very small signaling overhead for CSI feedback. The accuracy of the estimated data rates under MSR scheduling is more sensitive to the inaccurate SINR models than in the MMR and PFS cases. Therefore, sufficient and precise CSI is necessary for ensuring robust data rate estimations under MSR scheduling. In addition, we analyzed the factors that influence the estimation accuracy and assessed them by simulations. The impact of low SINRs on the estimation results is overcome greatly by the WS model. However, inaccurate channel measurements inevitably increase estimation errors. Thus, a sufficient number of RBs are required for calculating the reported RSRPs in order to improve the accuracy of data rate estimation.

4 Traffic Load Balancing Based on User Data Rate Estimation

In this chapter, we apply our work on the user data rate estimation in Chapter 3 to the design of user association schemes for traffic load balance in the multi-cell networks. In order to deliver high-quality mobile services with a minimum cost of base station (BS) deployment, wireless cellular networks have been trending towards increasing heterogeneity [117,118]. In such a heterogeneous cellular network (HCN), various low-power access points, such as femtocells and picocells, are overlaid with the conventional macrocells in a more targeted manner, as shown in Figure 4.1. This hybrid deployment of BSs is more flexible and specific for the expansion of cell-edge coverage and throughput improvement.

Among the most significant issues in HCNs, traffic load balancing is crucial for transmission efficiency and fairness. It refers to a balanced amount of users or applications accessing different types of cells and sharing the limited radio resources in the network. Hence, the users in various cells are able to enjoy fair and high qualities of services. In conventional cellular networks, each user accesses the BS that provides the best channel quality, namely, the highest signal-to-interference-plus-noise ratio (SINR). This user association scheme ensures the largest cell coverage with a given SINR target. However, it may lead to unbalanced traffic loads in HCNs. The macrocell BSs are normally built higher and have much larger transmit power than the ones in picocells and femtocells. Therefore, they have larger coverage areas and more served users. This unbalanced user association approach makes the small cells less useful, even though they are capable of providing more bandwidth for taking over the heavy traffic loads in macrocells.

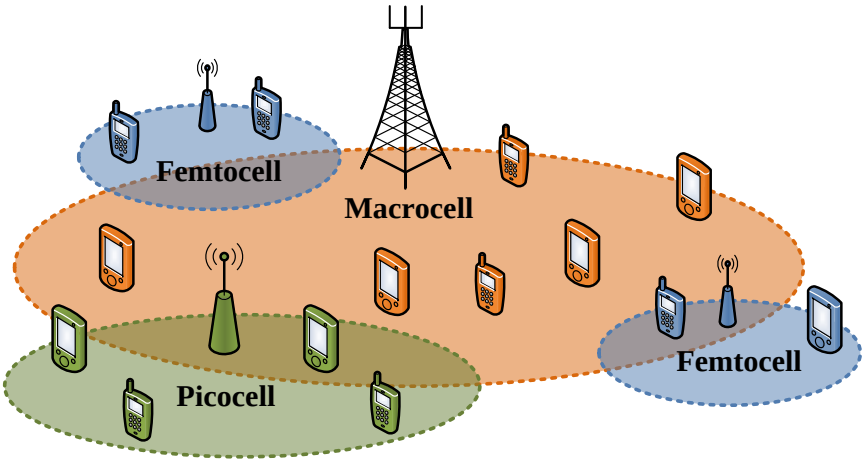


Figure 4.1: The hybrid deployment of BSs in a heterogeneous cellular network.

There has been a rich literature that treats the user association problem towards traffic load balance in HCNs. Two broad classes of their balancing strategies are cell range control and user access control. The former strategy tries to balance the traffic loads among cells by enlarging or shrinking the cell coverage ranges. This can be realized by the mean of the cell breathing technique which adjusts the transmit power of the BSs in different types of cells to control their transmission distances [119]. In addition, the biased association is another simple and effective small cell expansion method. Different bias factors are given to the BSs in various cells for associating with appropriate numbers of users and traffic loads [120]. However, the uneven distributions of the BSs and users in HCNs can diminish the effect of these cell range control techniques.

The user access control strategy considers the specific real-time channel conditions and user locations, and optimizes user association according to given objectives. A number of offline methods have been proposed to solve the user association problem for the entire network. In [121] and [122], the BSs and their channel bandwidths are dynamically assigned to the users for maximizing the logarithmic sum of the user data rates. The max-min rate (MMR) target is adopted in the user association problem in [123] for maintaining user fairness. In order to improve the weighted-sum rate (WSR), approximation algorithms have been proposed via the convex program relaxation or via discretized linear program relaxation in [124]. In [113], the user data rates under proportional fair scheduling (PFS) are estimated by the symmetric high-SINR channel modeling [125]. The estimation results are utilized for solving the optimization problem of user association with the logarithmic sum rate objective.

The offline traffic load balancing methods suffer from two problems in practice. Firstly, the globally optimal solutions are intractable due to the fact that the optimization problems for the whole HCN are normally NP-hard as proved in the aforementioned works. Therefore, the computational complexity for solving the problems increases significantly with the scale of the network and the user density in it. In addition, all users have to update their serving BSs according to the optimized solution and many of them may handover to alternative BSs whenever a substantial change takes place in the network, e.g., new user arrivals. Therefore, it is difficult to apply offline schemes to practical networks.

Different from the offline approaches, the online traffic load balancing schemes execute a partial change of the user association in a distributed manner according to the network status and the optimization target. In [119] and [126], users are assumed to be able to access multiple BSs dynamically. To improve the MMR performance, congestion load minimization algorithms have been proposed based on complete or limited knowledge of the user association and traffic load. In [127] and [128], the alpha-fairness metric has been adopted as the objective of the convex optimization problem for traffic load balancing, and an iterative distributed user association policy has been proposed for solving it. A primal-dual distributed algorithm is designed in [129] with the aim of maximizing the logarithmic sum rate. In [130], an approximation-based algorithm is proposed with a service-level-indicating metric for guaranteeing a given level of QoS in the HCN.

In the online traffic load balancing schemes proposed in the existing literature, the radio resources are allocated to the users equally or statically in each cell. This means that the impact of the dynamic resource allocation on the transmission performance has not been taken into account for the design of traffic load balancing schemes. As we discussed in Chapter 3, the estimated user data rates can be utilized for assisting system operations. In this chapter, we design online traffic load balancing schemes based on our data rate estimation.

The rest of this chapter is organized as follows. In Section 4.1, we firstly introduce a conventional cell rage control scheme, namely, maximum biased-received-power (Max-BRP). It is regarded as a benchmark user association strategy. In Section 4.2, we design iterative user handover algorithms with three different optimization objectives for traffic load balancing, including the max-sum rate (MSR), MMR and proportional fairness (PF). The corresponding dynamic resource allocation schemes and their user data rate estimation results are applied to our traffic load balancing schemes. In Section 4.3, we build a two-tier HCN for system-level simulations and evaluate the performance of our proposed schemes with different network deployments. Then, we apply our schemes to a simulation scenario which is based on the realistic site locations in the city of Los Angeles for its performance verification in Section 4.4. Finally, this chapter is summarized in Section 4.5.

4.1 The Max-BRP Scheme

To balance the numbers of users associated to different types of cells, the maximum BRP (Max-BRP) has been proposed as a simple cell range control method, which is an enhanced scheme based on the conventional maximum-SINR (Max-SINR) user association [120]. The cells in smaller sizes are given higher priorities for user access by using larger bias factors. We describe the Max-BRP scheme in detail as follows.

Remind that the user index set is

$$\mathbf{U} = \{u \,|u = 1, \ldots, U\}, \tag{4.1}$$

and the BS index set is

$$\mathbf{B} = \{b \,|b = 1, \ldots, B\}. \tag{4.2}$$

Note that set $\mathbf{B}$ contains various BSs in different types of cells in HCNs, including macrocells, femtocells, picocells, etc. We denote the user set served by BS b as $\mathbf{U}_b$. Since every user has only one serving BS in the network, there exists

$$\mathbf{U} = \underset{b \in \mathbf{B}}{\cup} \mathbf{U}_b. \tag{4.3}$$

To ensure a good channel quality, a user needs to choose the BS with the highest average delivered power in the Max-SINR scheme. Therefore, a BS has its associated

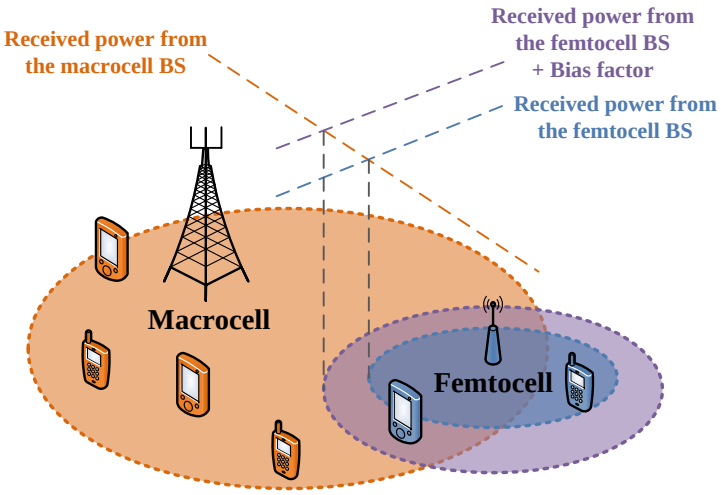


Figure 4.2: Illustration of the Max-BRP scheme for cell range control in HCNs.

user set as

$$\mathbf{U}_b = \left\{ u \,\middle|\, u \in \mathbf{U}, \arg\max_{i \in \mathbf{B}} \{p_{u,i}\} = b \right\}, \quad b \in \mathbf{B}, \tag{4.4}$$

where $p_{u,b}$ is the reference signal received power (RSRP) of BS b. It is detected by user u and reported via CSI feedback [110].

Based on the Max-SINR scheme, a bias factor is adopted in the Max-BRP scheme. Normally, the bias factors are identical for the same type of cells in one tier of the HCN. We denote the bias factor assigned to BS b as $\beta_b \geq 0$ dB. Thus, its associated user set is expressed as

$$\mathbf{U}_b = \left\{ u \,\middle|\, u \in \mathbf{U}, \arg\max_{i \in \mathbf{B}} \{p_{u,i}\,[\mathrm{dB}] + \beta_i\} = b \right\}, \quad b \in \mathbf{B}. \tag{4.5}$$

In Figure. 4.2, we present an example of the Max-BRP scheme in a two-tier HCN with one macrocell and one femtocell. Instead of changing the power transmitted by each BS as in the cell breathing methods, the Max-BRP scheme controls the coverage area of the femtocell by adjusting its bias factor. In order to relieve the high traffic load in the macrocell, the bias factor is positive so that users tend to access the femtocell. As the bias factor increases, more cell-edge users in the macrocell switch to the femtocell. Therefore, the Max-BRP scheme provides a convenient way for balancing traffic loads among different tiers in HCNs.

However, the uneven deployment of BSs and user distribution can diminish the effect of the Max-BRP scheme. For instance, some femtocells in the network may be more crowded than the others, and thus need to be assigned different bias factors. Even if the bias factor of each cell can be controlled independently, it can only adjust their coverage ranges roughly. The specific user distributions and the traffic loads in different cells are neglected. Thus, delicate and targeted control of the independent user access is more favorable to the traffic load balancing with given

network optimization objectives.

4.2 Traffic Load Balancing Schemes

In order to balance the traffic load in HCNs with various optimization objectives, we design three iterative user handover algorithms based on different dynamic resource allocation schemes in this section.

4.2.1 Throughput-Oriented Schemes

As we discussed, the uneven distributions of BSs and users may lead to unbalanced traffic loads in different types of cells and consequently decline system performance. To address this problem, it is necessary to control appropriately the users accessing each BS, i.e., the user association set $\mathbf{U}_b$, $b \in \mathbf{B}$. We design an online traffic load balancing scheme based on iterative user handover operations. The generalized utility function is considered firstly, followed by two specific optimization targets, namely, MSR and PF.

According to our analysis in Chapter 3, the ergodic user data rates under a given scheduling scheme can be estimated based on the reported user channel state information (CSI). The user data rates in a cell depend on the associated user set and can be expressed as a function of $\mathbf{U}_b$, i.e., $\overline{r}_u(\mathbf{U}_b)$, $u \in \mathbf{U}_b$.

We define a utility function of BS b as $\eta_b(\mathbf{U}_b)$ in the traffic load balancing problem. This utility value depends on the user data rates in the cell and changes with different associated user sets. The overall utility metric in the network is denoted as

$$\eta = \sum_{b \in \mathbf{B}} \eta_b(\mathbf{U}_b). \tag{4.6}$$

We denote the utility change of BS b when one of its user $u \in \mathbf{U}_b$ switches to another BS as

$$\eta_b^-(\mathbf{U}_b, u) = \eta_b(\mathbf{U}_b / \{u\}) - \eta_b(\mathbf{U}_b). \tag{4.7}$$

Similarly, the utility change when a new user $v \notin \mathbf{U}_b$ accesses BS b is denoted as

$$\eta_b^+(\mathbf{U}_b, v) = \eta_b(\mathbf{U}_b \cup \{v\}) - \eta_b(\mathbf{U}_b). \tag{4.8}$$

Note that the changes in utilities given in (4.7) and (4.8) can be either positive or negative, depending on the associated users and their channel states before and after a potential handover operation.

A user can detect the reference signals transmitted from the neighbor cells and reports their IRSRPs to its serving BS, which can be utilized for user data rate estimation. In addition, this implies that the reported neighbor cells offer relatively better channel conditions than the ones located farther away. Thus, we consider

the reported interfering cells in set $\mathbf{I'}_u$ as the potential handover targets for user u. Among them, we choose the neighbor cell with the joint maximum utility increment as the target cell that a user intends to switch to, which is denoted as

$$i_u^* = \arg\max_{i \in \mathbf{I'}_u} \left\{ \eta_i^+ \left(\mathbf{U}_i, u \right) + \eta_b^- \left(\mathbf{U}_b, u \right) \right\}, \quad u \in \mathbf{U}_b. \tag{4.9}$$

The corresponding utility increment brought by this handover operation is

$$\eta_u^* = \max_{i \in \mathbf{I'}_u} \left\{ \eta_i^+ \left(\mathbf{U}_i, u \right) + \eta_b^- \left(\mathbf{U}_b, u \right) \right\}, \quad u \in \mathbf{U}_b. \tag{4.10}$$

We define this potential utility increment as the handover factor of user u.

Of all users in $\mathbf{U}$, we use a greedy policy to choose the one with the largest positive handover factor $\eta_u^* > 0$ and execute its handover operation in each scheduling frame until there are no users offering positive handover factors. In addition, to avoid a user continuously switching among neighbor cells, a minimum staying duration is used to limit the user handover frequency. Specifically, a user cannot change its serving BS until it has been staying in the cell for at least T_h frames.

We propose two traffic load balancing schemes with the MSR and PF targets, respectively. In each of them, the corresponding resource scheduler is utilized. Their utility functions for handover are defined as follows.

- ▷ Max-Sum Rate
  In order to maximize the overall throughput in the network, the MSR scheduling scheme is utilized for dynamic resource allocation in each cell. In this case, the utility function per cell is defined as the sum of user data rates, i.e.,
  $$\eta_b \left(\mathbf{U}_b \right) = \sum_{u \in \mathbf{U}_b} \overline{r}_u \left(\mathbf{U}_b \right), \quad b \in \mathbf{B}. \tag{4.11}$$

- ▷ Proportional Fairness
  To balance the two performance targets, namely, the aggregate throughput and user fairness, PF is applied as the objective of the handover scheme. To this end, we use the PFS scheme for dynamic resource allocation per cell. It is proved that PFS results in the maximum logarithmic sum of user data rates in the long run [28]. Thus, we define the utility function per cell accordingly as follows.
  $$\eta_b \left(\mathbf{U}_b \right) = \sum_{u \in \mathbf{U}_b} \ln \left[\overline{r}_u \left(\mathbf{U}_b \right) \right], b \in \mathbf{B}. \tag{4.12}$$

In each iteration of the handover algorithm, the overall utility in the network must increase due to the positive handover factor. On the other hand, the utility cannot increase infinitely with limited radio resources and link capacities. Therefore, the algorithm must converge at the end, while no additional improvement of the utility is obtainable by changing the user association in the network.

Table 4.1: Simulation Parameters of the Two-Tier Experimental HCN

Parameter	Value
BS transmit power in macrocell/femtocell	43 dBm / 23 dBm
Path loss exponent in macrocell/femtocell	3.5 / 4
Number of resource blocks (RBs)	25 in 5 MHz
Noise power density	-174 dBm/Hz
Number of subcarries per RB (N_{sc})	12
Number of effective symbols (S_e)	10 per frame
Frame duration (T_s)	1 ms
Minimum staying time (T_h)	1,000 ms

4.2.2 A Fairness-Oriented Scheme

With the purpose of maximizing user fairness in the network, it is necessary to improve the data rates of the cell-edge users so that they are able to approach the average rate level in the network. Therefore, we adopt the MMR scheduler for dynamic resource allocation in each cell. The utility function in the fairness-oriented traffic load balancing scheme is defined as

$$\eta_b\left(\mathbf{U}_b\right) = \min_{u\in\mathbf{U}_b}\left\{\overline{r}_u\left(\mathbf{U}_b\right)\right\}, \quad b \in \mathbf{B}. \tag{4.13}$$

This implies that the lowest user data rate in the cell needs to be enhanced. Accordingly, we design the handover factor of each user as

$$\eta_u^* = \max_{i\in\mathbf{I'}_u}\left\{\min\left\{\eta_b\left(\mathbf{U}_b/\left\{u\right\}\right), \eta_i\left(\mathbf{U}_i\cup\left\{u\right\}\right)\right\} - \min\left\{\eta_b\left(\mathbf{U}_b\right), \eta_i\left(\mathbf{U}_i\right)\right\}\right\}, \quad u \in \mathbf{U}_b, \tag{4.14}$$

which is the maximum enhancement of the joint lower bound of the user data rates in the current cell of user u and its potential handover neighbor cells. Similar to the throughput-oriented schemes, we choose the user with the largest positive handover factor $\eta_u^* > 0$ in the network and switch it to the target neighbor cells iteratively until the algorithm converges.

4.3 Simulation and Numerical Results

In this section, system-level simulations are carried out by the Matlab software for evaluating the performance of our proposed traffic load balancing schemes and comparing it to the Max-BRP scheme [131]. We build a two-tier OFDMA-based experimental HCN with the system parameters and configurations given in Table 4.1. The femtocells in the second tier, which have relatively lower transmit power and smaller coverages, are randomly located overlapping with the macrocells in the first tier. Our proposed balancing schemes start from the user association results obtained by the Max-SINR scheme. The simulation time is set to 20,000 frames to guarantee

Table 4.2: BS and User Distributions of the Two-Tier Experimental HCN

Distribution	**Case 1**	**Case 2**	**Case 3**
Macrocell BS	PPP, in a circle radius = 750 m $\rho_1 = 1/\left(250^2\pi\right)\,\mathrm{m}^{-2}$		Hexagonal grid, 7 cells Inter-site distance 500 m
Femtocell BS	PPP $\rho_2 = 4\rho_1$		
User terminal	PPP	50% PPP in macrocells + 50% Matern-cluster in femtocells (radius = 20 m)	PPP
	$\rho_U = 20\rho_1$		

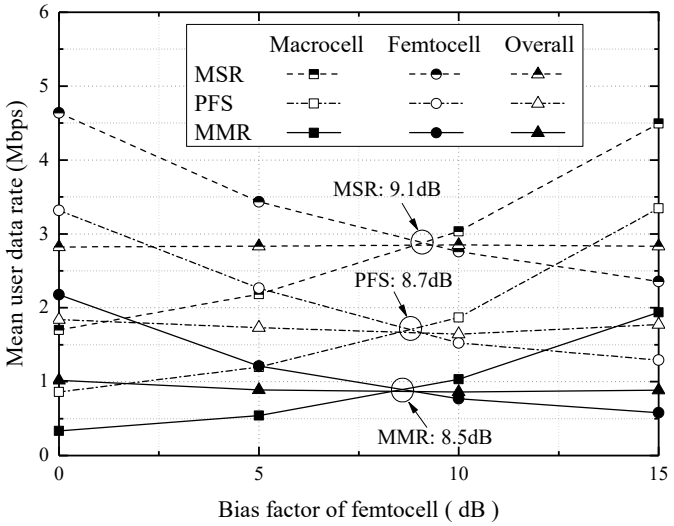


Figure 4.3: Mean user data rates with the Max-BRP scheme in case 1.

the convergence of the algorithms.

Three different cases of BS deployments and user distributions are developed as shown in Table 4.2. In case 1 and 2, the macrocell BSs are randomly located according to the Poisson point process (PPP) with a density of ρ_1 in a circle area. In case 3, there are 7 macrocells deployed in the hexagonal grid pattern, i.e., one BS in the center and six other BSs with the same distances around it. The femtocell BSs are all PPP-distributed with a density of $4\rho_1$ in the three cases. The density of user terminals is $20\rho_1$. They are also PPP-distributed in case 1 and 3, while half of the users are cluster-distributed around femtocells in order to simulate hotspots in case 2. We adopt the Matern-cluster process for modeling the user cluster distribution [132].

Firstly, we evaluate the throughput of the Max-BRP scheme with various resource schedulers, including MSR, PFS, and MMR. Figure 4.3 presents the mean user data rates in the scenario of case 1. The bias factor of the second tier controls the user preference for accessing the femtocells. More users transfer to the second tier as the bias factor increases, leading to fewer radio resources obtained per user and decline

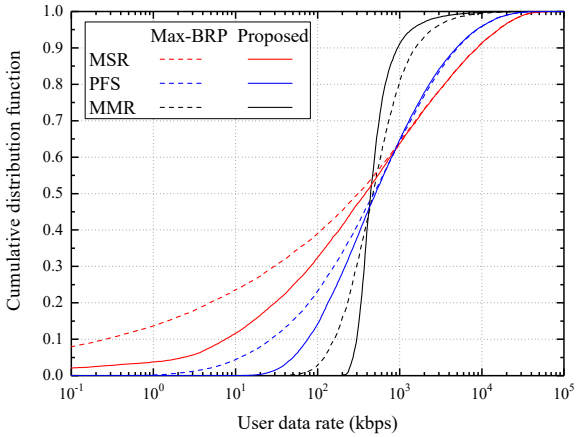
of the user data rates in femtocells. In contrast, the reduced number of users in macrocells obtain more bandwidth per user. Therefore, the mean user data rate in macrocell increases with the bias factor.

As shown in Figure 4.3, the variation of the throughput in the entire HCN is very slight as the bias factor changes. However, an appropriate bias factor can help to obtain balanced traffic load in each tier, i.e., the same level of the mean user data rates in both of macrocells and femtocells. With various resource schedulers, the optimal bias factors are different as noted in the figure. The MSR scheduler obtains the highest throughput while MMR results in the lowest one. This difference is ascribed only to their different resource scheduling mechanisms since the identical Max-BRP scheme is utilized.

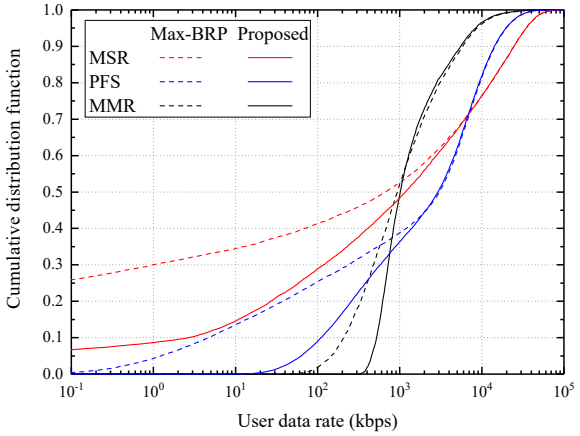
Figure 4.4 shows the cumulative distribution functions (CDFs) of the user data rates with the Max-BRP and our proposed traffic load balancing schemes. The results of the Max-BRP scheme are obtained by given the optimal bias factors as discussed in Figure 4.3. In comparison to the Max-BRP scheme, our proposed traffic load balancing schemes effectively decrease the proportion of low-bitrate users by delicately switching independent users from crowded cells to the ones with lower traffic loads. This cell-edge throughput improvement is more significant under the condition of unevenly distributed users. For instance, the ratio of the users with data rates lower than 0.1 kbps is reduced from 26% to 6% by our proposed MSR-based scheme in case 2. Moreover, as shown in Figure. 4.4, this improvement has nearly no impact on the performance of high-bitrate users in the cases of using the MSR and PFS schemes. Nevertheless, these two throughput-oriented schemes obtain a larger difference in user data rates and thus have poorer user fairness than the MMR-based scheme. The latter one reduces the user differences due to its fairness-oriented utility metric at the expense of decreasing the throughput of the high-bitrate users.

To further compare the throughput performance of the Max-BRP and our proposed traffic load balancing schemes, their mean user data rates in different cases are computed as shown in Figure 4.5. The performance is improved by our proposed schemes with the MSR and PF objectives, which are designed aiming at enhancing throughput. However, in order to increase the data rates for the low-bitrate users with the MMR policy, the benefits of the users in uncrowded cells may be reduced by sharing their radio resources with the transferred lower-bound users. Therefore, the MMR-based scheme reduces the mean data rates slightly in case 1 and 3 in comparison to the Max-BRP scheme.

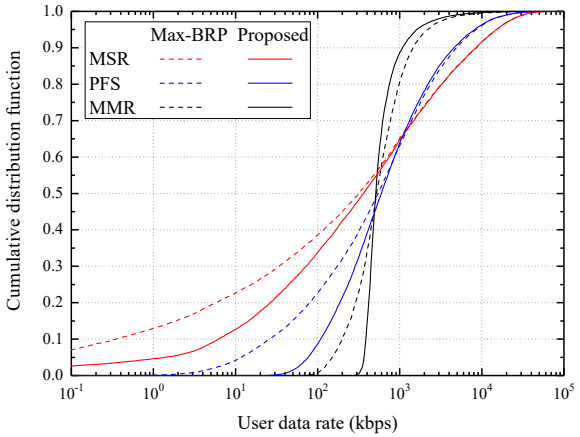
On the other hand, the overall throughput is much higher in case 2, where some users are cluster-distributed in femtocells. This is due to the better channel conditions of the femtocell users who are closer to their serving BSs than those with the entire PPP distributions. In addition, the regular hexagonal grid pattern of macrocell deployment in case 3 helps to improve the throughput by relief of the inter-cell interference at the edge areas of the macrocells. Thus, actual user distributions are crucial for guiding the deployment of HCNs. Specifically, it is necessary to plan the macrocells evenly to provide good large-scale coverage of the wireless network, while the femtocells act as supplementary access points for throughput improvement in



(a) Case 1.



(b) Case 2.



(c) Case 3.

Figure 4.4: The CDFs of user data rates with the Max-BRP and proposed traffic load balancing schemes.

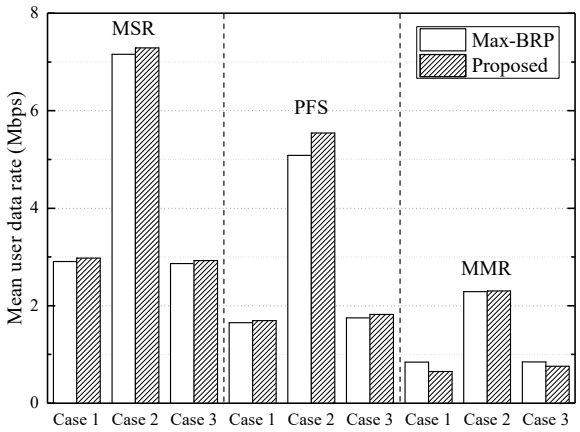


Figure 4.5: Mean user data rates with the Max-BRP and proposed traffic load balancing schemes.

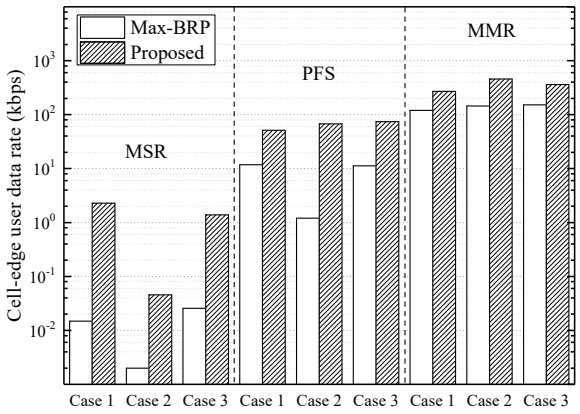
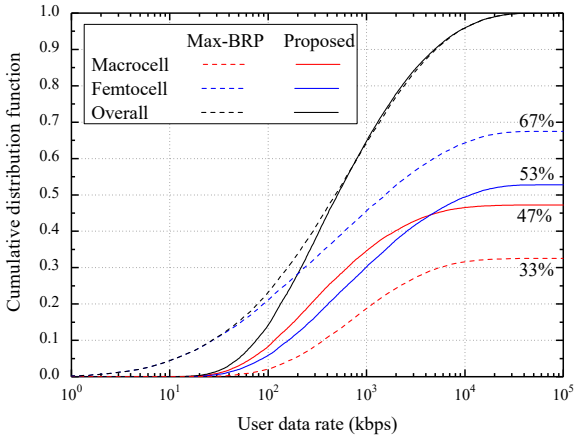


Figure 4.6: Cell-edge user data rates with the Max-BRP and proposed traffic load balancing schemes.

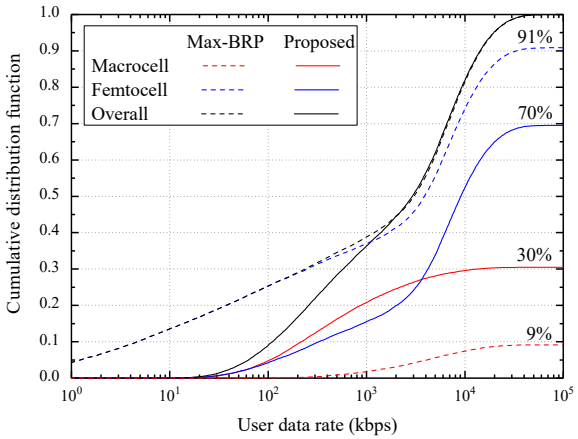
small user clusters.

We also compute the cell-edge user data rate that is defined as the 5th percentile of the lowest user data rate in the network, as shown in Figure 4.6. The MMR scheduler results in the highest cell-edge throughput. In contrast, the cell-edge performance is much worse under MSR scheduling, especially in the case of the uneven user distribution. However, this defect is relieved significantly by our proposed traffic load balancing schemes. For instance, in case 2, the cell-edge data rate is improved largely and better than the results obtained by the Max-BRP scheme in case 1 and 3. Combining all results shown in Figure 4.5 and 4.6, it is evident that the traffic load balancing scheme with the PF objective achieves a better balance between the overall and cell-edge throughput in comparison to the MSR- and MMR-based ones.

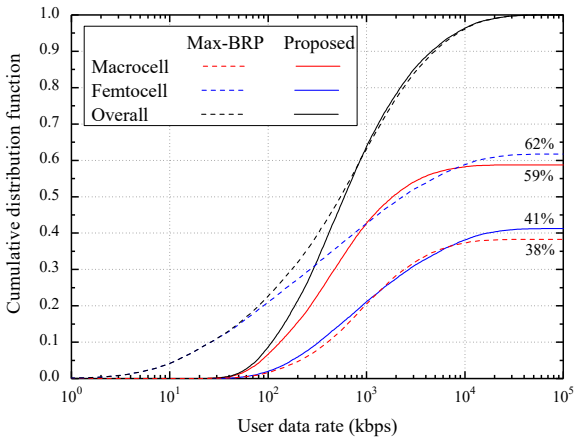
In order to investigate the traffic load transferred by our proposed scheme, we compute the CDFs of user data rates in macrocells and femtocells under PFS in different cases, as shown in Figure 4.7. Our proposed scheme is able to balance the traffic loads across the tiers and cells more effectively than the Max-BRP scheme.



(a) Case 1.



(b) Case 2.



(c) Case 3.

Figure 4.7: The CDFs of user data rates in different tiers of the HCN (PFS).

Table 4.3: Simulation Parameters of a Realistic HCN

Parameter		Value
Bandwidth		5 MHz @ 2GHz
BS Transmit power	Macrocell	43 dBm
	Femtocell	23 dBm
Path loss	Macrocell	140.7+36.7lg(d[km])
	Femtocell	128.1+37.6lg(d[km])
Standard deviation of shadowing	Macrocell	8 dB
	Femtocell	12 dB
Minimum distance	Macrocell	20 m
	Femtocell	2 m
Femtocell radius		30 m
Number of RBs (K)		25
Frame duration (T_s)		1 ms
Number of subcarriers each RB (N_{sc})		12
Number of effective OFDM symbols (S_e)		10 per frame
Minimum staying time (T_h)		1,000 ms

For instance, in case 1, 33% and 67% users access macrocells and femtocells with the Max-BRP scheme, respectively, while they are 47% and 53% by using our proposed scheme. In case 2, half of the users are cluster-distributed around the femtocell BSs, leading to more users accessing femtocells and consequently much heavier traffic load in the second tier. Therefore, the shortage of radio resources in femtocells results in a larger number of low-bitrate users. Nevertheless, our proposed scheme appropriately transfers the low-rate users in femtocells to macrocells so that the users around the hotspots can obtain more radio resources and better qualities of services. In case 3, the macrocell BSs are deployed in a hexagonal grid pattern, providing better large-scale coverage than the randomly located counterparts. Hence, more users prefer to access the macrocells due to the better channel qualities offered by them. The performance comparison between the Max-BRP and our proposed schemes indicates that the optimization of user association based on our user data rate estimation outperforms the conventional cell range control method.

4.4 An Application of the Proposed Scheme

In this section, we further verify the superiority of our traffic load balancing scheme with realistic BS distributions. The site locations in an 18x18 km^2 square area in the city of Los Angeles are adopted for simulations, as shown in Figure 4.8 [133]. We use the locations of the Wi-Fi sites as substitutions for the femtocell BSs. They have relatively lower transmit power and smaller coverages that are overlapped with the macrocells. All macrocell BSs keep working while femtocell BSs are activated randomly. The data rates of the users in the central 37 macrocells are computed and

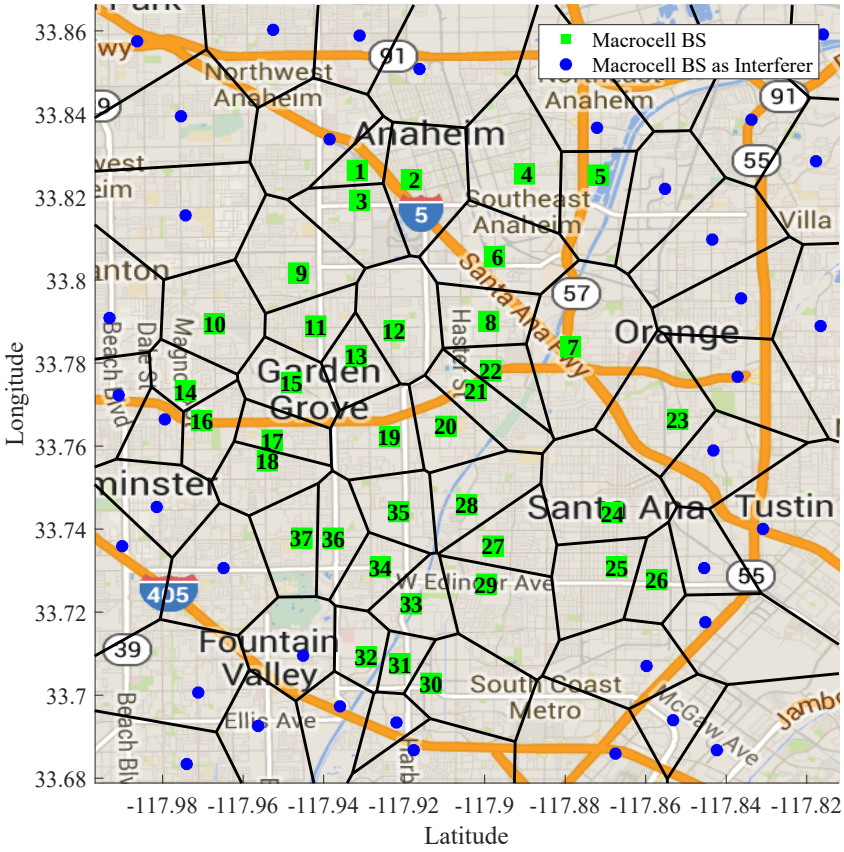
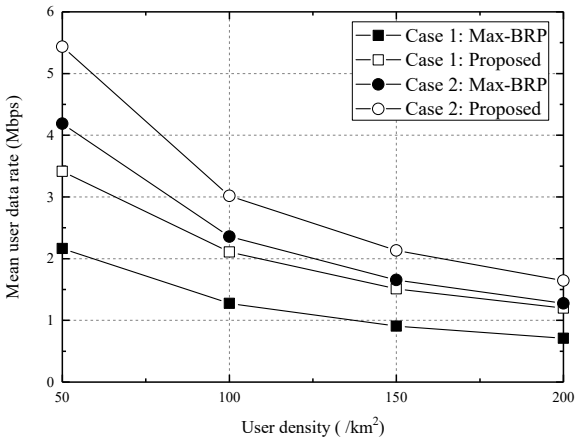


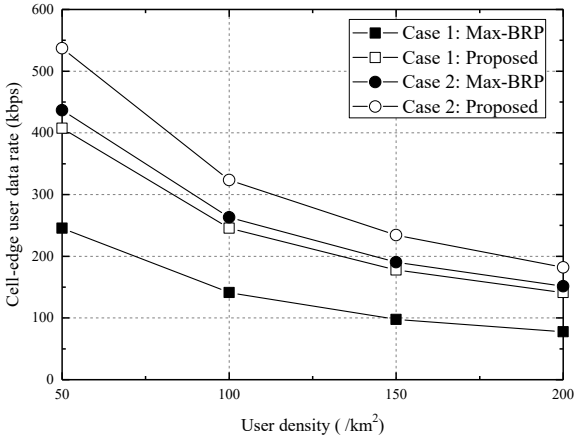
Figure 4.8: The macrocell BS locations in a square area of Los Angeles (Map source and copyright Google Maps, Google Inc.).

the BSs in other macrocells perform only as interference sources. Two different cases of user distributions are assumed. In case 1, users are all uniformly located. In order to simulate hotspots, half of the users are cluster-distributed around femtocells using the Matern-cluster process in case 2 [132]. The simulation parameters are listed in Table 4.3. The PF-based traffic load balancing scheme is utilized due to its good balance between the overall and cell-edge throughput as we clarified in Section 4.3.

Figure 4.9 presents the mean user data rates under different user densities. The number of activated femtocells is set to 3,000. The results of the Max-BRP scheme are obtained with the optimal bias factors as we explained in Section 4.3. Our proposed traffic load balancing scheme improves both of the mean and cell-edge user data rates. The mean user data rate decreases while there are more users in the network. This is because of the fewer obtainable radio resources per user. However, the decline of the mean data rate is not inversely proportional to the increasing user density. This is attributed to the multi-user diversity gain brought by using PFS that is an opportunistic scheduling scheme [36]. The system performance in case 2 is better than that in case 1 due to the cluster-distributed users with better channel qualities in the femtocells.



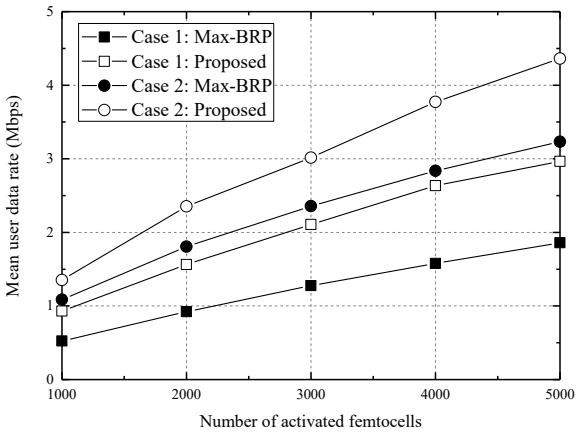
(a) Mean user data rate.



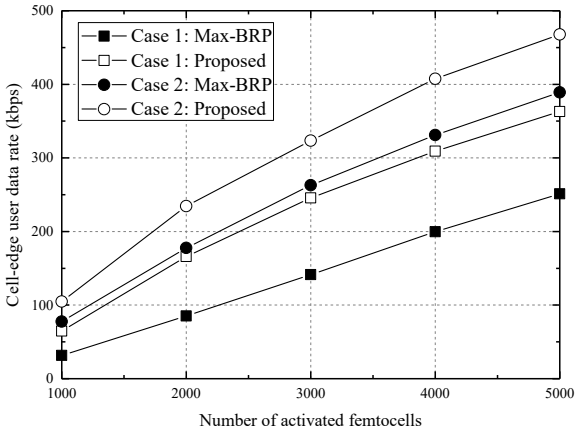
(b) Cell-edge user data rate.

Figure 4.9: User data rates under different user densities (3,000 activated femtocells).

Figure 4.10 shows the user data rates under different numbers of activated femtocells. The user density is $100/\text{km}^2$. The simulation results confirm again the superiority of our proposed traffic load balancing scheme. In addition, more activated femtocells in the network can provide dense access points and increase the spectrum reuse factor in the limited area. Every user has the opportunity to access a nearer femtocell BS and obtains more radio resources on average. Therefore, the throughput performance increases with the number of activated femtocells. However, inter-cell interference also increases while there are a large number of femtocell BSs working simultaneously. This reduces the additional benefit of denser activated femtocells and results in a nonlinear increment of the performance. Thus, considering the extra expense for deploying more femtocells in the HCN as well as an increased operating cost, it is necessary to plan the cells in small sizes in a targeted manner according to the actual user and application distributions in practice.



(a) Mean user data rate.



(b) Cell-edge user data rate.

Figure 4.10: User data rates under different numbers of activated femtocells (100 users /km^2).

4.5 Summary

In this chapter, we applied our estimation results of user data rates to designing traffic load balancing schemes in HCNs. In order to improve different performance objectives, three handover factors are designed for our online user transferring algorithms based on the corresponding resource schedulers, respectively. With the MSR target, the user handover operation depends on the corresponding improvement of the overall throughput in the network. With the aim of user fairness, the MMR scheduler is adopted and low-bitrate users are given priorities to transfer to the light-load cells so that the differences in user data rates can be reduced. The PF-oriented scheme utilizes the PFS scheme for resource allocation and improves the logarithmic sum of the user data rates.

The simulation results indicate that our proposed traffic load balancing schemes outperform the conventional Max-BRP scheme significantly. It effectively reduces

the traffic load imbalance among the macrocells and femtocells, and consequently enhances the data rates of cell-edge users. In addition, we utilize the realistic BS locations in the simulations for the performance evaluations in order to verify the superiority of our proposed scheme in actual networks. The simulation results confirm that our proposed scheme achieves a good balancing effect. Moreover, the simulation results also indicate that the traffic load and throughput in HCNs depend much on the user distributions and the deployment of BSs. Therefore, to improve the coverage and quality of service offered by HCNs, it is necessary to consider the actual application scenarios for the optimization of user association and handover.

5 Data Rate Estimation Under Bursty On-Off Traffic Flows

The rapid development of wireless communication systems enhances the transmission data rate significantly, prompting various emerging mobile services and applications. They have diverse traffic flows and bring about new requirements for system performance. For instance, the voice over IP (VoIP) service has the characteristics of bursty and persistence. It requires long-term stable connectivity, a certain level of bitrate, and a short end-to-end delay. In contrast, web browsing and instant message services have lower requirements on the delay index. Besides, it is desired to enhance the throughput performance in some real-time video transmission services for improving the picture quality. Therefore, the performance analysis based on different service types and their requirements is necessary for system optimization and providing a high quality of service (QoS).

In Chapter 3, we focus on the data rate performance analysis of different dynamic resource allocation schemes based on the saturated traffic model. It serves as a benchmark performance indicator to evaluate the maximum transmission capability of a wireless network. In comparison to the saturated traffic flows, the bursty traffic model is more realistic, especially for the performance analysis of dynamic service behaviors.

One typical bursty traffic model is the bursty on-off traffic flow. A user has data to be transmitted during the *on* period and is idle during the *off* period. This cycle can be continuous for many rounds. The lengths of *on* and *off* periods are determined by the type of user application as well as its traffic load. This traffic model can be applied to a number of streaming services at the session level, such as the video on demand (VOD). To offer a high QoS, it is desired to improve the data rate for each flow when they are active during the *on* periods. In addition, it is also necessary to guarantee fairness among multiple flows.

In this chapter, we extend our analytical work in Chapter 3 and focus on the performance analysis of dynamic resource allocation schemes under the bursty on-off traffic flows. It is a significant study for guiding the optimization and application of various schemes in the increasingly popular wireless streaming services. In the literature, there has not been any study on this issue to the best of our knowledge. The main difficulty of this performance analysis lies in the dynamic communication behaviors of the multiple user data flows. Some resource scheduler operates according to the historical scheduling results that are closely correlated to the scheduled user set. However, the active user set keeps changing due to the dynamic on-off process. Therefore, it is necessary to consider jointly the random traffic flows and time-varying user channel states in the performance analysis.

The rest of this chapter is organized as follows. Firstly, we elaborate on the bursty on-off traffic model in Section 5.1. Then, we derive the ergodic user data

rate with various scheduling schemes under bursty on-off traffic flows, including round-robin, max-min rate, and max-sum rate, in Section 5.2 to 5.4. We apply our analytical performance to the user data estimation and utilize the results obtained by system-level simulations to verify our analysis. In Section 5.5, we adopt the Gaussian approximation (GA) method to simplify the data rate estimation of the PFS scheme. We design a hybrid approximation method combining our multi-interference analysis in the case of saturated traffic flows and the GA method in order to improve the accuracy of data rate estimation. The impacts of traffic loads on various scheduling schemes are investigated as well. Finally, this chapter is summarized in Section 5.6.

5.1 Bursty On-Off Traffic Flows

The bursty traffic flow is modeled as a semi-Markov on-off process that is independent and identically distributed (i.i.d.) for different users [134]. We assume that the users can fully utilize the link capacity during their session periods. Therefore, they always have data to transmit when they are active in the *on* states.

The Pareto distribution is used for modeling the duration of the *on* state. It is expressed as

$$F_{\text{on}}(d) = 1 - \left(\frac{\beta}{d}\right)^{\alpha}, \quad d \geq \beta, \tag{5.1}$$

where d is the duration variable with a unit of second, $\beta > 0$ is the minimum period of the *on* state, and $\alpha > 1$ is the shape parameter. Thus, the mean duration of the *on* state is

$$D_{\text{on}} = \frac{\alpha\beta}{\alpha - 1}. \tag{5.2}$$

The duration of the *off* state is modeled by the exponential distribution, which is expressed as

$$F_{\text{off}}(d) = 1 - \exp(\lambda d), \quad d > 0, \tag{5.3}$$

where $\lambda > 0$ is the rate parameter. Thus, the mean *off* state period is

$$D_{\text{off}} = \lambda^{-1}. \tag{5.4}$$

The traffic load is defined as the duty cycle of the on-off process. It is calculated as

$$\rho = \frac{D_{\text{on}}}{D_{\text{on}} + D_{\text{off}}} = \left[1 + \frac{\alpha - 1}{\alpha\beta\lambda}\right]^{-1}. \tag{5.5}$$

Specifically, when the parameter $\rho \to 1$, all of the users always keep active and the traffic load in the system is consequently saturated.

5.2 Data Rate Analysis of the RR Scheduler

Under bursty on-off traffic flows, the active user set is dynamic due to the continuous changes in the user states. Thus, the obtained data rata of a user is influenced by the time-varying active user set. To analyze the ergodic user data rate in the long run, we utilize a semi-static approximation method introduced as follows.

In general, the session duration is on a scale of seconds or minutes for the online streaming services [134]. Thus, the change of the on-off state is at a much lower speed in comparison to the dynamic radio resource scheduling. Under this condition, the active user set keeps steady in a relatively long period that covers a large number of short scheduling frames. Therefore, the mean user data rate obtained during this period is approximately converged. We calculate the ergodic data rate of a user under every possible combination of active users based on our analysis in the case of saturated traffic flows. Then, the long-term mean data rate of a user is computed by the weighted sum of its data rates obtained in different active user sets in terms of their corresponding probabilities.

We first derive the ergodic user data rate with a basic scheduling scheme, namely, round-robin (RR). Its analytical performance is tractable since it uses neither historical scheduling records nor instantaneous CSI. Thus, it serves as a benchmark and is compared to the other scheduling schemes that will be analyzed in the following sections.

We consider the user set $\mathbf{U}_b$ that is associated with BS b. The number of users in it is denoted as $U_b = |\mathbf{U}_b|$. The probability of an active user set $\mathbf{V} \subseteq \mathbf{U}_b$ is calculated as

$$\Pr\{\mathbf{V}_{on} = \mathbf{V}\} = \rho^{|\mathbf{V}|}(1-\rho)^{|\mathbf{U}_b|-|\mathbf{V}|}. \tag{5.6}$$

As we introduced in Chapter 2, the RR scheduler distributes the same amount of radio resources to the multiple users in the long run. Therefore, the ergodic data rate of a user $u \in \mathbf{U}_b$ while it is active can be calculated as

$$\overline{r}_{u,rr}(\rho) = \rho^{-1} \sum_{\mathbf{V} \subseteq \mathbf{U}_b}^{u \in \mathbf{V}} \frac{\hat{r}_u}{|\mathbf{V}|} \Pr\{\mathbf{V}_{on} = \mathbf{V}\}, \tag{5.7}$$

where $\hat{r}_u$ is the mean data rate of user u per RB, as given by (3.40). Note that the sum in the formula includes all possible user subsets $\mathbf{V} \subseteq \mathbf{U}_b$ that contain user $u \in \mathbf{V}$. The radio resources are equally distributed to the users in each of them.

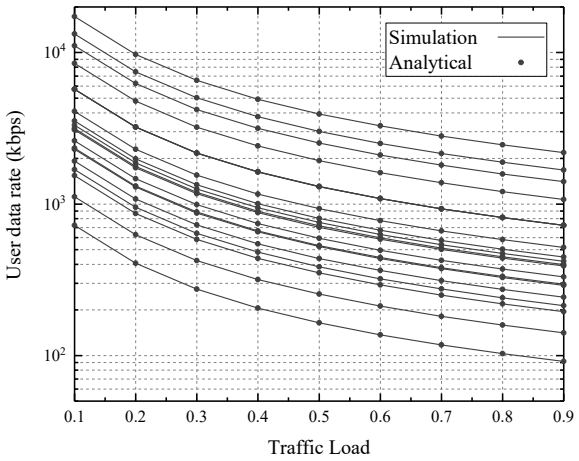


Figure 5.1: Simulation and analytical results of user data rates under bursty on-off traffic flows (RR, 20 users).

This formula can be further simplified as follows.

$$\begin{aligned}\overline{r}_{u,rr}(\rho) =& \rho^{-1} \sum_{n=0}^{U_b-1} \left(\begin{array}{c} n \\ U_b-1 \end{array} \right) \frac{\hat{r}_u}{n+1} \rho^{n+1} (1-\rho)^{U_b-n-1} \\ =& \frac{\hat{r}_u}{\rho U_b} \sum_{n=0}^{U_b-1} \left(\begin{array}{c} n+1 \\ U_b \end{array} \right) \rho^{n+1} (1-\rho)^{U_b-n-1} \\ =& \frac{\hat{r}_u}{\rho U_b} \left[1-(1-\rho)^{U_b} \right] \end{aligned} \tag{5.8}$$

The item in the square brackets is the probability that the system is not idle while one or more users are active. Therefore, it implies that the associated users equally share the system bandwidth while it is busy.

Specially, when the traffic load $\rho \to 1$, we have

$$\overline{r}_{u,rr}(1) = \frac{\hat{r}_u}{U_b}, \tag{5.9}$$

which is the identical result as in the saturated traffic flow case.

In contrast, when the traffic load $\rho \to 0$, it is rare that more than one user is active concurrently. Therefore, the ergodic user data rate becomes

$$\overline{r}_{u,rr}(\rho)\big|_{\rho \to 0} \approx \frac{\hat{r}_u}{\rho U_b} \left[1-(1-U_b\rho) \right] = \hat{r}_u. \tag{5.10}$$

This means that each user occupies the whole system bandwidth during its *on* state period while the user active probability is very low.

We apply our performance analysis to the user data rate estimation under bursty on-off traffic flows and carry out system-level simulations for its verification. The system configurations and deployment of the urban scenario in Section 3.5 are

utilized. The weighted-sum SINR model is adopted with maximum 8 neighbor cells reported per user. We set the parameters $\alpha = 1.5$ and $\beta = 10\,s$ for the *on* state. Hence, its mean duration is $D_{on} = 30\,s$. The rate parameter λ of the *off* state is set according to the test traffic load $\rho \in [0.1, 0.9]$.

In Figure 5.1, we present the data rates of 20 randomly distributed users under different traffic loads. The ergodic user data rates decrease as the traffic load increases due to more users are active simultaneously and share the limited system bandwidth. In addition, our analytical results are consistent with the simulation ones as shown in the figure, confirming the feasibility of using them for user data estimation under bursty traffic flows.

5.3 Data Rate Analysis of the MMR Scheduler

In the case of saturated traffic flows, the max-min rate (MMR) scheduler allocates the radio resource to the user with the lowest long-term averaged data rate. In this way, it maintains an approximately equal data rate level for the multiple users associated with one BS in the long run. Under bursty on-off traffic flows, this fair principle is applied to the active user sessions. Specifically, the radio resource is allocated to the active user that has the lowest bitrate for its session flow. Therefore, the active users obtain approximately the same data rates.

The average data rate in an active user subset $\mathbf{V} \subseteq \mathbf{U}_b$ can be calculated as

$$\overline{r}_{mmr}\left(\mathbf{V}\right) = \left[\sum_{v \in \mathbf{V}} \hat{r}_v^{-1}\right]^{-1}. \tag{5.11}$$

This analytical result relies on the fact the average user data rate converges during the *on* period due to the relatively slow changes in the user on-off states as we have discussed.

Thus, the ergodic data rate of a given user $u \in \mathbf{U}_b$ is derived as follows.

$$\begin{aligned}\overline{r}_{u,mmr}\left(\rho\right) =& \rho^{-1} \sum_{\mathbf{V} \subseteq \mathbf{U}_b}^{u \in \mathbf{V}} \left[\sum_{v \in \mathbf{V}} \hat{r}_v^{-1}\right]^{-1} \Pr\left\{\mathbf{V}_{on} = \mathbf{V}\right\} \\ =& \sum_{\mathbf{V} \subseteq \mathbf{U}_b}^{u \in \mathbf{V}} \frac{\rho^{|\mathbf{V}|-1}(1-\rho)^{|\mathbf{U}_b|-|\mathbf{V}|}}{\sum\limits_{v \in \mathbf{V}} \hat{r}_v^{-1}}\end{aligned} \tag{5.12}$$

We compare the estimated user data rates based on our analytical results to the ones obtained by simulations under different traffic loads, as shown in Figure. 5.2. Their high consistency verifies the correctness of our analysis. In addition, the ergodic user data rates approach the same level as the traffic load increases. The overall throughput gain of MMR over RR is shown in Figure 5.3. It decreases with the traffic load due to the fair scheduling of MMR.

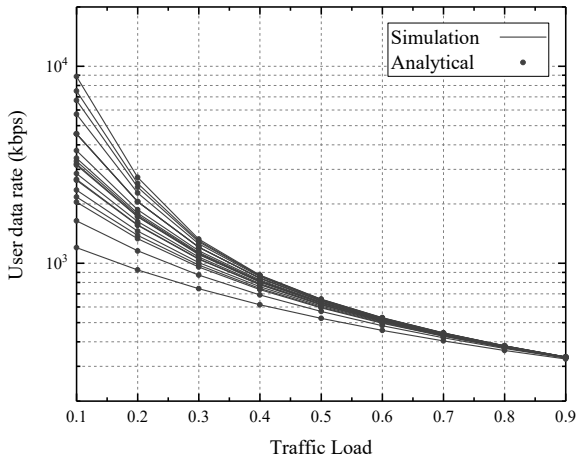


Figure 5.2: Simulation and analytical results of user data rates under bursty on-off traffic flows (MMR, 20 users).

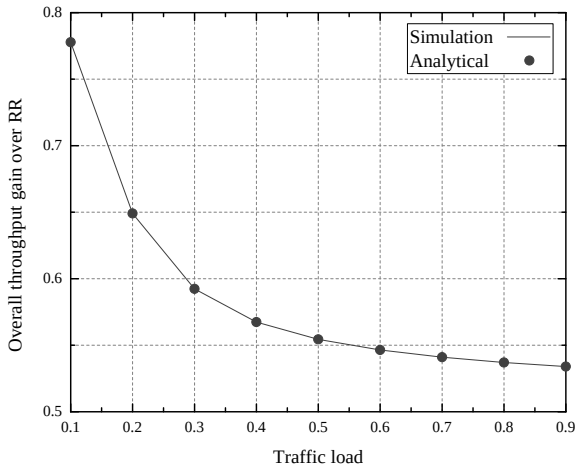


Figure 5.3: Overall throughput gain of the MMR scheduling scheme over RR (20 users).

When the traffic load is low, it is rare that multiple user sessions are activated concurrently. Thus, the MMR scheduler operates on one single active user most of the time. Under this condition, it loses the effect of keeping fairness among users and obtains similar scheduling results as the RR scheduler, i.e.,

$$\overline{r}_{u,mmr}\left(\rho\right)\big|_{\rho\to 0} \approx \hat{r}_u. \tag{5.13}$$

In contrast, as the traffic load increases, more users are active and scheduled by MMR simultaneously, resulting in better user fairness but lower overall throughput. Therefore, when the traffic load is heavy and close to 1, we have the identical result

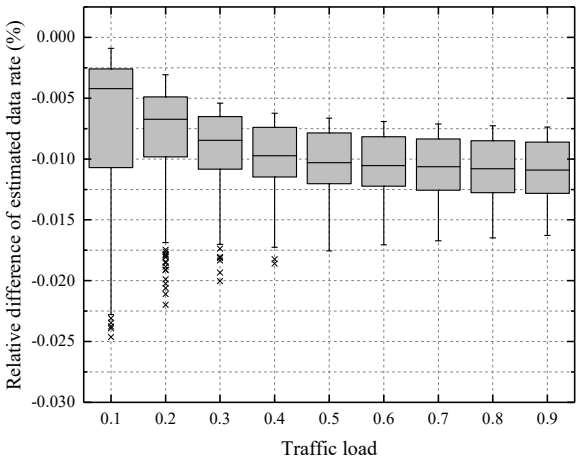


Figure 5.4: Relative differences between the estimated user data rates and simulation results (MMR, 20 users).

as in the saturated traffic case, i.e.,

$$\overline{r}_{u,mmr}(1) = \left[\sum_{u \in \mathbf{U}_b} \hat{r}_u^{-1}\right]^{-1}. \tag{5.14}$$

We compute the relative differences between our estimated user data rates and the simulation results, shown by the Tukey's box plots as in Figure. 5.4. The stochastic channel modeling with the incomplete CSI causes inevitable tiny errors in the data rate estimation, leading to slightly lower results than the actual ones. However, they are extremely close since the analytical results for MMR scheduling is nonsensitive to the probability distributions of user SINRs as we analyzed in Section 3.5. The relative differences are larger while the traffic load is at a lower level because very few users are active simultaneously under this condition. The period during which multiple users are scheduled by MMR simultaneously is short, resulting in relatively less convergence of the scheduling algorithm. However, this effect is negligible for practical usage while the overall estimation error is very small as shown in the figure.

5.4 Data Rate Analysis of the MSR Scheduler

Similar to that in the case of saturated traffic flows, the radio resource is allocated by the MSR scheduler to the user that is in the active user set and has the highest obtainable data rate. We consider a given user u in the active user set. It is scheduled instead of another user $v \in \mathbf{U}_b$ due to two possible reasons: if user v is in the *off* state, it has no data to be transmitted and thus is not a candidate user to be scheduled by the MSR scheduler; or if user v is in the *on* state but has a lower SINR and consequently a lower data rate than user u. Combining the above two cases in terms of their probabilities and considering all of the users in $\mathbf{U}_b$, the probability

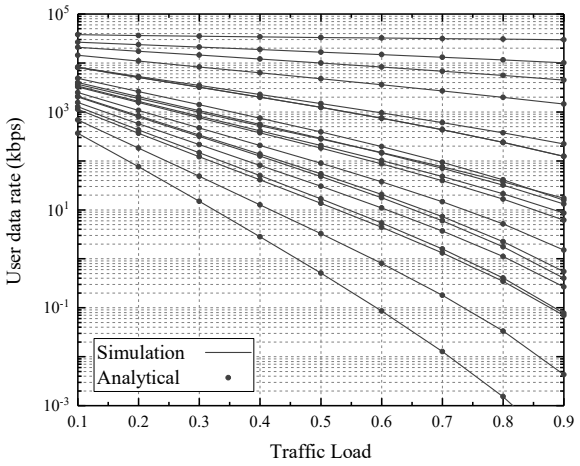


Figure 5.5: Simulation and analytical results of user data rates under bursty on-off traffic flows (MSR, 20 users).

distribution of the scheduled SINR while user u is in the *on* state is calculated as

$$f_{u,msr}(z,\rho) = f_{\Phi_u}(z) \prod_{v \in \{\mathbf{U}_b/u\}} [\rho F_{\Phi_v}(z) + (1-\rho)], \tag{5.15}$$

where $F_{\Phi_u}(z)$ and $f_{\Phi_u}(z)$ are cumulative distribution function (CDF) and probability density function (PDF) of the user SINR that have been derived in Section 3.2. Based on this PDF of the scheduled SINR and the data rate mapping function in (3.39), the ergodic user data rate under MSR scheduling and bursty on-off traffic flows is calculated as

$$\overline{r}_{u,msr}(\rho) = \int_0^\infty r(z) f_{u,msr}(z,\rho)\, dz. \tag{5.16}$$

In Figure 5.5, we present the simulation and analytical results of the user data rates under MSR scheduling. In contrast to the trend in the case of using MMR, the differences among the ergodic user data rates enlarge as the traffic load increases. This is attributed to the opportunistic scheduling mechanism of MSR. While there are multiple users activated concurrently, the user with the highest SINR is scheduled with the aim of maximizing overall throughput. Thus, the users with poor channel conditions are starved of radio resources. However, in the case of a lower traffic load, each user is more likely active alone. Under this condition, the radio resource is allocated to the only active user in the cell. Thus, a low-SINR user is able to obtain a larger bandwidth than in the heavy traffic load case. Specifically, the probability distribution of the scheduled SINR under a very low traffic load is approximately

$$f_{u,msr}(z,\rho)\big|_{\rho \to 0} \approx f_{\Phi_u}(z). \tag{5.17}$$

Therefore, the ergodic user data rate is identical to the one under RR scheduling,

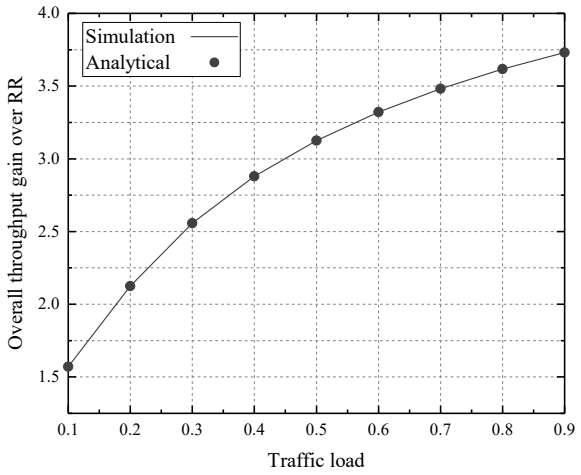


Figure 5.6: Overall throughput gain of the MSR scheduling scheme over RR (20 users).

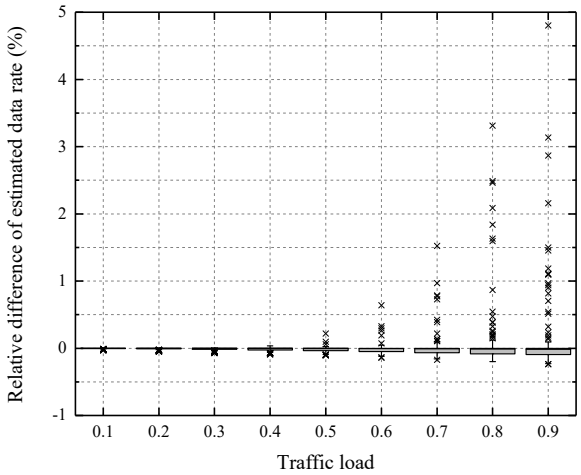


Figure 5.7: Relative differences between the estimated user data rates and simulation results (MSR, 20 users).

i.e,

$$\overline{r}_{u,msr}\left(\rho\right)\big|_{\rho\to 0} \approx \int_{0}^{\infty} r\left(z\right)dz = \hat{r}_u. \tag{5.18}$$

In Figure 5.6, we present the overall throughput gain of MSR over RR. It increases with the traffic load due to the larger multi-user diversity gain brought by the opportunistic scheduling while there are more users active and scheduled by MSR simultaneously. Hence, the MSR scheduler is able to improve the aggregated user data rate significantly at the expense of poor user fairness under a heavy traffic load.

In Figure 5.7, we present the relative differences between our estimated user data rates and the simulation results. They are less than 5% and keep at a very low level. The estimation is less accurate under the heavy traffic load. As we analyzed

in Section 3.5, the inaccuracy of SINR models has considerable influence on the data rate estimation under MSR scheduling. As shown in (5.15), the probability distributions of user SINRs have an increasing effect on the PDFs and CDFs of the scheduled SINRs as the traffic load parameter ρ becomes larger.

5.5 Data Rate Estimation of PFS

The PFS scheme relies on both of the instantaneous channel states and the historical scheduling results. Thus, its data rate analysis is more complex than the MSR and MMR schedulers under bursty traffic flows. We utilize again the semi-static approximation method to analyze its ergodic user data rates as follows.

We assume that a subset $\mathbf{V} \in \mathbf{U}_b$ is the active user set at a point in time. For a given user $u \in \mathbf{V}$, we denote $G_u(\mathbf{V})$ as its data rate gain by using PFS over RR under saturated traffic flows. Similar to the case of RR scheduling, the ergodic user data rate can be calculated as

$$\begin{aligned}\overline{r}_{u,pfs}(\rho) =& \rho^{-1} \sum_{\mathbf{V}\subseteq\mathbf{U}_b}^{u\in\mathbf{V}} \frac{\hat{r}_u G_u(\mathbf{V})}{|\mathbf{V}|} \Pr\{\mathbf{V}_{on} = \mathbf{V}\} \\ =& \hat{r}_u \sum_{\mathbf{V}\subseteq\mathbf{U}_b}^{u\in\mathbf{V}} \frac{G_u(\mathbf{V})}{|\mathbf{V}|} \rho^{|\mathbf{V}|-1}(1-\rho)^{|\mathbf{U}_b|-|\mathbf{V}|}. \end{aligned} \tag{5.19}$$

However, when the total number of users U_b is large, the computational complexity of this formula is very high since $G_u(\mathbf{V})$ is different for each subset $\mathbf{V}$ that satisfies $u \in (\mathbf{V} \subseteq \mathbf{U}_b)$. Therefore, it needs to be calculated independently for each of the $2^{(|\mathbf{U}_b|-1)}$ possible cases. In order to make the analysis tractable, there have been some research making efforts to simplify the channel models by introducing the symmetry to different users, e.g., the GA method [34].

In the rest of this section, we first derive the ergodic user data rates based on GA. In order to improve the data rate estimation accuracy with lower computational complexity, we then design a hybrid approximation model by carefully combining GA and our analytical results in the case of the saturated traffic flows.

5.5.1 Gaussian Approximation Method

With the aim of tractable analysis, $G_u(\mathbf{V})$ can be approximated so that it depends only on the number of users instead of the specific users in set $\mathbf{V}$. To this end, the instantaneous user data rates are modeled with the Gaussian distribution approximately in [33, 34]. Based on this model, the user data rate gain of PFS over RR under saturated traffic flows is approximately

$$\begin{aligned}\tilde{G}_{u,ga}(\mathbf{V}) =& 1+\frac{\hat{\sigma}_u}{\hat{r}_u}\left\{|\mathbf{V}|\int_{-\infty}^{\infty} z\frac{e^{-z^2/2}}{\sqrt{2\pi}}\left[F_{(0,1)}(z)\right]^{|\mathbf{V}|-1}dz\right\}\\ =& 1+\frac{\hat{\sigma}_u}{\hat{r}_u}\int_0^1 z d\left[F_{(0,1)}(z)\right]^{|\mathbf{V}|},\end{aligned} \tag{5.20}$$

where $F_{(0,1)}(z)$ is the CDF of the standard normal distribution, and $\hat{\sigma}_u$ is the standard deviation of the user data rate that is calculated based on the SINR model of user u.

Note that the integral part in (5.20) is a function which only depends on the number of active users. For ease of expression, we denote it as $L(N)$, i.e.,

$$L(N) = \int_0^1 z d\left[F_{(0,1)}(z)\right]^N. \tag{5.21}$$

Then, the estimated data rate gain in (5.20) is rewritten as

$$\tilde{G}_{u,ga}(\mathbf{V}) = 1+\frac{\sigma_u}{r_u}L(|\mathbf{V}|). \tag{5.22}$$

Substituting this formula into (5.19), we can solve the closed-form expression of the estimated user data rate by GA as follows,

$$\begin{aligned}\tilde{r}_{u,ga}(\rho) =& \sum_{u\in(\mathbf{V}\subseteq\mathbf{U}_b)}\left[\frac{\hat{r}_u\tilde{G}_{u,ga}(\mathbf{V})}{|\mathbf{V}|}\rho^{|\mathbf{V}|-1}(1-\rho)^{|\mathbf{U}_b|-|\mathbf{V}|}\right]\\ =& \hat{r}_u\sum_{n=1}^{|\mathbf{U}_b|}\left[\begin{pmatrix}|\mathbf{U}_b|-1\\ n-1\end{pmatrix}\frac{1+\hat{\sigma}_u L(n)}{n\hat{r}_u}\rho^{n-1}(1-\rho)^{|\mathbf{U}_b|-n}\right]\\ =& \frac{\hat{r}_u}{\rho|\mathbf{U}_b|}\left[1-(1-\rho)^{|\mathbf{U}_b|}\right]+\frac{\hat{\sigma}_u}{\rho|\mathbf{U}_b|}l(|\mathbf{U}_b|,\rho),\end{aligned} \tag{5.23}$$

where $l(N,\rho)$ represents

$$l(N,\rho) = \sum_{n=1}^{N}\left[\begin{pmatrix}N\\ n\end{pmatrix}\rho^n(1-\rho)^{N-n}L(n)\right]. \tag{5.24}$$

In particular, under saturated traffic flows, i.e., $\rho = 1$, we have

$$l(N,1) = L(N). \tag{5.25}$$

According to (5.23) and (5.7), we compute the performance gain of PFS over RR

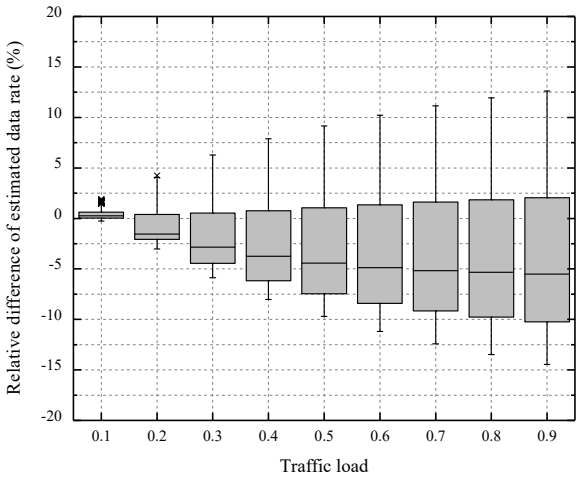


Figure 5.8: Relative differences of the GA-based data rate estimation under different traffic loads (PFS, 20 users).

under bursty on-off traffic flows as

$$\tilde{g}_{u,ga}\left(\mathbf{U}_b,\rho\right)=1+\frac{\hat{\sigma}_u}{\hat{r}_u}l\left(\left|\mathbf{U}_b\right|,\rho\right)\left[1-\left(1-\rho\right)^{\left|\mathbf{U}_b\right|}\right]^{-1}. \tag{5.26}$$

This formula implies that the scale of user data rate variation has a significant impact on the performance gain of PFS. The opportunistic scheduling takes more advantage of the time-varying user channel states while they are fluctuating greatly. In contrast, there is less benefit brought by PFS if the user channels are very steady. In addition, under a very low traffic load, i.e., $\rho \to 0$, we have

$$\tilde{g}_{u,ga}\left(\mathbf{U}_b,\rho\right)\big|_{\rho\to 0}\approx 1, \tag{5.27}$$

which indicates that the PFS scheme achieves nearly no performance improvement in comparison to RR while the multiple users are rarely active at the same time.

The relative differences of the GA-based data rate estimation under different traffic loads are presented in Figure 5.8. When the traffic load is low, the estimation error is small. However, it increases significantly with the traffic load. For instance, the relative differences of some estimated user data rates are even larger than 10 % when the traffic load reaches 0.9. This is due to the inaccurate estimation of the GA method in the case of a large number of users.

To further verify this, we set the traffic load $\rho = 1$ and calculate the relative differences with various numbers of users per cell, as shown in Figure 5.9. The GA-based approach results in larger estimation errors as the number of users increases. Thus, under a heavy traffic load, more users are likely to be active simultaneously, leading to inaccurate data rate estimation with the GA-based analysis. The main reason for this defect lies in the symmetric simplification of the multiple user channels. Although the different relative variations of user data rates are taken into account in the GA model, it is still far from the actual probabilities of

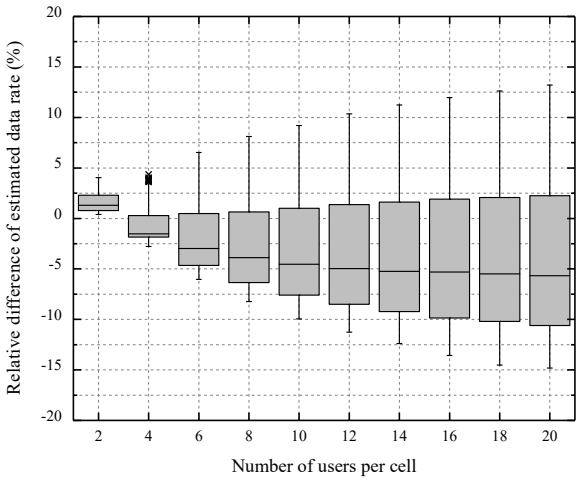


Figure 5.9: Relative differences of the GA-based data rate estimation under different numbers of users (PFS, $\rho = 1$).

user SINRs in the multi-cell network. They are normally asymmetric and dependent on the specific user locations and channel states.

5.5.2 Hybrid Approximation Method

In order to remedy the shortcomings of the GA-based analysis under a heavy traffic load, we utilize the analytical performance of PFS derived in Section 3.3.3. Different from GA, we have considered the probability distributions of independent user SINRs for performance analysis. Our stochastic channel models are developed based on multi-interference analysis (MIA) and achieve more accurate data rate estimation than the GA-based results. However, it is hard to apply the MIA-based results directly to the user data rate estimation under bursty on-off traffic flows due to the high combinatorial complexity as we explained at the beginning of this section.

In order to improve the accuracy of user data rate estimation with acceptable computational complexity, we design a hybrid approximation (HA) method by delicately combining the results obtained by GA and MIA. It has a higher tendency to use the GA-based data rate estimation under a low traffic load and to use the MIA-based ones under a high traffic load. Specifically, we formulate the estimated user data rate by HA as

$$\tilde{g}_{u,ha}\left(\mathbf{U}_b,\rho\right) = 1 + \left(1-\rho\right)\tilde{\eta}_{u,ga}\left(\mathbf{U}_b,\rho\right) + \rho\underbrace{\frac{\tilde{\eta}_{u,ga}\left(\mathbf{U}_b,\rho\right)}{\tilde{\eta}_{u,ga}\left(\mathbf{U}_b,1\right)}}_{(a)}\underbrace{\left[\overline{G}_u - 1\right]}_{(b)}, \tag{5.28}$$

where $\tilde{\eta}_u\left(\mathbf{U}_b,\rho\right)$ denotes the increment part of the PFS performance gain over RR estimated by GA, i.e,

$$\tilde{g}_{u,ga}\left(\mathbf{U}_b,\rho\right) = 1 + \tilde{\eta}_{u,ga}\left(\mathbf{U}_b,\rho\right), \tag{5.29}$$

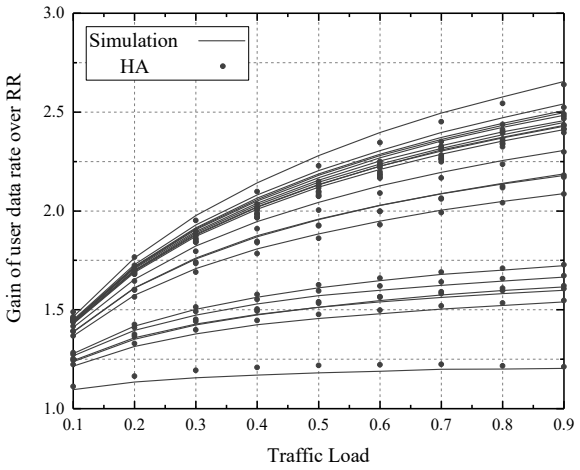


Figure 5.10: Simulation and analytical results of user data rate gains by using PFS over RR under bursty on-off traffic flows (20 users).

and $\overline{G}_u$ is the PFS performance gain calculated by MIA under saturated traffic flows and is computed as

$$\overline{G}_u = \overline{r}_{u,pfs}/\overline{r}_{u,rr}\,(1)\,. \tag{5.30}$$

It can be obtained by (3.55) and (5.9).

In this HA model, the performance gain increment estimated by GA, i.e., $\tilde{\eta}_u\,(\mathbf{U}_b, \rho)$, is weighted by $(1-\rho)$. Item (a) is a linear ratio of the performance gain increment between the unsaturated and saturated traffic flows, and item (b) is the performance gain increment estimated by MIA under saturated traffic flows. The combination of item (a) and (b) is weighted by the traffic load ρ.

Specially, we have the performance gain of PFS over RR while $\rho \to 0$ as

$$\tilde{g}_{u,ha}\,(\mathbf{U}_b, \rho)\big|_{\rho\to 0} \approx \tilde{g}_{u,ga}\,(\mathbf{U}_b, \rho)\big|_{\rho\to 0} \approx 1, \tag{5.31}$$

which implies that the performance gain of PFS is negligible under a very low traffic load. In this case, the ergodic user data rates are close to the ones under RR scheduling.

In contrast, the performance gain of PFS while $\rho = 1$ is

$$\tilde{g}_{u,ha}\,(\mathbf{U}_b, 1) = \overline{G}_u. \tag{5.32}$$

Therefore, HA yields identical results under a very high traffic load with the MIA-based ones.

In Figure 5.10, we present the simulation results of the user data rate gains by using PFS over RR and compare them to the ones obtained by our proposed HA model. The estimated results are close to the actual ones, indicating high reliability of the HA model. In addition, every user has its data rate gain larger than 1 and increasing with the traffic load due to the enhanced multi-user diversity gain while there are more concurrently active users. In comparison to the MSR and MMR schemes, PFS

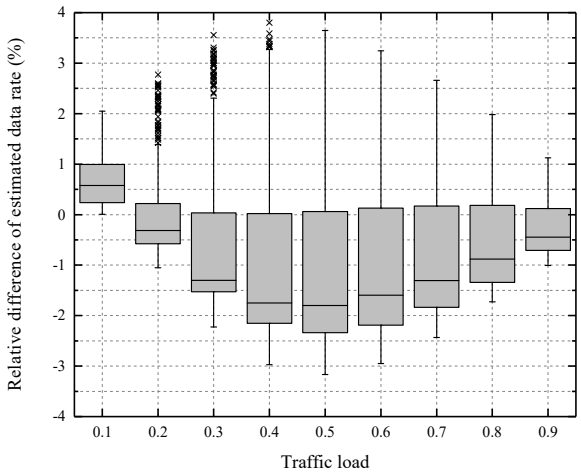


Figure 5.11: Relative differences of the HA-based data rate estimation under different traffic loads (PFS, 20 users).

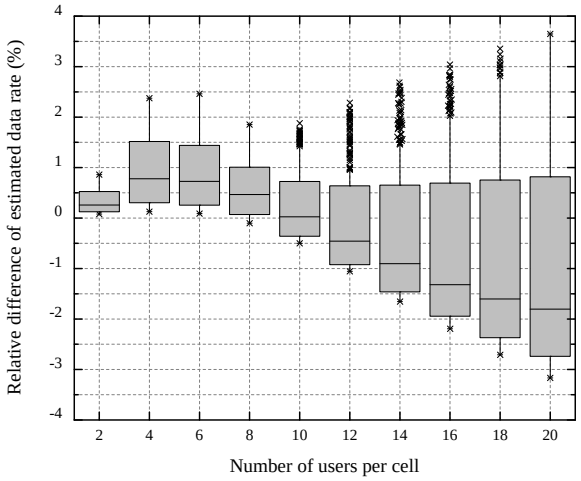


Figure 5.12: Relative differences of the HA-based data rate estimation under different numbers of users (PFS, $\rho = 0.5$).

obtains a better balance between the overall efficiency and user fairness and thus is preferable for multi-user scheduling under bursty traffic flows.

We compute the relative differences of the estimated user data rates by HA, as shown in Figure 5.11. Compared to the GA method, it achieves significant improvement in terms of estimation accuracy, especially under a high traffic load. The estimation errors are lower than 4% under various traffic loads. We further investigate the estimation accuracy with $\rho = 0.5$ since the estimation errors are relatively larger under the median traffic load as shown in Figure 5.11. The estimation results under different numbers of users are presented in Figure 5.12. When there are fewer users in the network, most of the estimated user rates are slightly larger than the actual ones due to the overestimation caused by GA under this condition, which can be observed in Figure 5.9. The estimation errors are always within the range of $\pm 4\%$

as the number of users changes, which is a significant reduction in comparison to the pure GA-based approach. Thus, the HA model is more accurate and favorable to practical user data rate estimation.

5.6 Summary

In this chapter, we extended our performance analysis of resource allocation schemes to the case of bursty on-off traffic flows. We analyzed the transmission performance of the RR, MMR, and MSR scheduling schemes. The analytical results were applied to user data rate estimation and were verified to be very accurate by simulation results. The estimated results under MSR have higher errors due to its high sensitivity to the SINR models than MMR and RR. In order to make the analytical performance of PFS tractable, we utilized the GA method and design a hybrid approximation model by delicately combining the GA-based analysis with our MIA results under saturated traffic flows. The simulation results have confirmed that it can significantly improve the accuracy of user data rate estimation.

In addition, multiple user flows are rarely active in parallel when the traffic load is very low. Hence, each user has a great chance to occupy the whole system bandwidth when it is active alone. This results in similar performance as in the case of RR even though various scheduling schemes are utilized. Therefore, under a low traffic load, multi-user scheduling schemes have only minor impacts on the system performance and can be omitted for simpler implementation in practice.

6 Dynamic Channel and Power Allocation for Single-Channel NOMA

In this chapter, we focus on the dynamic channel and power allocation (DCPA) problem for downlink single-channel NOMA (SC-NOMA) systems. We adopt the proportional fairness as the objective of DCPA. The proportional fair scheduling (PFS) has been widely accepted in NOMA systems, which can achieve a good balance between the overall and cell-edge user throughput as we have discussed. A common practice for solving the DCPA problem in SC-NOMA systems is based on the decoupled approach. Specifically, the power allocation is implemented for every candidate multiplexed user set in the first stage. Then, the optimal user set which offers the highest scheduling factor is selected in the second stage.

In the literature, there have been several power allocation and user set selection schemes proposed for the SC-NOMA systems as we have introduced in Chapter 2. However, they suffer from very high computational complexity for practical usage. In addition, the analytical performance of DCPA in SC-NOMA systems has not been studied so far but is significant for its optimization and application. Therefore, we are motivated to design low-complexity DCPA schemes and develop the analytical model for their performance analysis.

In this chapter, we consider two SC-NOMA system models defined as follows.

- ▷ *Practical SC-NOMA system*: The maximum number of multiplexed users per frame is controlled by a pre-defined parameter S due to the limited processing capability of the SIC receivers. Normally, the limitation is 2 or 3 in practice and the corresponding systems are referred to as 2-user and 3-user SC-NOMA, respectively.
- ▷ *Ideal SC-NOMA system*: We relax the limitation on the number of multiplexed users per frame in an ideal SC-NOMA system. Thus, an arbitrary number of users can be multiplexed within the same channel simultaneously.

Based on the assumption of the ideal SC-NOMA system, we derive a closed-form solution of the optimal power allocation for PFS in Section 6.2. The performance of this solution is proved to be the upper bound for PFS in practical SC-NOMA systems. According to our derivation, we design a low-complexity algorithm to jointly select the optimal multiplexed users and determine their assigned power. In order to reduce the computational complexity for user set selection (USS) in practical SC-NOMA systems, we design two USS schemes based on our optimal power allocation, including one optimal tree-searching-based USS scheme and a suboptimal preselection-based USS scheme, in Section 6.3.

In Section 6.4, we develop an analytical model of the upper bound data rate performance obtained in the ideal SC-NOMA system. The ergodic user data rates are derived based on our stochastic channel model in Chapter 3. The influence of partial channel state information (CSI) on the analytical result is studied as well. We utilize the analytical upper bound performance to estimate user data rates and overall throughput in the 2-user and 3-user practical SC-NOMA systems. In Section 6.5, the system-level simulations are carried out to verify our analysis of the upper bound performance. We also confirm that using the analytical results for user data rate estimation in practical SC-NOMA systems is feasible. Various influence factors on the estimation accuracy, including SIC limitation, the number of users, partial and imperfect CSI, are carefully investigated. In Section 6.6, we summarize this chapter at the end.

6.1 System Model

In the OMA system, only one user is assigned the bandwidth and power in each scheduling frame. The instantaneous SINR of a user u is given in (3.7) and can be further expressed as

$$\Phi_u = \frac{H_{u,b} p_t}{\sum_{i \in \mathbf{I}_u} P_{u,i} + \sigma_N}, \tag{6.1}$$

where $H_{u,b}$ is the comprehensive channel gain of user u, including the large-scale slow fading and the Rayleigh fast fading, and p_t is the transmit power budget of the downlink SC-NOMA system. For ease of expression, all of the power values in the following parts are normalized by p_t. We define the SINR in (6.1) as the instantaneous user channel quality indicator (CQI). It is the maximum achievable SINR while the user occupies the scheduling frame and full transmit power alone.

In the NOMA system, SIC is adopted to allow superposition of multiple user signals with different transmit power levels [71]. In each scheduling frame, the BS assigns the transmit power to multiple users in $\mathbf{U}_b$. Focusing on the DCPA problem for a given BS b, we neglect the subscript b hereinafter. We denote the power allocated to user u as p_u. It satisfies that

$$\sum_{u \in \mathbf{U}} p_u = p_t = 1, \quad \text{and} \quad p_u \in [0, 1]\,. \tag{6.2}$$

Specially, if $p_u = 0$ in a given scheduling frame, user u has no power assigned and it is not scheduled in the frame.

We denote the set of multiplexed users in the considered scheduling frame as $\mathbf{s} = \{\, c\,(i) |\, i = 1, \ldots, s\}$, where $s = |\mathbf{s}|$ is the number of multiplexed users. It is a subset of $\mathbf{U}$ which contains the users with non-zero allocated power, i.e.,

$$\mathbf{s} = \{\, u \,|p_u > 0, u \in \mathbf{U}\,\}\,. \tag{6.3}$$

For ease of the following derivation, we assume that the multiplexed users in $\mathbf{s}$ are

sorted in descending order of their instantaneous CQIs, i.e.,

$$\Phi_{c(i)} \geq \Phi_{c(i+1)}, \quad i = 1, \ldots, (s-1). \tag{6.4}$$

According to [31], the SIC order at every receiver is always to first decode the users with the worse channel qualities to ensure the decoding correctness. Thus, the i-th user in the sorted user vector $\mathbf{s}$ decodes and cancels successively the interference signals of user $c(i+1) \sim c(s)$ in reverse order [1, 70]. Note that this SIC decoding order is optimal. The proof is given later in Section 6.2.3.

For the implementation of SIC, a multiplexed user with a higher CQI has to be informed of the modulation and coding schemes (MCSs) and the power allocated to the users that have lower CQIs. If the number of multiplexed users per frame is large, the signalling overhead and SIC processing complexity at user terminals become very high [31]. Therefore, the number of multiplexed users power frame is limited by a predefined parameter. We denote it as S and have $s \leq S$. In common practice, the limitation S is set to 2 or 3 [11, 77, 80]. Particularly, we have $S = 1$ in OMA systems.

According to [71], the instantaneous obtainable data rate of user $c(i)$ after SIC is calculated as

$$r_{c(i)} = \begin{cases} B \ln\left(1 + \Phi_{c(i)} p_{c(i)}\right), & i = 1, \\ B \ln\left(1 + \dfrac{\Phi_{c(i)} p_{c(i)}}{\Phi_{c(i)} \sum\limits_{j=1}^{i-1} p_{c(j)} + 1}\right), & i = 2, \ldots, s. \end{cases} \tag{6.5}$$

where B is the bandwidth of the SC-NOMA system. The unit of the user data rate in (6.5) is nat/s. The first user in the multiplexed user vector has the highest instantaneous CQI and cancels the interference signals of all other users. Thus, its decoded signal is interference-free. The other users can only eliminate the interference signals of the users with relatively lower CQIs and regard the rest interference signals as noise during their signal decoding.

We define the cumulative power (CP) allocated to the first i users in $\mathbf{s}$ as

$$q_i = \sum_{j=1}^{i} p_{c(j)}, \quad i = 1, \ldots, s. \tag{6.6}$$

Note that q_i is the power of the interference that user $c(i+1)$ cannot eliminate by SIC. Without loss of generality, we define $q_0 = 0$ for ease of expression in the following derivation. According to (6.2) and (6.3), we have $q_s = 1$ and

$$q_{i-1} < q_i, \quad i = 1, \ldots, s. \tag{6.7}$$

By (6.7), we define a vector of CPs in ascending order as $\mathbf{q} = \{q_0, q_1 \ldots, q_s\}$. Thus, the power ratio assigned to the i-th user in the multiplexed user vector $\mathbf{s}$ can be

calculated as

$$p_{c(i)} = q_i - q_{i-1}, \quad i = 1, \ldots, s. \tag{6.8}$$

Substituting (6.8) into (6.5), we derive a concise form of the instantaneous obtainable data rate of user $c(i)$ as

$$r_{c(i)} = B \ln \left(\frac{1 + \Phi_{c(i)} q_{c(i)}}{1 + \Phi_{c(i)} q_{c(i-1)}} \right), \quad i = 1, \ldots, s. \tag{6.9}$$

Therefore, the scheduling factor of PFS can be rewritten as a function of two vector variables, $\mathbf{s}$ and $\mathbf{q}$, i.e.,

$$\omega(\mathbf{s}, \mathbf{q}) = \sum_{i=1}^{s} \frac{r_{c(i)}}{R_{c(i)}}, \tag{6.10}$$

where $R_{c(i)}$ is the EMA data rate defined as in (2.3). In each scheduling frame, it is necessary to optimize the multiplexed users and the power allocated to their signals with the aim of maximizing $\omega(\mathbf{s}, \mathbf{q})$.

6.2 Relaxed DCPA Problem for Ideal SC-NOMA

In this section, we solve the optimization problem of DCPA for the ideal SC-NOMA system at first.

6.2.1 Relaxed Optimization Problem of DCPA

We assume that the limitation on the number of multiplexed users per frame is relaxed, equivalently, $S = U$. Under this condition, an arbitrary number of users can be scheduled simultaneously in each frame. Hence, we have an ideal NOMA system in which the cost for the high-order SIC is neglected [99]. The relaxed optimization problem of DCPA in the ideal SC-NOMA system is formulated as follows.

$$P6.1: \quad \max_{\mathbf{s}, \mathbf{q}} \ \omega(\mathbf{s}, \mathbf{q}), \tag{6.11a}$$

$$s.t. \quad \mathbf{s} \subseteq \mathbf{U}, \tag{6.11b}$$

$$q_{i-1} < q_i, \quad i = 1, \ldots, s. \tag{6.11c}$$

We expand the scheduling factor of PFS in (6.11a) as follows,

$$\begin{aligned} \omega(\mathbf{s}, \mathbf{q}) &= B \sum_{i=1}^{s} \left[\ln\left(1 + q_i \Phi_{c(i)}\right) - \ln\left(1 + q_{i-1} \Phi_{c(i)}\right) \right] R_{c(i)}^{-1} \\ &\overset{(6.7)}{=} B \sum_{i=1}^{s} \int_{q_{i-1}}^{q_i} \frac{\Phi_{c(i)}}{R_{c(i)} \left(1 + x \Phi_{c(i)}\right)} dx. \end{aligned} \tag{6.12}$$

As shown in (6.7), the CP values in $\mathbf{q}$ increases with index i. Thus, the integral terms in (6.12) must be positive. We denote the derivative function in the integral part of (6.12) as

$$\pi_u(x) = \frac{\Phi_u}{R_u(1 + x\Phi_u)}, \quad x \in [0, 1]. \tag{6.13}$$

Note that $\pi_u(x)$ is integrable everywhere within the range $x \in [0, 1]$. Substituting (6.13) into (6.12), we rewrite the scheduling factor as

$$\omega(\mathbf{s}, \mathbf{q}) = B \sum_{i=1}^{s} \int_{q_{i-1}}^{q_i} \pi_{c(i)}(x)\, dx. \tag{6.14}$$

6.2.2 Optimal Solution to the Relaxed DCPA Problem

We present the maximum objective of $P6.1$ and the optimal solutions of $\mathbf{s}$ and $\mathbf{q}$ for obtaining it in Theorem 6.1, followed by its proof.

Theorem 6.1. *The maximum objective of the optimization problem $P6.1$ is*

$$\hat{\omega} = B \int_{0}^{1} \max_{u \in \mathbf{U}} \pi_u(x)\, dx, \tag{6.15}$$

and it is obtained with $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$, i.e., $\hat{\omega} = \omega(\hat{\mathbf{s}}, \hat{\mathbf{q}})$, where

$$\begin{aligned} \hat{\mathbf{s}} &= \{\hat{c}(1), \hat{c}(2), \ldots, \hat{c}(\hat{s})\} \\ &= \left\{ \arg\max_{u \in \mathbf{U}} \pi_u(x) \,|x \in (0, 1) \right\}, \quad \hat{s} = |\hat{\mathbf{s}}|, \end{aligned} \tag{6.16}$$

and

$$\hat{\mathbf{q}} = \{\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_{\hat{s}-1}, \hat{q}_{\hat{s}}\}, \tag{6.17}$$

where

$$\hat{q}_0 = 0, \tag{6.18}$$

$$\hat{q}_i = \frac{R_{\hat{c}(i)}\Phi^{-1}_{\hat{c}(i)} - R_{\hat{c}(i+1)}\Phi^{-1}_{\hat{c}(i+1)}}{R_{\hat{c}(i+1)} - R_{\hat{c}(i)}}, \quad i = 1, \ldots, \hat{s} - 1, \tag{6.19}$$

$$\hat{q}_{\hat{s}} = 1. \tag{6.20}$$

In order to prove Theorem 6.1, we deduce the following lemmas. In the first step, we prove that $\hat{\omega}$, i.e., the integral expression in (6.15) is an upper bound of the objective in $P6.1$ as follows.

Lemma 6.1. *$\hat{\omega}$ is an upper bound of the objective function $\omega(\mathbf{s}, \mathbf{q})$ in $P6.1$, i.e.,*

$$\omega(\mathbf{s}, \mathbf{q}) \leq \hat{\omega}. \tag{6.21}$$

Proof. According to (6.14), we prove (6.21) as follows.

$$\begin{aligned}\omega(\mathbf{s},\mathbf{q}) =& B\sum_{i=1}^{s}\int_{q_{i-1}}^{q_q}\pi_{c(i)}(x)\,dx\\ \leq& B\sum_{i=1}^{s}\int_{q_{i-1}}^{q_i}\max_{j=1}^{s}\pi_{c(j)}(x)\,dx\\ =& B\int_{0}^{1}\max_{j=1}^{s}\pi_{c(j)}(x)\,dx\\ \leq& B\int_{0}^{1}\max_{u\in\mathbf{U}}\pi_u(x)\,dx\\ =&\hat{\omega}\end{aligned} \tag{6.22}$$

□

In the second step, we prove that $\hat{\omega}$ is obtained with $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ given in (6.16) and (6.17), respectively. Since the objective function $\omega(\mathbf{s},\mathbf{q})$ is the integral of $\pi_{c(i)}$ as shown in (6.14), we study the relationships among the derivative functions of different users as follows. Although the possibility that two users have the same instantaneous CQI is very low in reality, we consider this special case firstly for the sake of analysis completeness.

Lemma 6.2. *For two users, $u, v \in \mathbf{U}$, which have $\Phi_u = \Phi_v$,*

Case 1) *if $R_v < R_u$, then $\pi_u(x) < \pi_v(x)$;*

Case 2) *if $R_v > R_u$, then $\pi_u(x) > \pi_v(x)$;*

Case 3) *if $R_v = R_u$, then $\pi_u(x) = \pi_v(x)$.*

Proof. If $\Phi_u = \Phi_v$, we have

$$\frac{\pi_u(x)}{\pi_v(x)} = \frac{R_v}{R_u}. \tag{6.23}$$

Thus, the three cases in Lemma 6.2 hold. □

According to case 1 and 2 in Lemma 6.2, when multiple users in $\mathbf{U}$ have the same instantaneous CQI, only the one with the minimum R_u, equally, the maximum derivative function, may be selected by (6.16). Coincidentally, when their EMA data rates are also the same as shown in case 3, they have identical derivative functions. Under this condition, we can randomly select one of them as a candidate user and neglect the rest. This choice has no influence on maximizing the scheduling factor

as shown in (6.14). Thus, the CQIs of the optimal multiplexed users in vector $\hat{\mathbf{s}}$ are mutually different and strictly decreasing, i.e.,

$$\Phi_{\hat{c}(i)} > \Phi_{\hat{c}(i+1)}, \quad i = 1, \ldots, \hat{s} - 1. \tag{6.24}$$

In the following lemmas, we consider the relationships among the derivative functions under the condition that $\Phi_u \neq \Phi_v$, $u, v \in \mathbf{U}$.

Lemma 6.3. *For two users, $u, v \in \mathbf{U}$, which have $\Phi_u > \Phi_v$,*

Case 1) *if they meet the condition:*

$$0 < \frac{\Phi_v}{\Phi_u} < \frac{R_v}{R_u} < \frac{1 + \Phi_u^{-1}}{1 + \Phi_v^{-1}} < 1, \tag{6.25}$$

then it holds that

$$\pi_u(x) = \pi_v(x), \quad x = \theta_{u,v}, \tag{6.26}$$
$$\pi_u(x) > \pi_v(x), \quad x \in (0, \theta_{u,v}), \tag{6.27}$$
$$\pi_u(x) < \pi_v(x), \quad x \in (\theta_{u,v}, 1), \tag{6.28}$$

where

$$\theta_{u,v} = \theta_{v,u} = \frac{R_u \Phi_u^{-1} - R_v \Phi_v^{-1}}{R_v - R_u}; \tag{6.29}$$

Case 2) *if they meet the condition:*

$$\frac{R_v}{R_u} \leq \frac{\Phi_v}{\Phi_u}, \tag{6.30}$$

then it holds that

$$\pi_u(x) < \pi_v(x), \quad x \in (0, 1);$$

Case 3) *if they meet the condition:*

$$\frac{R_v}{R_u} \geq \frac{1 + \Phi_u^{-1}}{1 + \Phi_v^{-1}}, \tag{6.31}$$

then it holds that

$$\pi_u(x) > \pi_v(x), \quad x \in (0, 1).$$

Proof. Assuming $\Phi_u > \Phi_v, u, v \in \mathbf{U}$, and letting $\pi_u(x) = \pi_v(x)$, we have the x coordinate of the intersection point of the two derivative functions as

$$x = \theta_{u,v} = \theta_{v,u} = \frac{R_u \Phi_u^{-1} - R_v \Phi_v^{-1}}{R_v - R_u}. \tag{6.32}$$

Case 1) If the intersection point is in the range $\theta_{u,v} \in (0, 1)$, we have the

equivalence that

$$\begin{cases} \Phi_u > \Phi_v, \\ 0 < \dfrac{R_u\Phi_u^{-1} - R_v\Phi_v^{-1}}{R_v - R_u} < 1 \end{cases} \Leftrightarrow 0 < \frac{\Phi_v}{\Phi_u} < \frac{R_v}{R_u} < \frac{1+\Phi_u^{-1}}{1+\Phi_v^{-1}} < 1. \tag{6.33}$$

Under this condition, the relationship between $\pi_u(x)$ and $\pi_v(x)$ when $x \neq \theta_{u,v}$ is given as follows.

▷ When $x \in (0, \theta_{u,v})$, it holds that

$$\begin{aligned} &\pi_u(x) - \pi_v(x) \\ &= \frac{R_v\Phi_v^{-1} - R_u\Phi_u^{-1} - x(R_u - R_v)}{R_uR_v\left(\Phi_u^{-1}+x\right)\left(\Phi_v^{-1}+x\right)} \\ &\overset{(6.33)}{>} \frac{R_v\Phi_v^{-1} - R_u\Phi_u^{-1} - \theta_{u,v}(R_u - R_v)}{R_uR_v\left(\Phi_u^{-1}+x\right)\left(\Phi_v^{-1}+x\right)} \\ &\overset{(6.32)}{=} 0. \end{aligned} \tag{6.34}$$

▷ When $x \in (\theta_{u,v}, 1)$, on the contrary, it holds that

$$\pi_u(x) - \pi_v(x) < 0.$$

Case 2) If user u and v have relationship as

$$\frac{R_v}{R_u} \leq \frac{\Phi_v}{\Phi_u},$$

then we have

$$\frac{\pi_u(x)}{\pi_v(x)} = \frac{\Phi_u}{\Phi_v}\frac{R_v}{R_u}\frac{(1+x\Phi_v)}{(1+x\Phi_u)} < 1.$$

Case 3) If user u, v have relationship as

$$\frac{R_v}{R_u} \geq \frac{1+\Phi_u^{-1}}{1+\Phi_v^{-1}}, \tag{6.35}$$

then there are two subcases as follows.

▷ When $R_v \geq R_u$, it holds that

$$\frac{\pi_u(x)}{\pi_v(x)} = \frac{R_v}{R_u}\frac{\left(\Phi_v^{-1}+x\right)}{\left(\Phi_u^{-1}+x\right)} > 1. \tag{6.36}$$

▷ When $R_v < R_u$, we have

$$\frac{1+\Phi_u^{-1}}{1+\Phi_v^{-1}} \leq \frac{R_v}{R_u} < 1. \tag{6.37}$$

Then, it holds that

$$\begin{aligned}
&\left[\pi_u(x)\right]^{-1} - \left[\pi_v(x)\right]^{-1} \\
=&\left(1+\Phi_u^{-1}\right)R_u - \left(1+\Phi_v^{-1}\right)R_v + (1-x)(R_v - R_u) \\
\leq&(1-x)(R_v - R_u) \\
<&0.
\end{aligned} \tag{6.38}$$

Hence, combining (6.36) and (6.38), we have $\pi_u(x) > \pi_v(x)$ in Case 3. In summary, the three cases in Lemma 6.3 are proved completely. □

Based on Lemma 6.3, we deduce the following lemmas.

Lemma 6.4. *In the user vector* $\hat{\mathbf{s}}$*, it holds that*

$$\pi_{\hat{c}(i)}(x) > \pi_{\hat{c}(i+1)}(x), \quad x \in \left(0, \theta_{\hat{c}(i),\hat{c}(i+1)}\right), \tag{6.39}$$

$$\pi_{\hat{c}(i)}(x) < \pi_{\hat{c}(i+1)}(x), \quad x \in \left(\theta_{\hat{c}(i),\hat{c}(i+1)}, 1\right), \tag{6.40}$$

$$i = 1, 2, \ldots, \hat{s} - 1.$$

Proof. According to (6.24), the adjacent user $\hat{c}(i)$ and $\hat{c}(i+1)$ in $\hat{\mathbf{s}}$ must meet one of the three cases in Lemma 6.3. If they meet condition (6.30) in case 2, we have

$$\pi_{\hat{c}(i)}(x) < \pi_{\hat{c}(i+1)}(x), \quad x \in (0,1); \tag{6.41}$$

or if they meet condition (6.31) in case 3, we have

$$\pi_{\hat{c}(i)}(x) > \pi_{\hat{c}(i+1)}(x), \quad x \in (0,1). \tag{6.42}$$

The above two cases conflict with (6.16) since one of the two users must have a lower derivative function than the other one within the range $x \in (0,1)$ and cannot be included in $\hat{\mathbf{s}}$. Thus, there exist the relationships (6.39) and (6.40) between the derivative functions of two adjacent users in $\hat{\mathbf{s}}$ according to case 1 of Lemma 6.3. □

Lemma 6.5. *In the user vector* $\hat{\mathbf{s}}$*, it holds that*

$$\theta_{\hat{c}(i-1),\hat{c}(i)} < \theta_{\hat{c}(i),\hat{c}(i+1)}, \quad i = 2, \ldots, \hat{s} - 1. \tag{6.43}$$

Proof. By Lemma 6.4, the derivative function of user $\hat{c}(i)$ satisfies

$$\pi_{\hat{c}(i)}(x) < \pi_{\hat{c}(i-1)}(x), \quad x \in \left(0, \theta_{\hat{c}(i-1),\hat{c}(i)}\right), \tag{6.44}$$

$$\pi_{\hat{c}(i)}(x) < \pi_{\hat{c}(i+1)}(x), \quad x \in \left(\theta_{\hat{c}(i),\hat{c}(i+1)}, 1\right), \tag{6.45}$$

$$i = 2, \ldots, \hat{s} - 1.$$

We assume that the i-th user in $\hat{\mathbf{s}}$ leads to

$$\theta_{\hat{c}(i-1),\hat{c}(i)} \geq \theta_{\hat{c}(i),\hat{c}(i+1)}. \tag{6.46}$$

Due to (6.44) and (6.45), it is clear that

$$\pi_{\hat{c}(i)}(x) < \max\left\{\pi_{\hat{c}(i-1)}(x), \pi_{\hat{c}(i+1)}(x)\right\}, \quad x \in (0,1).$$

This conflicts with (6.16) and user $\hat{c}(i)$ cannot be included in the user vector $\hat{\mathbf{s}}$. Therefore, assumption (6.46) is false and (6.43) holds. □

Based on Lemma 6.1, 6.4 and 6.5 presented above, we prove Theorem 6.1 as follows.

Proof. By Lemma 6.4 and 6.5, the selected users in $\hat{\mathbf{s}}$ have their derivative functions to be the maximum in the following intervals of $x \in (0,1)$.

$$\arg\max_{u\in\mathbf{U}} \pi_u(x) = \begin{cases} \hat{c}(1), & x \in \left(0, \theta_{\hat{c}(1),\hat{c}(2)}\right), \\ \hat{c}(i), & x \in \left(\theta_{\hat{c}(i-1),\hat{c}(i)}, \theta_{\hat{c}(i),\hat{c}(i+1)}\right), \quad \text{and } i = 2,\ldots,\hat{s}-1, \\ \hat{c}(\hat{s}), & x \in \left(\theta_{\hat{c}(\hat{s}-1),\hat{c}(\hat{s})}, 1\right). \end{cases} \tag{6.47}$$

Comparing (6.17) and (6.29), it is clear that

$$\hat{q}_i = \theta_{\hat{c}(i),\hat{c}(i+1)}, \tag{6.48}$$

where $\theta_{\hat{c}(i),\hat{c}(i+1)}$ is the x coordinate of the intersection point of $\pi_{\hat{c}(i)}(x)$ and $\pi_{\hat{c}(i+1)}(x)$. Thus, according to (6.14), the scheduling factor obtained with $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ is the integral of the multiple derivative functions that are maximum within $\hat{s}$ adjacent intervals of $x \in (0,1)$, respectively. It is calculated as

$$\begin{aligned} \omega(\hat{\mathbf{s}}, \hat{\mathbf{q}}) &= B \sum_{i=1}^{\hat{s}} \int_{\hat{q}_{i-1}}^{\hat{q}_i} \pi_{\hat{c}(i)}(x)\, dx \\ &\overset{(6.47)}{=} B \sum_{i=1}^{\hat{s}} \int_{\hat{q}_{i-1}}^{\hat{q}_i} \max_{u\in\mathbf{U}} \pi_u(x)\, dx \\ &= B \int_0^1 \max_{u\in\mathbf{U}} \pi_u(x)\, dx \\ &= \hat{\omega}. \end{aligned} \tag{6.49}$$

By Lemma 6.1, $\hat{\omega}$ in (6.15) is an upper bound of the scheduling factor. In addition, it is obtained with $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ according to (6.49). Therefore, $\hat{\omega}$ is the maximum objective in $P6.1$. The optimal solutions to $P6.1$ are $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ as presented in (6.16) and (6.17). □

6.2.3 Optimal SIC Decoding Order

Remind that the above derivation is based on the SIC order given in Section 6.1, i.e., a SIC receiver always decode the signals of the users with the lower instantaneous CQIs firstly. We prove that this decoding order is a necessary condition for the optimal PFS in downlink NOMA systems as follows.

Proof. Consider two adjacent users $c(i)$ and $c(i+1)$ in the multiplexed user vector $\mathbf{s}$ and there exists $\Phi_{c(i)} > \Phi_{c(i+1)}$.

We assume a different decoding order for the two users as follows. Both of user $c(i)$ and $c(i+1)$ first decode and cancel the interference signals of user $c(i+2) \sim c(s)$ successively. Then, user $c(i+1)$ with a lower CQI decodes and cancels the interference signals of user $c(i)$ that has a higher CQI before it decodes its own signal. In this case, to guarantee that user $c(i+1)$ is able to decode the signal of user $c(i)$ correctly, the BS needs to choose the MCS for user $c(i)$ depending on the CQI of $c(i+1)$. Since user $c(i)$ has a better channel quality, it can decode its own signal without SIC. Thus, the weighted sum of their data rates is calculated as follows.

$$\begin{aligned}
\omega_{i,i+1} &= \frac{r_{c(i)}}{R_{c(i)}} + \frac{r_{c(i+1)}}{R_{c(i+1)}} \\
&= \frac{B}{R_{c(i)}} \ln\left[1 + \frac{\Phi_{c(i+1)} p_{c(i)}}{1 + \Phi_{c(i+1)}\left(q_{i-1} + p_{c(i+1)}\right)}\right] + \frac{B}{R_{c(i+1)}} \ln\left[1 + \frac{\Phi_{c(i+1)} p_{c(i+1)}}{1 + \Phi_{c(i+1)} q_{i-1}}\right] \\
&= \frac{B}{R_{c(i)}} \ln\left[\frac{1 + \Phi_{c(i+1)} q_{i+1}}{1 + \Phi_{c(i+1)}\left(q_{i-1} + p_{c(i+1)}\right)}\right] + \frac{B}{R_{c(i+1)}} \ln\left[\frac{1 + \Phi_{c(i+1)}\left(q_{i-1} + p_{c(i+1)}\right)}{1 + \Phi_{c(i+1)} q_{i-1}}\right]
\end{aligned} \tag{6.50}$$

Note that we have $p_{c(i)} + p_{c(i+1)} = q_{i+1} - q_{i-1}$ according to (6.8). The derivative of $\omega_{i,i+1}$ w.r.t. $p_{c(i+1)}$ is

$$\frac{\partial \omega_{i,i+1}}{\partial p_{c(i+1)}} = \frac{B \Phi_{c(i+1)}}{1 + \Phi_{c(i+1)}\left(q_{i-1} + p_{c(i+1)}\right)} \left(\frac{1}{R_{c(i+1)}} - \frac{1}{R_{c(i)}}\right). \tag{6.51}$$

Case 1) If $R_{c(i)} > R_{c(i+1)}$, then $\omega_{i,i+1}$ is a monotonically increasing function of $p_{c(i+1)}$. Thus, it is maximum when $p_{c(i+1)} = q_{i+1} - q_{i-1}$. In this case, $p_{c(i)} = 0$, i.e., user $c(i)$ has no power assigned and is not scheduled.

Case 2) If $R_{c(i)} < R_{c(i+1)}$, then $\omega_{i,i+1}$ is a monotonically decreasing function of $p_{c(i+1)}$. It is maximum when $p_{c(i+1)} = 0$, i.e., user $c(i+1)$ is not scheduled. Thus, $p_{c(i)} = q_{i+1} - q_{i-1}$.

Case 3) If $R_{c(i)} = R_{c(i+1)}$, then their weighted sum rate $\omega_{i,i+1}$ is equivalent to their sum rate. In this case, user $c(i+1)$ is not scheduled since allocating power to user $c(i)$ that has a higher instantaneous CQI is more efficient for improving the scheduling factor.

In the above three cases, only one of the two users is scheduled. Thus, it does not carry out SIC to eliminate the interference signal of the other one. This indicates that in the optimal solution for PFS, the SIC with the assumed decoding order shown above does not exist for any two adjacent users in the multiplexed user vector $\mathbf{s}$. By recursion, in every user receiver, the optimal SIC decoding order is always to first decode the signals of the users with worse instantaneous CQIs. □

6.2.4 Algorithm for DCPA in Ideal SC-NOMA

Based on the above derivation, we design now an algorithm to calculate $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ jointly as in Algorithm 6.1. According to the optimal solution presented in Theorem 6.1, the strategy of our algorithm is to select the users that provide the maximum derivative functions within the range $x \in (0, 1)$.

Algorithm 6.1 The Optimal Solution to $P6.1$

1: $\hat{c}(1) = \arg\max_{u \in \mathbf{U}} \pi_u(0)$,
2: $\mathbf{s}^{(1)} = \{\hat{c}(1)\}$, $\mathbf{q}^{(1)} = \{0\}$,
3: $\mathbf{V}_1 = \left\{ u | \, \theta_{\hat{c}(1),u} \in (0,1), u \in (\mathbf{U}/\hat{c}(1)) \right\}$,
4: $i = 1$.
5: **while** $\mathbf{V}_i \neq \varnothing$ **do**
6: $\quad \hat{c}_{i+1} = \arg\min_{u \in \mathbf{V}_i} \theta_{\hat{c}(i),u}$,
7: $\quad \mathbf{s}^{(i+1)} = \{\mathbf{s}_n, \hat{c}(i+1)\}$, $\mathbf{q}^{(i+1)} = \left\{\mathbf{q}^{(i)}, \theta_{\hat{c}(i),\hat{c}(i+1)}\right\}$,
8: $\quad \mathbf{V}_{i+1} = \left\{ u \left| \theta_{\hat{c}(i+1),u} \in (0,1), u \in (\mathbf{V}_i/\hat{c}(i+1)) \right. \right\}$,
9: $\quad i = i + 1$.
10: **end while**
11: $\hat{\mathbf{s}} = \mathbf{s}^{(i)}$, $\hat{\mathbf{q}} = \left\{\mathbf{q}^{(i)}, 1\right\}$.
12: Note: If there are multiple maximums in line 1 or multiple minimums in line 6, the algorithm chooses the user with the lowest instantaneous CQI due to Lemma 6.3. In this case, if there are multiple users with the same lowest CQI, they must also have an identical averaged data rate R_u. Thus, the algorithm randomly chooses one of them due to Lemma 6.2.

The algorithm starts from $x = 0$ and selects the first user $\hat{c}(1)$ that has the largest $\pi_u(0)$. Then, we add it to the multiplexed user vector $\mathbf{s}^{(1)}$ and calculate the intersection point $\theta_{\hat{c}(1),u}$ of the derivative functions of user $\hat{c}(1)$ and each of the other users $u \in (\mathbf{U}/\hat{c}(1))$. By Lemma 6.4, a user may be scheduled only when $\theta_{\hat{c}(1),u} \in (0, 1)$. Thus, we remove the invalid users and denote the updated rest candidate user set in the first step as $\mathbf{V}_1$.

In the i-th iteration of the while loop, we select user $\hat{c}(i+1)$ with the minimum intersection point $\theta_{\hat{c}(i),\hat{c}(i+1)}$ and add it to $\mathbf{s}^{(i+1)}$ because the derivative function is monotonically decreasing. According to (6.48), the optimal CP equals to $\theta_{\hat{c}(i),\hat{c}(i+1)}$. We add it to the CP vector $\mathbf{q}^{(i+1)}$. Then, we update the rest candidate user set $\mathbf{V}_{i+1}$. In every iteration, we select one user and remove the invalid users. By Lemma 6.4

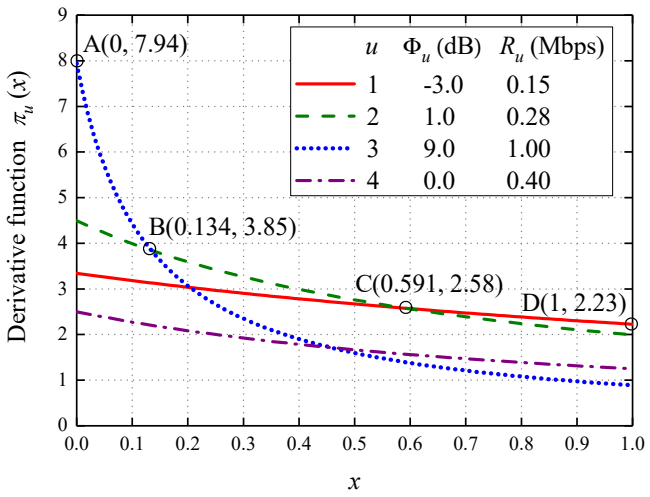


Figure 6.1: A 4-user example of Algorithm 6.1.

and 6.5, $\Phi_{\hat{c}(i)}$ decreases and $\theta_{\hat{c}(i),\hat{c}(i+1)}$ increases. Therefore, the algorithm must be convergent and stops when there are no valid users in $\mathbf{V}_i$ to process. The outcomes of the algorithm are the optimal solutions as given in (6.16) and (6.17).

The computational complexity of Algorithm 6.1 is no more than $(U+1)\,U/2$ for USS. Therefore, it is much lower than the FUSC method which has a complexity of $2^U - 1$ [10, 11]. Moreover, it is unnecessary to sort the users in terms of their instantaneous CQIs before its execution. Therefore, the computational complexity can be further reduced.

We present an example of the algorithm and derivative functions with $U = 4$ in Figure 6.1. By using Algorithm 6.1, we obtain the optimal multiplexed user vector $\hat{\mathbf{s}} = \{3,\ 2,\ 1\}$ and the corresponding CP vector $\hat{\mathbf{q}} = \{0,\ 0.134,\ 0.591,\ 1\}$. The maximum scheduling factor is the integral along the segments A-B-C-D. User 2 and 4 meet condition (6.31) in Lemma 6.3. Therefore, we have $\pi_4(x) < \pi_2(x)$ in the range $x \in (0, 1)$. In this case, user 4 must not be selected into the optimal multiplexed user vector $\hat{\mathbf{s}}$.

6.3 DCPA Problem for Practical SC-NOMA

Considering the limitation on the number of multiplexed users per frame as we discussed in Section 6.1, the optimization problem of DCPA for the practical SC-NOMA system is formulated as follows.

$$P6.2: \quad \max_{\mathbf{s},\mathbf{q}} \ \omega(\mathbf{s},\mathbf{q}), \tag{6.52a}$$

$$s.t. \quad \mathbf{s} \subseteq \mathbf{U}, \tag{6.52b}$$

$$q_{i-1} < q_i, \quad i = 1, \ldots, s. \tag{6.52c}$$

$$s \leq S. \tag{6.52d}$$

Note that the optimal solution to $P6.1$ is based on the assumption that $S = U$. However, parameter S in $P6.2$ is normally a smaller integer for practical SIC implementation. Since the number of scheduled users s is limited by S, it is possible that $\hat{s} > S \geq s$. In this case, the maximum scheduling factor in $P6.1$ may be unobtainable in $P6.2$. Therefore, the optimal transmission performance of PFS in the ideal SC-NOMA system under the condition $S = U$ is an upper bound for practical SC-NOMA.

Due to the integer limitation on the size of the multiplexed user set in (6.52d), $P6.2$ is a mixed-integer-nonlinear-programming problem and very difficult to solve. A straightforward approach to obtain the optimal multiplexed user set is using FUSC, i.e., comparing the optimal scheduling factors of all candidate user sets each of which contains no more than S users. The optimal power allocation for each user set can be computed by the algorithm developed for the relaxed problem in Section 6.2. However, as we discussed, the FUSC method costs extremely high computational complexity, especially when the number of users is large. In fact, some candidate user sets can be omitted for comparison since they are infeasible for the optimal solution to $P6.2$. In order to reduce the computational complexity for DCPA for practical SC-NOMA systems, we propose two USS schemes based on our optimal power allocation in Section 6.2, including an optimal tree-searching-based USS (TSU) scheme and a suboptimal preselection-based USS (PSU) scheme.

6.3.1 Tree-Searching-Based User Set Selection

The power set of the user set $\mathbf{U}$ contains all of its 2^U subsets, including the empty set [135]. We can build a forest to represent all of the possible nonempty user subsets by the following rules:

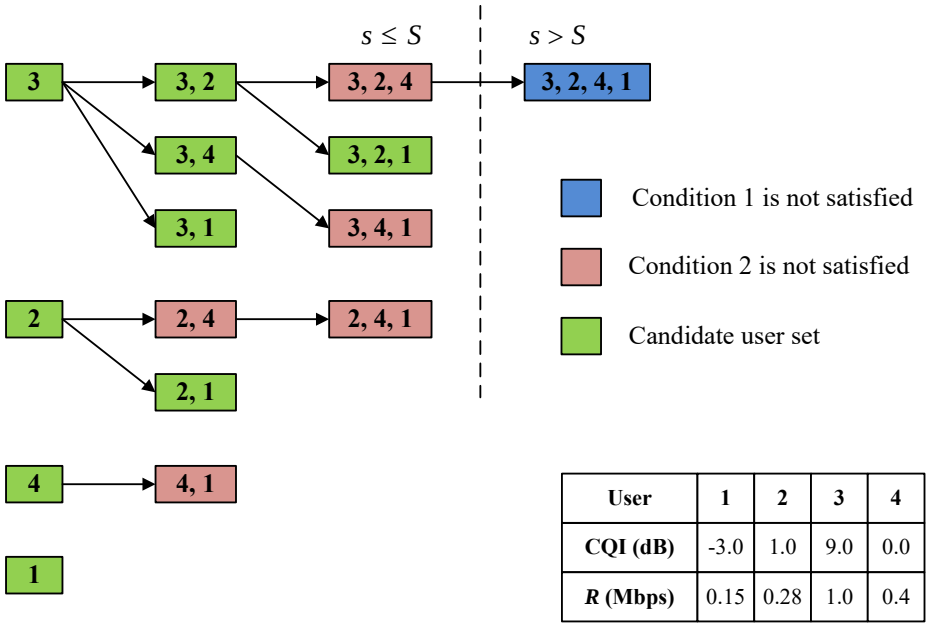
- ▷ Each tree root node has only one unique user in $\mathbf{U}$;
- ▷ Each child node has a unique candidate user set in which users are sorted in descending order of their CQIs;
- ▷ The child nodes of a node with the user set $\mathbf{s} = \{c(1), \ldots, c(s)\}$ are

$$\left\{ \mathbf{s} \cup \{u\} \middle| \, \Phi_{c(s)} > \Phi_u, u \in \mathbf{U} \right\}; \tag{6.53}$$

In Figure 6.2, we present the tree structures of the 4-user example given in Figure 6.1, in which the maximum number of multiplexed users is $S = 3$. The level of each tree node is the same as the number of multiplexed users in it.

In order to reduce the number of compared candidate user sets, we can avoid comparing the user sets which must not be the optimal one. Among the multiple nodes in the forest, we only compute and compare the scheduling factors of the candidate user set in the nodes that meet the following conditions:

Condition 1) The level of the node is no larger than S.



User	1	2	3	4
CQI (dB)	-3.0	1.0	9.0	0.0
R (Mbps)	0.15	0.28	1.0	0.4

Figure 6.2: Illustration of the tree structures of candidate user sets ($U = 4$, $S = 3$).

Condition 2) The user set in the node meets Lemma 6.5.

The first condition guarantees that the number of multiplexed users is no more than the limitation S. Then, given a user set satisfying Condition 1, the optimal solution to the power allocation problem in the ideal SC-NOMA system can be applied to it. Thus, if this user set is the optimal one, it must meet Lemma 6.5. Comparing to a node $\mathbf{s}$, one of its child node $\mathbf{s} \cup \{c(s+1)\}$ has one extra user $c(s+1)$ at the end of its multiplexed user vector. This user has a lower CQI than the others in $\mathbf{s}$. Therefore, we need to calculate only the intersection point of $\pi_{c(s)}(x)$ and $\pi_{c(s+1)}(x)$, i.e., $\theta_{c(s),c(s+1)}$, and check if it meets (6.43) and in the range $\theta_{c(s),c(s+1)} \in (0,1)$. In addition, a child node includes a subsequence of the users in its parent node. Therefore, if the user set in a node does not meet Condition 2, all of the child nodes in its branch do not, either. Then, this whole branch can be omitted. Note that the TSU scheme can only remove partial infeasible user sets for the optimal solution. However, as the scale of the problem increases significantly with U and S, it becomes superior to the FUSC scheme in terms of computational complexity.

Based on the above two conditions, we can adopt the depth-first searching algorithm to visit the valid tree nodes and compare the scheduling factors offered by the candidate user sets in them. As shown in Figure 6.2, the node with four users inside, i.e., $\{3,2,4,1\}$, does not meet Condition 1 and is omitted. The derivative functions of user 4 and 1 have no intersection point in the range $x \in (0,1)$, as shown in Figure 6.1. Therefore, Condition 2 is not satisfied by the nodes that contain this pair of users, including $\{4,1\}$, $\{3,4,1\}$, and $\{2,4,1\}$. Thus, they are removed. Similarly, the nodes with both of user 4 and 2 are omitted as well. The rest user sets are candidates that need to be compared for obtaining the overall maximum

scheduling factor in $P6.2$.

6.3.2 Preselection-Based User Set Selection

To further reduce the computational complexity for USS, we utilize Algorithm 6.1 for preselection of the candidate users. By combining it with the FUSC method, we propose a preselection-based USS (PSU) scheme as presented in Algorithm 6.2. Specifically, we obtain the optimal multiplexed users for $P6.1$ with Algorithm 6.1 in the first step. Then, only these selected user are considered for the optimization problem $P6.2$. Among them, we choose the optimal S users for multiplexing by FUSC. The overall computational complexity of our proposed PSU scheme for USS is extremely low and is bounded by

$$\left(\begin{array}{c} U+1 \\ 2 \end{array} \right) + \left(\begin{array}{c} \hat{s} \\ S \end{array} \right). \tag{6.54}$$

Therefore, the PSU scheme is favorable to the DCPA in practical SC-NOMA systems.

Algorithm 6.2 Preselection-based USS (PSU) scheme

1: Solve $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ by Algorithm 6.1.
2: **if** $\hat{s} \leq S$, **then**
3: $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ are valid solutions and can be applied to scheduling directly.
4: **else**
5: Use the FUSC method to select the optimal subsequence $\tilde{\mathbf{s}} \subset \hat{\mathbf{s}}$, which contains S users and has the highest ω as follows.

$$\begin{aligned} (\tilde{\mathbf{s}},\ \tilde{\mathbf{q}}) &= \arg\max_{(\mathbf{s},\,\mathbf{q})} \omega\left(\mathbf{s}, \mathbf{q}\right), \\ s.t. \quad \mathbf{s} &= \{c\left(1\right), \ldots, c\left(S\right)\} \subset \hat{\mathbf{s}}, \\ \mathbf{q} &= \{0, q_1, \ldots, q_{S-1}, 1\}, \\ q_i &= \theta_{c(i),c(i+1)}, \quad i = 1, \ldots, S-1. \end{aligned}$$

6: **end if**

Note that the PSU scheme does not guarantee that the optimal multiplexed user set is selected since the preselection may omit the users in the optimal solution to $P6.2$. However, the preselected users have their derivative functions being maximum in certain ranges of $x \in (0, 1)$. Therefore, they have larger chances to be the ones in the optimal solution of $P6.2$. Hence, the user set solved by the PSU scheme is identical with the optimal one in most cases and yields close-to-optimal performance. This is verified later by simulation results in Section 6.5.

If we set $S = 2$ in the example given in Figure 6.1, the preselected user set obtained in the first stage is $\hat{\mathbf{s}} = \{3, 2, 1\}$. Then, the optimal 2-user subset $\tilde{\mathbf{s}} = \{3, 2\}$ is selected by FUSC. The corresponding optimal CP vector is $\tilde{\mathbf{q}} = \{0, 0.134, 1\}$.

6.4 Upper Bound Performance Analysis

In this section, we analyze the upper bound performance obtained in the ideal SC-NOMA system based on our optimal solution to the relaxed DCPA problem $P6.1$ and stochastic channel modeling.

Under the fluctuating wireless channels, the instantaneous CQI of a user is a random variable. According to our SINR models developed in Chapter 3, we have the cumulative distribution function (CDF) and probability density function (PDF) of CQI as

$$F_{\Phi_u}(\phi) = \Pr\{\Phi_u < \phi\}, \quad \text{and} \tag{6.55}$$

$$f_{\Phi_u}(\phi) = \frac{\partial F_{\Phi_u}(\phi)}{\partial \phi}, \quad \phi > 0. \tag{6.56}$$

The value of a derivative function at a certain point of $x \in (0,1)$, i.e., $\pi_u(x)$ in (6.13), is a random variable depending on Φ_u. We derive its conditional CDF given x as follows,

$$\begin{aligned} F_{\pi_u}(y\,|x) =& \Pr\{\pi_u(x) < y\} \\ =& \Pr\left\{\frac{\Phi_u}{(\Phi_u x + 1) R_u} < y\right\} \\ =& \Pr\left\{\Phi_u < \frac{y R_u}{1 - y R_u x}\right\} \\ =& F_{\Phi_u}\left(\frac{y R_u}{1 - y R_u x}\right), \\ & y \in \left(0, \frac{1}{x R_u}\right), \quad x \in (0,1). \end{aligned} \tag{6.57}$$

Accordingly, the conditional PDF of the derivative function is derived as

$$f_{\pi_u}(y\,|x) = f_{\Phi_u}\left(\frac{g R_u}{1 - y R_u x}\right) \frac{R_u}{(1 - y R_u x)^2}. \tag{6.58}$$

We denote the ergodic data rate of user u as $\overline{r}_u = \mathbb{E}[r_u]$. Thus, the expectation of the overall throughput in the cell is expressed as

$$\overline{r}_\Sigma = \sum_{u \in \mathbf{U}} \overline{r}_u. \tag{6.59}$$

By Theorem 3.2, when $\tau \gg 1$, we have the data rate approximation as

$$\overline{r}_u = \mathbb{E}[R_u] \approx R_u. \tag{6.60}$$

Therefore, the expectation of the scheduling factor per user approximately equals

to 1, i.e.,

$$\overline{\omega}_u = \mathbb{E}\left[\frac{r_u}{R_u}\right] \approx 1. \tag{6.61}$$

This implies that PFS maintains weighted fairness among users. Thus, each user has its scheduling factor ω_u fluctuating around 1. On the other hand, we can calculate $\overline{\omega}_u$ with (6.57) and (6.58) as follows,

$$\begin{aligned}
\overline{\omega}_u &= B \int_0^1 \int_0^{\frac{1}{xR_u}} f_{\pi_u}(y\,|x\,)\, y \prod_{v\in(\mathbf{U}/u)} \Pr\{\pi_u(x) < y\}\, dydx \\
&= B \int_0^1 \int_0^{\frac{1}{xR_u}} f_{\pi_u}(y\,|x\,) \prod_{v\in(\mathbf{U}/u)} F_{\pi_v}(y\,|x\,)\, y\, dy\, dx \\
&\overset{(6.60)}{\approx} B \int_0^1 \int_0^{\frac{1}{x\overline{r}_u}} f_{\Phi_u}\left(\frac{y\overline{r}_u}{1-y\overline{r}_u x}\right) \frac{y\overline{r}_u}{(1-y\overline{r}_u x)^2} \prod_{v\in(\mathbf{U}/u)} F_{\Phi_v}\left(\frac{y\overline{r}_v}{1-y\overline{r}_v x}\right) dy\, dx \\
&\overset{(6.61)}{\approx} 1.
\end{aligned} \tag{6.62}$$

This is the mean value of ω_u under the condition that $\pi_u(x)$ is the maximum and correspondingly user u is selected into $\hat{\mathbf{s}}$ by (6.16). Assuming ergodicity of the radio channels, the expectation of user data rate $\overline{r}_u$ can be obtained by solving (6.62). Substituting the estimated CDFs and PDFs of the instantaneous user CQIs into (6.62), we have the equations of the estimated user data rates. The closed-form solution to (6.62) is unobtainable, we can nevertheless calculate the results by numerical methods [114].

In addition, the average power ratio allocated to user u can be calculated with the estimated user data rates as follows,

$$\overline{p}_u = \mathbb{E}[p_u] = \int_0^1 \int_0^{\frac{1}{x\overline{r}_u}} f_{\pi_u}(y\,|x\,) \prod_{v\in(\mathbf{U}/u)} F_{\pi_v}(y\,|x\,) dydx. \tag{6.63}$$

Although the analytical performance obtained by (6.62) is an upper bound, it can be utilized to estimate user data rate and overall throughout in practical SC-NOMA systems. In the following section, the estimation accuracy is evaluated by comparing it to the simulation results.

6.5 Simulations and Numerical Results

In this section, the data rate performance of different DCPA schemes is evaluated by system-level simulations in Matlab [131]. In the OMA system ($S = 1$), we use the

Table 6.1: Simulation Parameters of the SC-NOMA System

Parameter	**Value**
Inter site distance	500 m
Minimum link distance	35 m
Bandwidth	200 kHz @ 2.0 GHz
BS transmit power (p_t)	29 dBm
BS transmit antenna gain	15 dBi
Path loss	128.1+37.6lg(d[km])
Standard deviation of shadow fading	8 dB
Fast fading	Rayleigh model
Noise power density	-174 dBm/Hz
Noise figure	5 dB
Frame duration	1 ms
Averaging coefficient (τ)	1000
User number per cell (U)	2∼15

conventional PFS scheme to select the scheduled user with the maximum scheduling factor in each frame. In practical SC-NOMA systems with $S = 2$ and 3, we compare four PF-oriented schemes, including FTPA in [10], TTPA in [11], and our proposed TSU and PSU schemes. The TTPA scheme obtains the optimal multiplexed users and power allocation, while FTPA outputs only suboptimal solutions. When $S = U$, we use Algorithm 6.1 to solve $\hat{\mathbf{s}}$ and $\hat{\mathbf{q}}$ for the ideal SC-NOMA system. Then, its analytical performance is calculated and compared to the simulation results of the practical SC-NOMA systems with the PSU scheme.

The simulation parameters are configured according to a typical downlink multi-cell network developed in [136] and are listed in Table 6.1. A downlink cellular network with 37 cells is deployed in a hexagonal grid pattern. To avoid the edge effect, only the performance of the central cell is computed and the other 36 neighbor BSs act as interferers. User terminals are uniformly randomly distributed in the cell. The number of reported IRSRPs I_R is set to 8 [110]. The decay factor in the FTPA scheme is set to 0.7, which is the optimal value for PFS as given in [10].

In Figure 6.3, we present the change of EMA data rate per user over time and compare it to the ergodic user data rate. The EMA data rate is initialized according to the expectation of the spectrum efficiency calculated with the mean user CQI, i.e.,

$$R_u\left(t=0\right)=\frac{\ln\left(1+\phi_u\right)}{U}, \tag{6.64}$$

where ϕ_u is the mean CQI of user u. As we proved in Theorem 3.2, the EMA data rate approximately equals to the ergodic data rate although it is fluctuating, as shown in Figure 6.3. In addition, the EMA data rates become steady after about 5,000 frames. Thus, we compute the statistic performance over 20,000 frames after 5,000 initial frames.

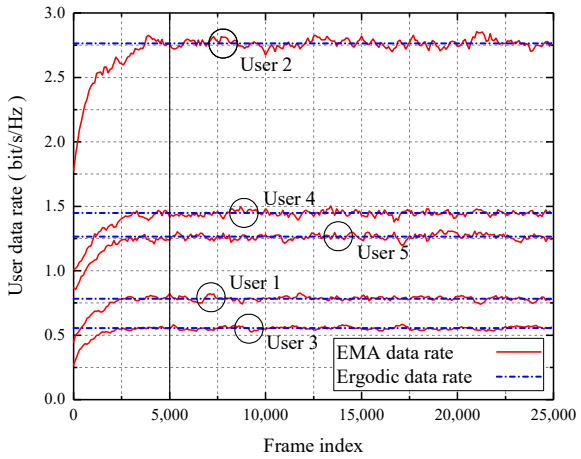


Figure 6.3: Comparison of the EMA and ergodic user data rates (PSU, $U = 5$, and $S = 2$).

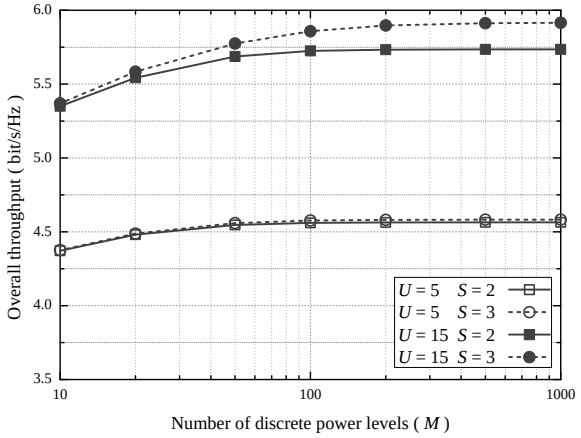
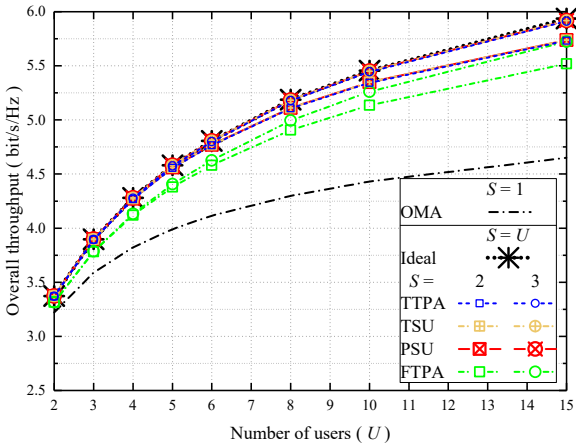
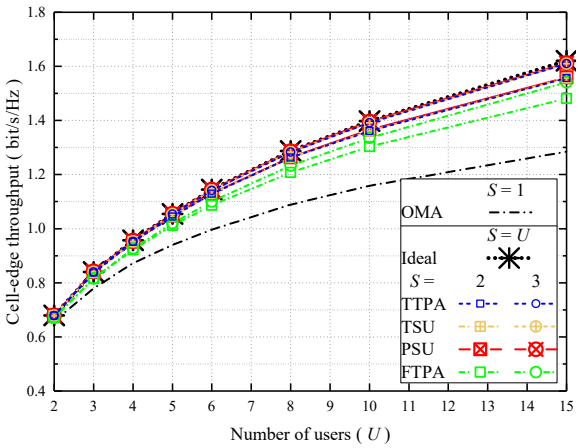


Figure 6.4: The overall throughput of SC-NOMA with different numbers of discrete power levels (TTPA).

In NOMA systems, the precision of power allocation is determined by the number of discrete power levels, which is denoted as an integer M. In Figure 6.4, we present the overall throughput of the SC-NOMA system with different power allocation precisions. The performance is improved as the parameter M increases due to a more delicate control of the allocated power. However, the computational complexity for the searching-based TTPA scheme is $O\left(M^2 S\right)$ [11]. Therefore, it raises sharply with the number of discrete power levels. As shown in Figure 6.4, the additional improvement of the throughput brought by the increment of M is limited when it is large enough, e.g., $M > 100$. Thus, in the following simulations, we set the parameter $M = 200$ for a good balance between the system performance and computational complexity of TTPA.



(a) Overall throughput



(b) Cell-edge throughput

Figure 6.5: Simulation results of the throughput performance in SC-NOMA systems.

6.5.1 Simulation Results

The simulation results of the throughput performance in SC-NOMA systems are presented in Figure 6.5. The cell-edge throughput shown in Figure 6.5(b) is defined as the mean throughput of the lowest 5% users. In comparison to the OMA system, SC-NOMA systems improve both overall and cell-edge throughput. The TTPA and TSU schemes obtain higher throughput than FTPA because of their optimal power allocation for PFS. In addition, the simulation results indicate that the PSU scheme obtains close-to-optimal performance that is nearly identical with TTPA and TSU in practical SC-NOMA systems. The performance increases when there are more users in the cell, owing to the multi-user diversity gain brought by PFS. In particular, we have the upper bound performance while $S = U$ in the ideal SC-NOMA system, which is consistent with our analysis in Section 6.4. The performance in the practical SC-NOMA cases ($S = 2$, 3) is close to the upper bound, especially when there are fewer users.

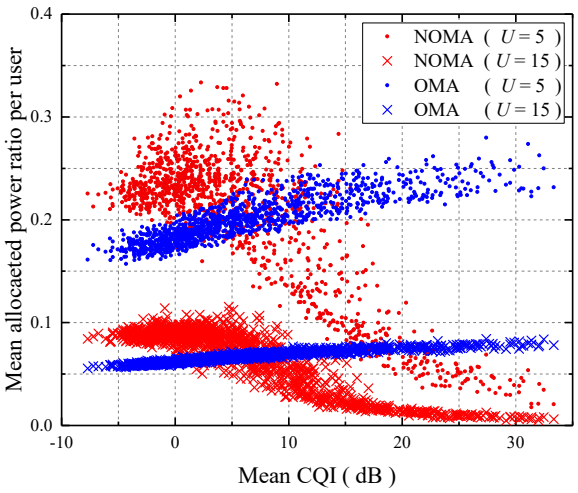
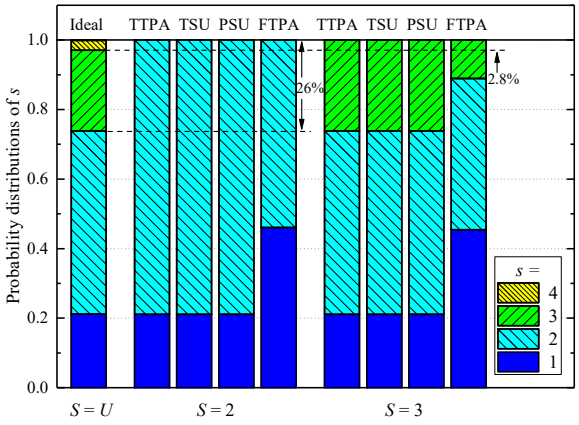


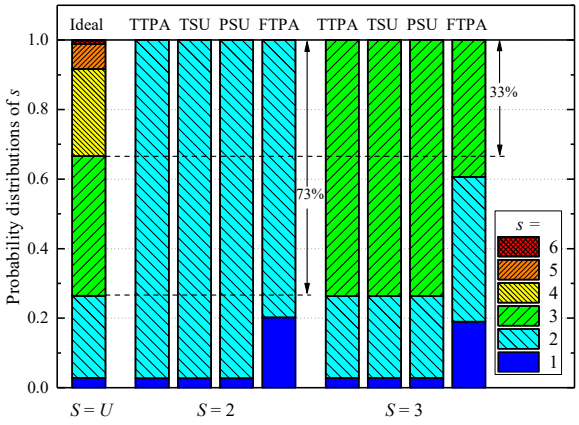
Figure 6.6: Mean allocated power ratio per user (PSU, $S = 2$).

Figure 6.6 presents the mean allocated power ratio of each user in the simulations. Compared to the case of $U = 5$, less power is obtained per user while $U = 15$ due to more users sharing the limited transmit power budget. The conventional PFS in the OMA system allocates similar power ratios to the users with different mean CQIs. The high-CQI users obtain slightly more power when there are very few users. However, the SC-NOMA system allocates more power to the low-CQI users as shown in the figure. This effectively enhances the cell-edge user data rates and consequently improve the user fairness in the network. In addition, due to the extra multi-user diversity gain brought by PFS in the power domain, the over throughput in the network is also guaranteed as shown in Figure 6.5.

To further study the impacts of U and S on the throughput performance, we present in Figure 6.7 the probability distributions of the multiplexed user numbers s. The TTPA, TSU, and PSU schemes obtain identical results and multiplex more users than FTPA. The numbers of multiplexed users with these schemes are limited by S in the practical SC-NOMA systems. In the 3-user SC-NOMA case, it is allowable to multiplex one extra user than the 2-user case. Thus, the performance is higher when $S = 3$ because of the additional candidate user sets. The ideal SC-NOMA system with $S = U$ has more users multiplexed simultaneously, especially when the number of users is higher. For instance, in the scenario with $U = 5$ as shown in Figure 6.7(a), the probability results are $\Pr\{\hat{s} > 2\} = 26\%$ and $\Pr\{\hat{s} > 3\} = 2.8\%$ in the ideal SC-NOMA system. When $U = 15$, the probabilities increase to $\Pr\{\hat{s} > 2\} = 73\%$ and $\Pr\{\hat{s} > 3\} = 33\%$, respectively. This is due to that a larger number of candidate users in the cell increase the chance to multiplex more users simultaneously without any limitation in the ideal SC-NOMA system. Therefore, as U increases, the performance gaps between the ideal and practical SC-NOMA systems become larger. However, due to the performance proximity as shown in Figure 6.5, it is feasible to use the analytical performance of the upper bound for data rate estimation in practical SC-NOMA systems. Therefore, we evaluate the estimation accuracy by simulations in Section 6.5.2.



(a) $U = 5$



(b) $U = 15$

Figure 6.7: Probability distributions of the multiplexed user numbers per frame.

To assess the computational complexity for USS, we calculate the average number of compared candidate user sets, as shown in Figure 6.8. In the OMA system, only one user is scheduled every time. Thus, the number of its compared user sets is U. In the FTPA and TTPA schemes, the FUSC method is adopted for USS [10, 11]. It considers all possible combinations of users and thus leads to the highest complexity. When $S > 2$, the TSU scheme can reduce the number of compared user sets by removing some infeasible ones. The computational complexity is significantly reduced by our proposed PSU scheme and is close to Algorithm 6.1. Moreover, it is nearly linear in the number of users and insensitive to the change of parameter S. This is due to many invalid users that are removed by Algorithm 6.1. Moreover, TSU, PSU, and Algorithm 6.1 calculate the optimal CP vector according to the closed-form solutions. Thus, they cost the lowest computational complexity for power allocation. In contrast, the TTPA scheme obtains the optimal allocated power with a more complex searching method.

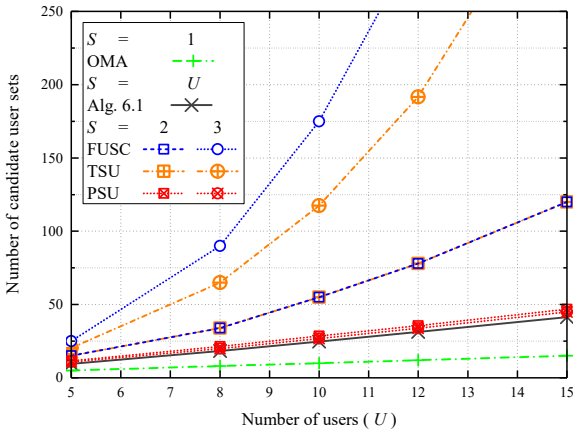


Figure 6.8: Computational complexity for USS with various DCPA schemes.

6.5.2 Data Rate Estimation

We investigate the relative differences of the estimated user data rates with various influence factors. The analytical results of the upper bound performance are compared to the simulation ones obtained with the PSU scheme in practical SC-NOMA systems.

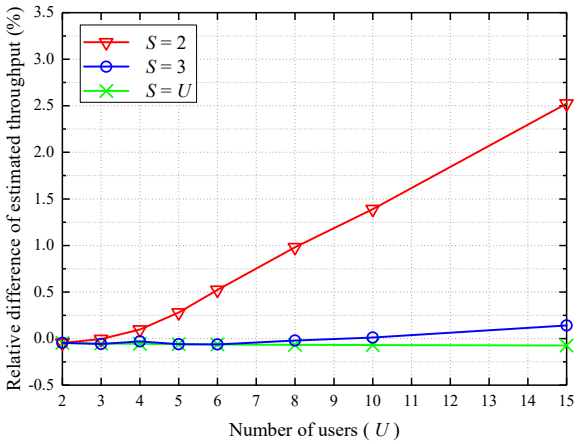
6.5.2.1 Number of Users

In Figure 6.9(a), we present the relative difference of the estimated overall throughput. The analytical results are very close to the simulation ones while $S = U$, verifying our analysis of the upper bound performance. When $S = 2$ and 3, the estimation error increases with the number of users because of the enlarged gaps of the multi-user diversity gains between the practical and ideal SC-NOMA systems. Nevertheless, the throughput estimation is very accurate even in the scenario with 15 users, where the relative error is lower than 0.2% for the 3-user SC-NOMA system and is 2.5% for the 2-user case.

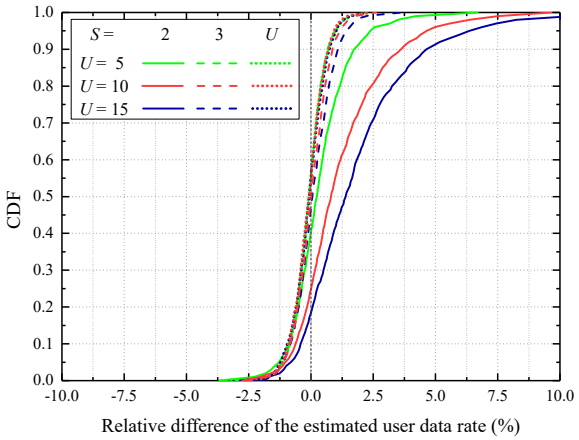
We calculate the relative difference of the estimated data rate per user and present their CDF in Figure 6.9(b). The 3-user SC-NOMA system obtains better performance than the 2-user case, which is closer to the upper bound. Thus, the estimation results are more accurate while $S = 3$. Nevertheless, the estimation accuracy is feasible for practical applications when $S = 2$. For instance, when there are 15 users, more than 92% statistical relative differences are within the range of $\pm 5.0\%$.

6.5.2.2 Partial Channel State Information

We investigate the influence of partial reported CSI on the data rate estimation accuracy. The relative differences of the estimated overall throughput with different



(a) Estimated overall throughput

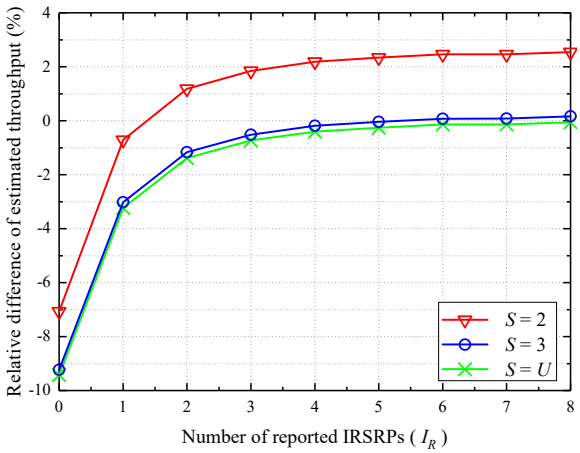


(b) Estimated data rate per user

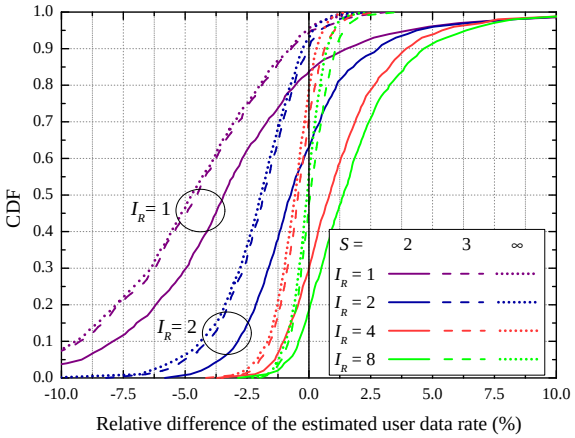
Figure 6.9: Relative differences of the estimated user data rates under different numbers of users.

numbers of reported IRSRPs are presented in Figure 6.10(a). With fewer IRSRPs reported per user, the influence of inaccurate channel models on the data rate estimation becomes more evident. The estimated throughput is lower than the actual result when I_R is very small. The relative differences are reduced within the range $\pm 3.0\%$ when $I_R > 1$, and barely change when $I_R > 3$, indicating that the estimated probability distributions of the user CQIs are accurate enough.

We calculate the relative differences of the estimated user data rates and their CDFs with different numbers of reported IRSRPs, as shown in Figure 6.10(b). When $I_R = 1$, the estimated user CQIs are inaccurate due to the lack of CSI. Thus, the estimated data rates have larger deviations. This inaccuracy drawback is remedied by increasing I_R and becomes negligible in comparison to other influencing factors (e.g., U and S) when I_R is larger than 4. As shown in Figure 6.10(b), the gap between the estimation errors with $I_R = 4$ and 8 is less than 1.0%. Therefore, it is reasonable to set I_R to a smaller number than 8 so that fewer IRSRPs are reported



(a) Estimated overall throughput



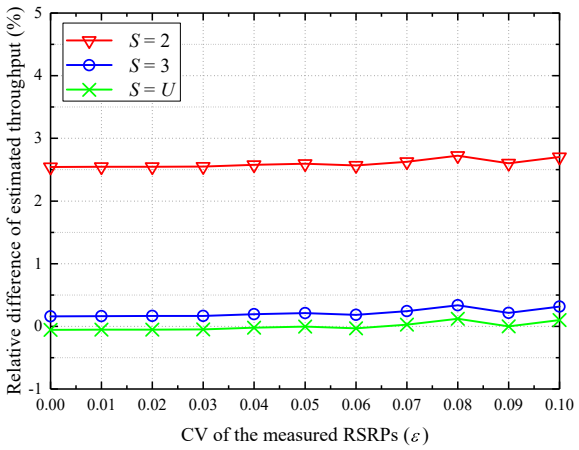
(b) Estimated data rate per user

Figure 6.10: Relative differences of the estimated user data rates with partial IRSRPs ($U = 15$).

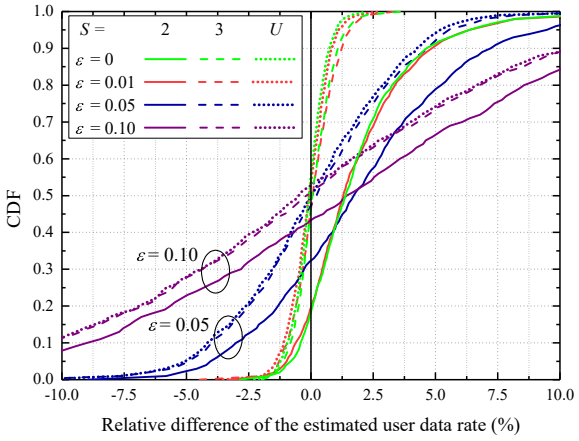
for the sake of reduction in signalling overhead.

6.5.2.3 Imperfect CSI Measurement

We consider the estimation error caused by the imperfect measurements of RSRPs. Each reported RSRP is the mean value of multiple received reference signal power samples. Under Rayleigh fading channels, the reported RSRP follows the Erlang distribution. We denote its coefficient of variation (CV) as $\varepsilon > 0$ in the imperfect measurements. In Figure 6.11(a), we present the relative differences of the estimated overall throughput with different variations of the measured RSRPs. The estimated overall throughput is insensitive to the inaccuracy of the measured RSRPs. However, as shown in Figure 6.11(b), the results of the estimated user data rates indicate that imperfect CSI measurements enlarge the error range. When $\varepsilon = 0.10$, only



(a) Estimated overall throughput



(b) Estimated data rate per user

Figure 6.11: Relative differences of the estimated user data rates with imperfect CSI measurement ($U = 15$, $I_R = 8$).

76.6% and 77.8% statistical relative differences are within the range of ±10.0% in the 2-user and 3-user SC-NOMA systems, respectively. As ε is reduced by half ($\varepsilon = 0.05$), these proportions rise to 96.1% and 99.4%. Therefore, high-precision channel measurements are always needed for improving the data rate estimation accuracy.

6.6 Summary

In this chapter, we studied the DCPA problem for SC-NOMA systems with the PF objective and analyzed its performance. Firstly, we derived the optimal solution to the DCPA problem in the ideal SC-NOMA system. Its performance is an upper bound and serves as a benchmark. Based on this optimal DCPA solution, we designed a low-complexity algorithm for jointly solving the power allocation and

USS problems. In order to reduce the computational complexity for USS in practical SC-NOMA systems, we further designed two USS schemes based on our optimal power allocation, namely, the TSU and PSU schemes. Then, the upper bound performance obtained by the optimal DCPA solution in the ideal SC-NOMA system has been analyzed based on the stochastic channel models.

The simulation results indicate that the NOMA system outperforms OMA in terms of both overall and cell-edge throughput. The performance gain increases with U and S due to the increased multi-user diversity gain brought by PFS. In particular, the upper bound performance is achieved when $S = U$, which is consistent with our analysis. In addition, our proposed TSU scheme obtains the optimal performance and costs lower computational complexity than the FUSC scheme when the number of multiplexed users is larger than 2. Moreover, the PSU scheme obtains close-to-optimal performance and its USS complexity is nearly linear in the number of users. By comparing our analytical performance to the simulation results, it is confirmed that the data rate estimation based on our analysis is very accurate.

We also investigated the impact of partial and imperfect CSI on the data rate estimation accuracy. With more than 4 reported IRSRPs, the estimation errors caused by inaccurate CQI models can be relieved. Hence, only a small amount of CSI is necessary for ensuring the estimation accuracy so that the signalling overhead can be reduced. The imperfect CSI measurements have a negligible influence on the estimated overall throughput but lead to a larger deviation of the estimated user data rates. Thus, it is necessary to enhance the channel measurement precision in order to guarantee high estimation accuracy.

7 Dynamic Channel and Power Allocation for Multi-Channel NOMA

In this chapter, we extend our research on the dynamic channel and power allocation (DCPA) schemes in the SC-NOMA system to the multi-channel NOMA (MC-NOMA) system in which multiple subchannels are utilized for NOMA transmission. Due to the variance and fluctuation of wireless channel states in the frequency domain, multi-channel diversity gain can be achieved by inter-channel power allocation in MC-NOMA systems. The DCPA problem is more complex for MC-NOMA since it is necessary to design the intra- and inter-channel power allocation jointly. Same as the DCPA problems in the SC-NOMA system, we adopt proportional fairness (PF) as the optimization objective.

In the literature, there have been several DCPA schemes proposed for downlink MC-NOMA systems as we introduced in Chapter 2, such as the matching game algorithm, iteration-based dynamic programming, monotonic optimization, and difference of convex programming. The main drawback of these schemes lies in their high computational complexity when the numbers of users and subchannels are large. On the other hand, the performance of MC-NOMA systems has not been studied analytically to the best of our knowledge. Hence, a comprehensive study on the DCPA schemes and their performance analysis is significant and desired for MC-NOMA.

In downlink MC-NOMA systems, a subchannel can be allocated to multiple user signals by using SIC for multi-user detection. On each subchannel, the multi-user diversity is determined by not only independent users but also their combinations. Similar to the SC-NOMA case in Chapter 6, we consider two MC-NOMA system models that are defined as follows.

- ▷ *Practical MC-NOMA system*: The maximum number of multiplexed users per subchannel is controlled by a pre-defined parameter S due to the limited processing capability of SIC receivers.
- ▷ *Ideal MC-NOMA system*: The limitation on the number of multiplexed users per subchannel is relaxed. Hence, an arbitrary number of users can be multiplexed within the same subchannel simultaneously.

In Section 7.1, we derive the optimal solution to the DCPA problem for the ideal MC-NOMA system based on our work in Chapter 6. The basic decomposition method is utilized in order to solve the intra- and inter-channel power allocation problems in two stages, respectively. The optimal solution to the intra-channel power

allocation can be obtained as in the SC-NOMA system. We design a low-complexity water-filling algorithm for solving the optimal inter-channel power allocation. The performance of this optimal DCPA scheme in the ideal MC-NOMA system is proved to be an upper bound for practical MC-NOMA. In Section 7.3, we utilize this optimal DCPA scheme and propose a user-preselection (UP)-based scheme for practical MC-NOMA systems with the aim of reducing the computational complexity.

Then, we analyze the upper bound performance obtained in the ideal MC-NOMA system in Section 7.4. In Section 7.5, the performance of our proposed UP-based DCPA scheme is evaluated by simulations and compared to the optimal scheme in the literature as well as the upper bound. We also investigate the impacts on multi-user and multi-channel diversities on the system performance. In addition, we use the analytical results of the upper bound performance to estimate the user data rates in practical MC-NOMA systems and evaluate its estimation accuracy under different system configurations. Finally, we summarize this chapter in Section 7.6.

7.1 System Model

We consider a downlink MC-NOMA system where the BS transmits to multiple user receivers with K subchannels. The index set of subchannels is denoted as

$$\mathbf{K} = \{\, k |\, k = 1, \ldots, K \}\,. \tag{7.1}$$

For ease of expression, all power values in the system model are normalized by the average power budget per subchannel. Hence, the total transmit power budget over the K subchannels is $p_t = K$. In each scheduling frame, the BS assigns power to the user signals and subchannels. In the following part, we focus on the DCPA problem in a given scheduling frame.

All of the K subchannels are assumed to have independent and identically distributed (i.i.d.) channel states for a given user [88]. The instantaneous channel quality indicator (CQI) of user u on subchannel k is defined as

$$\Phi_{k,u} = H_{k,u}/\sigma_{k,u}, \tag{7.2}$$

where $H_{k,u}$ is the comprehensive channel gain of user u on subchannel k, including the large-scale slow fading and the Rayleigh fast fading, and $\sigma_{k,u}$ is the normalized power of useless signals on subchannel k, including the additive white Gaussian noise and inter-cell interference that is not canceled by SIC. Thus, the instantaneous CQI $\Phi_{k,u}$ is the SINR of user u on subchannel k while one unit of normalized power is allocated to its signal on the subchannel.

The multiplexed users on subchannel k are sorted in descending order of their instantaneous CQIs and are collected in a user vector. It is denoted as

$$\mathbf{s}_k = \{\, c_k\,(i)\,|i = 1, \ldots, s_k \}\,, \tag{7.3}$$

where

$$\Phi_{k,c_k(i)} \geq \Phi_{k,c_k(i+1)}, \quad i = 1, \ldots, s_k - 1, \tag{7.4}$$

$s_k = |\mathbf{s}_k|$ is the number of multiplexed users on subchannel k, and $c_k(i)$ is the user with the i-th highest CQI in vector $\mathbf{s}_k$. The transmit power allocated to each user on subchannel k is expressed as

$$p_{k,u} > 0, \quad \forall u \in \mathbf{s}_k; \tag{7.5}$$

$$p_{k,u} = 0, \quad \forall u \in (\mathbf{U}/\mathbf{s}_k). \tag{7.6}$$

In downlink scenarios, SIC is carried out by user receivers on each subchannel for decoding. The i-th user in $\mathbf{s}_k$ decodes and cancels successively the interference signals of user $c_k(i+1) \sim c_k(s_k)$ in reverse order [70]. Due to the finite processing capability of the SIC receiver in practice, the number of multiplexed users per subchannel is limited by a predefined parameter that is denoted as S. Thus, we have $s_k \leq S$, $\forall k \in \mathbf{K}$. The limitation S is normally set to a small integer for reducing the implementation difficulty of SIC, e.g., $S = 2$ or 3, which means that at most 2 or 3 users can be multiplexed on each subchannel [88, 96].

We denote the cumulative power (CP) allocated to the first i users in $\mathbf{s}_k$ as

$$q_{k,i} = \sum_{j=1}^{i} p_{k,c_k(j)}, \quad i = 1, \ldots, s_k. \tag{7.7}$$

Thus, the total power allocated to subchannel k can be expressed as

$$p_k = q_{k,s_k} = \sum_{u \in \mathbf{U}} p_{k,u}. \tag{7.8}$$

For ease of derivation, we define $q_{k,0} = 0$, $k \in \mathbf{K}$. Due to the positive power values allocated to the multiplexed users in $\mathbf{s}_k$, we have the relationship as follows,

$$q_{k,i-1} < q_{k,i}, \quad i = 1, \ldots, s_k. \tag{7.9}$$

By (7.9), we define an ascending CP vector as

$$\mathbf{q}_k = \{q_{k,i} \,|i = 0, \ldots, s_k\}. \tag{7.10}$$

Thus, the power allocated to the i-th user in $\mathbf{s}_k$ is

$$p_{k,c(i)} = q_{k,i} - q_{k,i-1}, \quad i = 1, \ldots s_k. \tag{7.11}$$

The instantaneous obtainable data rate of the i-th multiplexed user on subchannel k is calculated as

$$r_{k,c_k(i)} = B_{sc} \ln \left(\frac{1 + q_{k,i} \Phi_{k,c_k(i)}}{1 + q_{k,i-1} \Phi_{k,c_k(i)}} \right), \quad i = 1, \ldots s_k. \tag{7.12}$$

where B_{sc} is the bandwidth per subchannel. The unit of the user data rate in (7.12)

is nat/s. If user u is not multiplexed on subchannel k, we have

$$r_{k,u} = 0, \quad \forall u \in (\mathbf{U}/\mathbf{s}_k)\,. \tag{7.13}$$

With (7.12) and (7.13), we denote the data rate of user u on subchannel k as a function $r_{k,u}(\mathbf{s}_k, \mathbf{q}_k)$. Thus, the instantaneous data rate of user u over all subchannels is

$$r_u(\mathbf{S}, \mathbf{Q}) = \sum_{k \in \mathbf{K}} r_{k,u}(\mathbf{s}_k, \mathbf{q}_k), \tag{7.14}$$

where $\mathbf{S} = \{\mathbf{s}_k | \, k \in \mathbf{K}\}$ and $\mathbf{Q} = \{\mathbf{q}_k | \, k \in \mathbf{K}\}$.

The PF scheduling factor is defined as

$$\begin{aligned} \omega(\mathbf{S}, \mathbf{Q}) &= \sum_{u \in \mathbf{U}} \frac{r_u(\mathbf{S}, \mathbf{Q})}{R_u} \\ &= \sum_{k \in \mathbf{K}} \sum_{u \in \mathbf{U}} \frac{r_{k,u}(\mathbf{s}_k, \mathbf{q}_k)}{R_u} \\ &= \sum_{k \in \mathbf{K}} \omega_k(\mathbf{s}_k, \mathbf{q}_k), \end{aligned} \tag{7.15}$$

where $\omega_k(\mathbf{s}_k, \mathbf{q}_k)$ is the scheduling factor function of subchannel k, and R_u is the EMA data rate of user u. It is updated in each scheduling scheme by

$$\begin{aligned} R_u(t+1) &= \left(1 - \frac{1}{\tau}\right) R_u(t) + \frac{1}{\tau} r_u(t) \\ &= \left(1 - \frac{1}{\tau}\right) R_u(t) + \frac{1}{\tau} \sum_{k \in \mathbf{K}} r_{k,u}(t). \end{aligned} \tag{7.16}$$

where $r_u(t)$ is the obtained data rate of user u in the t-th scheduling frame. It is the sum of the instantaneous data rates over the K subchannels.

7.2 Relaxed DCPA Problem for Ideal MC-NOMA

In this section, we develop the optimal DCPA scheme for the ideal MC-NOMA system and design a low-complexity algorithm for inter-channel power allocation based on our power allocation solution for the ideal SC-NOMA system in Section 6.2.

The optimization problem of DCPA in the ideal MC-NOMA system is formulated

as follows.

$$P7.1: \quad \max_{\mathbf{S},\mathbf{Q}} \omega\left(\mathbf{S},\mathbf{Q}\right) \tag{7.17a}$$

$$s.t., \quad \mathbf{s}_k \subseteq \mathbf{U},\ \forall k \in \mathbf{K}, \tag{7.17b}$$

$$q_{k,i-1} < q_{k,i},\ i = 1,\ldots,s_k,\ \forall k \in \mathbf{K}, \tag{7.17c}$$

$$\sum_{k\in\mathbf{K}} q_{s_k} \leq K. \tag{7.17d}$$

Constraint (7.17c) ensures that the multiplexed users on each subchannel have positive allocated power. Constraint (7.17d) limits the total transmit power over all subchannels by the power budget $p_t = K$.

The optimal solution to $P7.1$ can be solved by the basic decomposition method in [137] due to the fact that

- ▷ the overall scheduling factor is the sum of $\omega_k\left(\mathbf{s}_k,\mathbf{q}_k\right)$ over all subchannels as shown in (7.15);
- ▷ given total transmit power $p_k \in [0,K]$ allocated to subchannel k, its scheduling factor function $\omega_k\left(\mathbf{s}_k,\mathbf{q}_k\right)$ is determined only by the intra-channel power allocation.

We describe the decomposed subproblem and master problem as in the following two stages.

Stage 1) For every subchannel, given any power $p_k \in [0,K]$ allocated to it, we solve the optimal intra-channel power allocation and obtain the function of the maximum obtainable scheduling factor w.r.t. p_k. This subproblem is formulated as

$$P7.1.1: \quad \max_{\mathbf{s}_k,\mathbf{q}_k} \omega_k\left(\mathbf{s}_k,\mathbf{q}_k\right) \tag{7.18a}$$

$$s.t., \quad \mathbf{s}_k \subseteq \mathbf{U}, \tag{7.18b}$$

$$q_{k,i-1} < q_{k,i},\ i = 1,\ldots,s_k, \tag{7.18c}$$

$$q_{s_k} = p_k. \tag{7.18d}$$

The maximum scheduling factor function of subchannel k in $P7.1.1$ is denoted as $\hat{\omega}_k\left(p_k\right)$, where $p_k \in [0,K]$.

Stage 2) Based on the maximum scheduling factor functions obtained on the K subchannels in the first stage, we solve the optimal inter-channel power allocation with the objective of maximizing their sum. This master problem

is formulated as

$$P7.1.2: \quad \max_{\mathbf{p}} \sum_{k\in\mathbf{K}} \hat{\omega}_k\left(p_k\right) \tag{7.19a}$$

$$s.t. \quad p_k \geq 0,\ k \in \mathbf{K}, \tag{7.19b}$$

$$\sum_{k\in\mathbf{K}} p_k \leq K, \tag{7.19c}$$

where $\mathbf{p} = \{p_k | \, k \in \mathbf{K}\}$. Its optimal solution for $P7.1.2$ is denoted as $\mathbf{p}^* = \{p_k^* | \, k \in \mathbf{K}\}$.

Finally, by substituting p_k^* into $P7.1.1$ in the first stage, we can obtain the optimal solutions on each subchannel, which are denoted as $\mathbf{s}_k^*$ and $\mathbf{q}_k^*$. Accordingly, the optimal solutions to $P7.1$ are denoted as $\mathbf{S}^* = \{\mathbf{s}_k^* | \, k \in \mathbf{K}\}$ and $\mathbf{Q}^* = \{\mathbf{q}_k^* | \, k \in \mathbf{K}\}$.

7.2.1 Optimal Solution to Subproblem *P*7.1.1

Based on the optimal solution of the power allocation for the SC-NOMA system, we solve the subproblem $P7.1.1$ as follows. We further derive the scheduling factor function $\omega_k\left(\mathbf{s}_k, \mathbf{q}_k\right)$ as

$$\begin{aligned}
\omega_k\left(\mathbf{s}_k, \mathbf{q}_k\right) =& B_{sc} \sum_{i=1}^{s_k} \frac{1}{R_{c_k(i)}} \ln\left(\frac{1+q_{k,i}\Phi_{k,c_k(i)}}{1+q_{k,i-1}\Phi_{k,c_k(i)}}\right) \\
=& B_{sc} \sum_{i=1}^{s_k} \int_{q_{k,i-1}}^{q_{k,i}} \frac{\Phi_{k,c_k(i)}}{R_{c_k(i)}\left(1+x\Phi_{k,c_k(i)}\right)} dx \\
=& B_{sc} \sum_{i=1}^{s_k} \int_{q_{k,i-1}}^{q_{k,i}} \pi_{k,c_k(i)}\left(x\right) dx,
\end{aligned} \tag{7.20}$$

where $\pi_{k,c_k(i)}$ is a derivative function defined as

$$\pi_{k,u}\left(x\right) = \frac{\Phi_{k,u}}{R_u\left(1+x\Phi_{k,u}\right)}, \quad x \in [0, K]. \tag{7.21}$$

It is positive, continuous and monotonically decreasing (PCMD).

By Theorem 6.1, the maximum scheduling factor function of subchannel k with power p_k allocated on it is calculated as

$$\hat{\omega}_k\left(p_k\right) = B_{sc} \int_0^{p_k} \max_{u\in\mathbf{U}} \pi_{k,u}\left(x\right) dx. \tag{7.22}$$

For ease of expression, we denote the maximum derivative function of subchannel k

in (7.22) as

$$\hat{\pi}_k(x) = \max_{u \in \mathbf{U}} \pi_{k,u}(x), \quad x \in [0, K]. \tag{7.23}$$

Therefore, $\hat{\pi}_k(x)$ is the derivative of the maximum scheduling factor function $\hat{\omega}_k(p_k)$ on subchannel k w.r.t. p_k. It is also a PCMD function. Therefore, $\hat{\omega}_k(p_k)$ is a continuously differentiable and monotonically increasing concave function.

In order to derive $\hat{\pi}_k(x)$, we first assume that $p_k = K$ and utilize the optimal DCPA solution that have been developed for the ideal SC-NOMA system in Section 6.2. Under this condition, the optimal multiplexed user vector is

$$\begin{aligned} \hat{\mathbf{s}}_k &= \{\hat{c}_k(1), \hat{c}_k(2), \ldots, \hat{c}_k(\hat{s}_k)\} \\ &= \left\{ \arg\max_{u \in \mathbf{U}} \pi_{k,u}(x) \,|x \in [0, K] \right\}, \end{aligned} \tag{7.24}$$

where $\hat{s}_k = |\hat{\mathbf{s}}_k|$. The optimal intra-channel power allocation is

$$\hat{\mathbf{q}}_k = \{\hat{q}_{k,0}, \hat{q}_{k,1}, \ldots, \hat{q}_{k,\hat{s}_k-1}, \hat{q}_{k,\hat{s}_k}\}, \tag{7.25}$$

where

$$\hat{q}_{k,i} = \begin{cases} 0, & i = 0, \\ \theta_k(\hat{c}_k(i), \hat{c}_k(i+1)), & i = 1, \ldots, \hat{s}_k - 1, \\ K, & i = \hat{s}_k, \end{cases} \tag{7.26}$$

where $\theta_k(u, v)$ is defined as

$$\theta_k(u, v) = \frac{R_u^{-1}\Phi_{k,v}^{-1} - R_v^{-1}\Phi_{k,u}^{-1}}{w_v - R_u^{-1}}, \quad u, v \in \mathbf{U}. \tag{7.27}$$

The detailed procedure for solving $\hat{\mathbf{s}}_k$ and $\hat{\mathbf{q}}_k$ can be referred to Algorithm 6.1. In Section 6.5, its computational complexity has been verified to be linear in the number of users U for each subchannel.

With $\hat{\mathbf{s}}_k$ and $\hat{\mathbf{p}}_k$, the maximum derivative function $\hat{\pi}_k(x)$ in (7.23) is expressed as

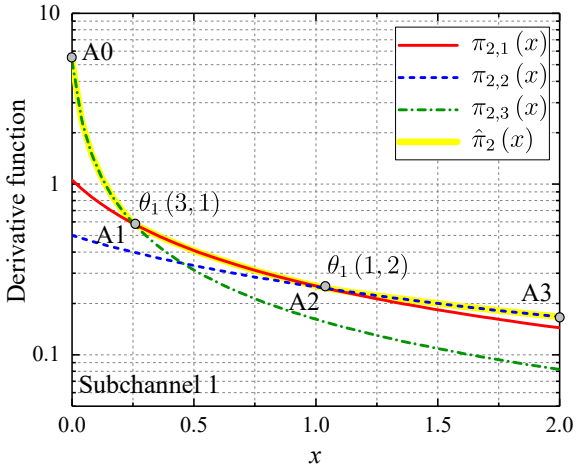
$$\begin{aligned} \hat{\pi}_k(x) = \pi_{k,\hat{c}_k(i)}(x), \quad & x \in [\hat{q}_{k,i-1}, \hat{q}_{k,i}], \\ & i = 1, \ldots, \hat{s}_k. \end{aligned} \tag{7.28}$$

Therefore, it consists of the derivative functions of the users in $\hat{\mathbf{s}}_k$, which are maximum within $\hat{s}_k$ adjacent intervals of $x \in [0, K]$. These selected users are in descending order of their CQIs as x increases. By substituting (7.28) into (7.22), we can obtain the function of the maximum scheduling factor on subchannel k w.r.t. p_k, i.e., $\hat{\omega}_k(p_k)$.

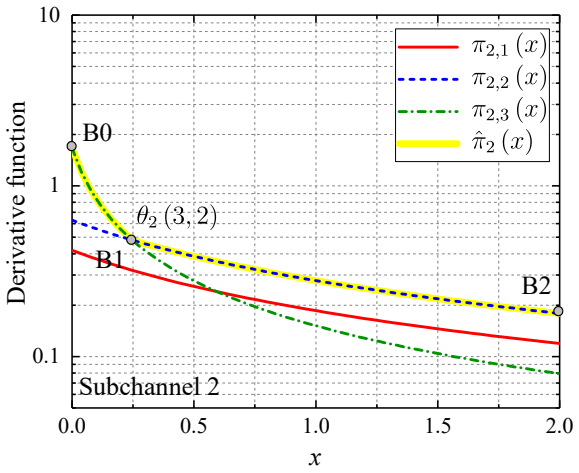
To illustrate the above relationship among derivative functions, we present an example of an MC-NOMA system with 2 subchannels and 3 users in Figure 7.1. The user EMA data rates and CQIs on the subchannels are given in Table 7.1. On subchannel 1, $\hat{\pi}_1(x)$ consists of the derivative functions that are maximum within 3 adjacent intervals of $x \in [0, K]$, as shown by the multi-segment curve

Table 7.1: An Example of the Derivative Functions in the MC-NOMA System

User	**EMA Data Rate**	**Instantaneous CQI**	
u	R_u	$\Phi_{1,u}$	$\Phi_{2,u}$
1	3	5 dB	1 dB
2	2	0 dB	1 dB
3	6	15 dB	10 dB



(a) Subchannel 1



(b) Subchannel 2

Figure 7.1: An example of the derivative functions in the MC-NOMA system ($U = 3$, $K = 2$).

A0-A1-A2-A3. The corresponding user vector is $\hat{\mathbf{s}}_1 = \{3, 1, 2\}$ and the CP vector is $\hat{\mathbf{q}}_1 = \{0, 0.253, 1.052, 2\}$. On subchannel 2, there are only 2 users in $\hat{\mathbf{s}}_2 = \{3, 2\}$ because $\pi_{2,1}(x)$ is lower than $\pi_{2,2}(x)$ and thus not maximum anywhere within $x \in [0, K]$. In this case, user 1 cannot be selected for user multiplexing on

subchannel 2. Therefore, $\hat{\pi}_2(x)$ consists of only $\pi_{2,2}(x)$ and $\pi_{2,3}(x)$, as shown by the multi-segment curve B0-B1-B2 in Figure 7.1. The CP vector on subchannel 2 is $\hat{\mathbf{q}}_2 = \{0, 0.247, 2\}$.

7.2.2 Optimal Solution to Master Problem *P*7.1.2

Due to the concavity of the function $\hat{\omega}_k(p_k)$, the master problem $P7.1.2$ is a water-filling inter-channel power allocation problem and can be solved by the method of Lagrange multiplier [138]. We construct the Lagrange function as

$$\mathcal{L}(\mathbf{p}, \rho) = \sum_{k \in \mathbf{K}} \hat{\omega}_k(p_k) - \rho \left(\sum_{k \in \mathbf{K}} p_k - K \right). \tag{7.29}$$

Although the derivative of $\hat{\omega}_k(p_k)$ is PCMD, it may be composed of multiple derivative functions with different parameters as shown in (7.28). Therefore, the classic water-filling algorithm cannot be directly applied to solving $P7.1.2$. We design a low-complexity algorithm as in Algorithm 7.1 to solve the optimal water-level, which is denoted as

$$L^* = 1/\rho^*, \tag{7.30}$$

where ρ^* is the optimal multiplier in (7.29). The optimal inter-channel power allocation $\mathbf{p}^*$ is derived at the end of the algorithm. Note that a subchannel may have no power assigned while its maximum derivative function $\hat{\pi}_k(x)$ is lower than the optimal multiplier ρ^*. This case happens if all of the users have very low CQIs on the subchannel. We denote the set of subchannels that have positive allocated power as $\mathbf{K}^*$, i.e.,

$$\mathbf{K}^* = \{k \,|\, p_k^* > 0, k \in \mathbf{K}\}. \tag{7.31}$$

The inverse function of the maximum derivative function exists because it is PCMD. It is denoted as $\hat{\pi}_k^{-1}(x)$ and can be computed by the piecewise function that consists of multiple segments of the derivative functions belonging to the multiplexed users on subchannel k.

In Figure 7.2, we illustrate the procedure of Algorithm 7.1 with the example given in Figure 7.1. First, $\hat{\pi}_k(x)$ is calculated at each CP point in $\hat{\mathbf{q}}_k$ and is denoted as $\rho_{k,i}$. The number of multiplexed users in the optimal user vector is initialized as $s_k^* = 0$. Then, in the while loop, we select the maximum ρ_{k,s_k^*} in each iteration. We set the multiplier ρ^+ to this value and judge if the corresponding sum power of all subchannels exceeds the total power limit K. As shown in Figure 7.2, the order of selected points for ρ^+ is A0-B0-A1-B1-A2. In each iteration, if the sum power does not exceed K, the number of the multiplexed users in the selected subchannel is increased by one and the subchannel is added into $\mathbf{K}^*$. Otherwise, the while loop is terminated.

In the example shown in Figure 7.2, when point A2 $(0.244, 4.387)$ is selected, the corresponding sum power with $\rho^+ = 0.244$ is $P_s = 2.308$ and is larger than $K = 2$. Then, the while loop stops. The numbers of multiplexed users are $s_1^* = 2$ and $s_2^* = 2$,

Algorithm 7.1 Optimal Solution to $P7.1.2$

$\mathbf{K}^* = \varnothing$
for $k \in \mathbf{K}$ **do**
  $\rho_{k,i} = \hat{\pi}_k\left(\hat{q}_{k,i}\right)$, $i = 0, \ldots s_k$
  $s_k^* = 0$
end for
while 1 **do**
  $\rho^+ = \max\limits_{k \in \mathbf{K}} \rho_{k,s_k^*}$
  $k^+ = \arg\max\limits_{k \in \mathbf{K}} \rho_{k,s_k^*}$
  $P_s = \hat{\pi}_{k^+}^{-1}\left(\rho^+\right) + \sum\limits_{k \in \mathbf{K}^*}^{k \neq k^+} \hat{\pi}_k^{-1}\left(\rho^+\right)$
  if $P_s < K$ **then**
    $s_{k^+}^* = s_{k^+}^* + 1$
    $\mathbf{K}^* = \mathbf{K}^* \cup \{k^+\}$
  else
    BREAK
  end if
end while
The optimal water level:

$$L^* = \frac{1}{\rho^*} = \frac{K + \sum\limits_{k \in \mathbf{K}^*} \Phi_{k,\hat{c}_k\left(s_k^*\right)}^{-1}}{\sum\limits_{k \in \mathbf{K}^*} R_{\hat{c}_k\left(s_k^*\right)}^{-1}} \tag{7.32}$$

The optimal inter-channel power allocation ($\mathbf{p}^*$):
for $k \in \mathbf{K}$ **do**
  if $k \in \mathbf{K}^*$ **then**
    $p_k^* = \hat{\pi}_k^{-1}\left(\rho^*\right) = L^* R_{\hat{c}_k\left(s_k^*\right)}^{-1} - \Phi_{k,\hat{c}_k\left(s_k^*\right)}^{-1}$
  else
    $p_k^* = 0$
  end if
end for

i.e., the first two users in the $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$ are selected, respectively. Therefore, we use the derivative functions, $\pi_{1,\hat{c}(2)}\left(x\right) = \pi_{1,1}\left(x\right)$ and $\pi_{2,\hat{c}(2)}\left(x\right) = \pi_{2,2}\left(x\right)$, to calculate the optimal multiplier ρ^* and the optimal water-level L^* as in (7.32). Then, the optimal inter-channel power allocation can be obtained by substituting ρ^* into the inverse functions of $\pi_{1,1}\left(x\right)$ and $\pi_{2,2}\left(x\right)$. At the end of the algorithm, we obtain $\mathbf{p}^* = \{0.928, 1.072\}$.

After solving the subproblem in $P7.1.1$ and the master problem in 7.1.2, we substitute the optimal p_k^* into $\hat{\omega}_k\left(p_k\right)$ and obtain the maximum scheduling factor

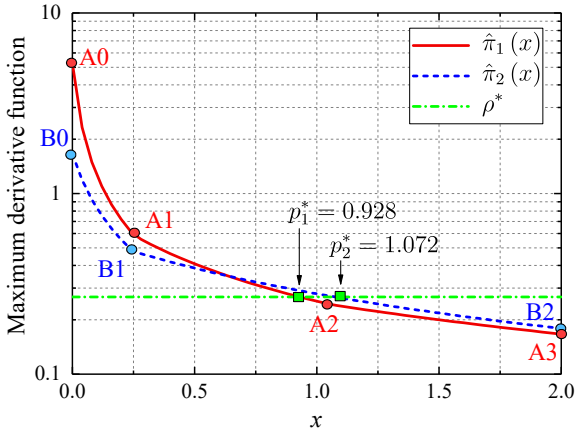


Figure 7.2: An example of the maximum derivative functions in the MC-NOMA system.

per subchannel. The optimal user vector per subchannel is

$$\mathbf{s}_k^* = \{\hat{c}_k(1), \hat{c}_k(2), \ldots, \hat{c}_k(s_k^*)\}, \tag{7.33}$$

which contains the first s_k^* users in $\hat{\mathbf{s}}_k$. The optimal CP vector for each subchannel is given as

$$\mathbf{q}_k^* = \left\{\hat{q}_{k,0}, \hat{q}_{k,1}, \ldots, \hat{q}_{k,s_k^*-1}, p_k^*\right\}, \tag{7.34}$$

Thus, the optimal solution for the example shown in Figure 7.2 is given as

$$\begin{aligned}
\mathbf{s}_1^* &= \{3, 1\}, \\
\mathbf{s}_2^* &= \{3, 2\}, \\
\mathbf{q}_1^* &= \{0, 0.253, 0.928\}, \\
\mathbf{q}_2^* &= \{0, 0.247, 1.072\}.
\end{aligned}$$

The computational complexity of Algorithm 7.1 is mainly determined by its while loop. When every subchannel has $s_k^* = \hat{s}_k$, the number of iterations reaches its maximum, i.e.,

$$\sum_{k \in \mathbf{K}} \hat{s}_k + 1. \tag{7.35}$$

Since $\hat{s}_k \leq U$, the number of iterations is bounded by $(KU + 1)$. In each iteration, K subchannels are compared in line 7 of the algorithm. Therefore, the overall computational complexity of Algorithm 7.1 is $O\left(K^2 U\right)$.

7.3 Suboptimal DCPA for Practical MC-NOMA

In this section, we address the DCPA problem for practical MC-NOMA systems with a limited number of multiplexed users per subchannel. With this additional

constraint, we formulate the optimization problem of DCPA as follows.

$$P7.2: \quad \max_{\mathbf{S},\mathbf{Q}} \omega(\mathbf{S},\mathbf{Q}) \tag{7.36a}$$

$$s.t., \quad \mathbf{s}_k \subseteq \mathbf{U},\ \forall k \in \mathbf{K}, \tag{7.36b}$$

$$q_{k,i-1} < q_{k,i},\ i = 1,\ldots,s_k,\ \forall k \in \mathbf{K}, \tag{7.36c}$$

$$\sum_{k\in\mathbf{K}} q_{s_k} \le K. \tag{7.36d}$$

$$s_k \le S, \quad \forall k \in \mathbf{K}, \tag{7.36e}$$

where parameter $S < U$. Note that $P7.1$ is a special case of $P7.2$ when $S = U$, i.e., all of the users can be multiplexed on each subchannel. Without the limitation on s_k in $P7.1$, it has a larger solution space than $P7.2$. Therefore, the performance of the ideal MC-NOMA system is an upper bound for the cases where $S < U$.

Similar to the optimal DCPA scheme for the ideal MC-NOMA system, we utilize the decomposed method to solve $P7.2$. However, since s_k is not limited by S in $P7.1$, its optimal solution may be infeasible for $P7.2$. In order to limit the number of multiplexed users per subchannel, we design a user-preselection (UP)-based DCPA scheme as follows.

We first assume that the total power budget is equally allocated to every subchannel, i.e., $p_k = 1$, $\forall k \in \mathbf{K}$. In this case, we have K parallel SC-NOMA systems, each of which generates an independent DCPA problem. In the literature, there have been several DCPA schemes proposed with the PF target for practical SC-NOMA systems, such as the TTPA scheme [11] and the iterative water-filling scheme [10]. We utilize the PSU scheme proposed in Section 6.3 due to its advantage of extremely low complexity. The detailed description of the PSU scheme is given in Algorithm 6.2. The multiplexed user and CP vectors on subchannel k obtained by PSU are denoted as $\tilde{\mathbf{s}}_k$ and $\tilde{\mathbf{q}}_k$, respectively.

Then, we consider only the users in $\tilde{\mathbf{s}}_k$ on subchannel k as its candidate multiplexed users for $P7.1.2$ and set the power allocated to other users as 0 on this subchannel, i.e.,

$$p_{k,u} \ge 0, \quad \forall u \in \tilde{\mathbf{s}}_k; \tag{7.37}$$

$$p_{k,u} = 0, \quad \forall u \notin \tilde{\mathbf{s}}_k. \tag{7.38}$$

Since $|\tilde{s}_k| = \tilde{\mathbf{s}}_k \le S$ now, the optimal solution to $P7.1.2$ in Section III can be utilized for $P7.2$. In this case, the number of iterations for the while loop in Algorithm 1 is bounded by $(KS+1)$. Thus, its computational complexity is $O\left(K^2S\right)$. Note that the user preselection is carried out for each subchannel without consideration of the inter-channel power allocation. Therefore, the above UP-based DCPA scheme obtains a suboptimal solution for $P7.2$.

7.4 Upper Bound Performance Analysis

In this section, we analyze the data rate performance of the ideal MC-NOMA system based on our optimal DCPA solution developed in Section 7.2.

7.4.1 Derivative Functions

Under fluctuating channel states, the derivative function $\pi_{k,u}(x)$ is a random variable determined by the EMA data rate R_u, the instantaneous user CQI $\Phi_{k,u}$ and variable x. We derive its conditional cumulative distribution function (CDF) given x as

$$\begin{aligned} F_{\pi_{k,u}}(y\,|x) =& \Pr\{\pi_{k,u}(x) < y\} \\ =& \Pr\left\{\frac{\Phi_{k,u}}{(\Phi_{k,u}x+1)R_u} < y\right\} \\ =& \Pr\left\{\Phi_{k,u} < \frac{yR_u}{1-yR_ux}\right\} \\ =& F_{\Phi_{k,u}}\left(\frac{yR_u}{1-yR_ux}\right), \\ & y \in \left(0, \frac{1}{xR_u}\right), \quad x \in [0, K]. \end{aligned} \tag{7.39}$$

where $F_{\Phi_{k,u}}$ is the CDF of the CQI of user u on subchannel k. Thus, the conditional probability density function (PDF) of $\pi_{k,u}(x)$ is derived as follows,

$$\begin{aligned} f_{\pi_{k,u}}(y\,|x) &= \frac{\partial F_{\pi_{k,u}}(y\,|x)}{\partial y} \\ &= f_{\Phi_{k,u}}\left(\frac{gR_u}{1-yR_ux}\right)\frac{R_u}{(1-yR_ux)^2}. \end{aligned} \tag{7.40}$$

Then, the conditional CDF of the maximum derivative function given x on each subchannel is calculated as

$$\begin{aligned} F_{\hat{\pi}_k}(y\,|x) &= \Pr\{\hat{\pi}_k(x) < y\} \\ &= \Pr\left\{\max_{k\in\mathbf{K}}\{\pi_{k,u}(x)\} < y\right\} \\ &= \prod_{u\in\mathbf{U}} \Pr\{\pi_{k,u}(x) < y\} \\ &= \prod_{u\in\mathbf{U}} F_{\pi_{k,u}}(y\,|x). \end{aligned} \tag{7.41}$$

It is identical for all subchannels due to the i.i.d. channel states. In addition, the maximum derivative function $\hat{\pi}_k(x)$ is PCMD as we have clarified in Section 7.2.

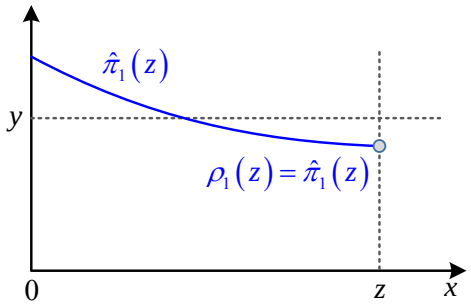


Figure 7.3: The relationship between $\rho_1(z)$ and $\hat{\pi}_1(z)$.

Therefore, its reverse function exists and has a conditional CDF as

$$\begin{aligned} F_{\hat{\pi}_k^{-1}}(x|y) &= \Pr\left\{\hat{\pi}_k^{-1}(y) < x\right\} \\ &= \Pr\left\{\hat{\pi}_k(x) < y\right\} \\ &= F_{\hat{\pi}_k}(y|x). \end{aligned} \tag{7.42}$$

This indicates that the conditional CDF of the maximum derivative function is identical with the conditional CDF of its reverse function on each subchannel.

7.4.2 Optimal Multiplier

Without loss of generality, we consider the first m subchannels in $\mathbf{K}$ and $m < K$. We assume that a power budget $z \in [0, K]$ is allocated to these m subchannels. Under this condition, the optimal multiplier for their inter-channel power allocation problem is denoted as a function of z, i.e., $\rho_m(z)$.

In particular, there is only one subchannel when $m = 1$. It obtains all of the transmit power z and its derivative function is shown in Figure 7.3. In this case, we have

$$\rho_1(z) = \hat{\pi}_1(z), \quad z \in [0, K]. \tag{7.43}$$

Therefore, $\rho_1(z)$ is a PCMD function as $\hat{\pi}_1(z)$.

When $m > 1$, we derive $\rho_m(z)$ and prove that it is also PCMD by induction as follows.

Proof. We first assume that $\rho_m(z)$ is a PCMD function when $m = n \geq 1$. To calculate $\rho_{n+1}(z)$, we consider three cases:

Case 1) If there exists $\hat{\pi}_{n+1}(x) \geq \rho_n(z - x)$, $\forall x \in [0, z]$, then all of the power z is allocated to the $(n+1)$-th subchannel. Therefore, $\rho_{n+1}(z) = \hat{\pi}_{n+1}(z)$ and it is PCMD as $\hat{\pi}_{n+1}(z)$.

Case 2) If there exists $\hat{\pi}_{n+1}(x) \leq \rho_n(z-x)$, $\forall x \in [0,z]$, then all of the power z is allocated to the first n subchannels. Therefore, $\rho_{n+1}(z) = \rho_n(z)$ and it is PCMD as $\rho_n(z)$.

Case 3) Otherwise, there exists a point $a \in (0,z)$ such that

$$\rho_{n+1}(z) = \hat{\pi}_{n+1}(a) = \rho_n(z-a). \tag{7.44}$$

Point a is the x coordinate of the intersection point of the monotonically decreasing function $\hat{\pi}_{n+1}(x)$ and the monotonically increasing function $\rho_n(z-x)$ w.r.t. x. In this case, the optimal power allocated to subchannel $(n+1)$ is a, and the rest power $(z-a)$ is allocated to the first n subchannels.

We denote the optimal multiplier in this case as $y = \rho_{n+1}(z)$. Since both $\hat{\pi}_{n+1}(x)$ and $\rho_n(x)$ are PCMD functions, their reverse functions exist and are also PCMD. By (7.44), we have the reverse function of $\rho_{n+1}(z)$ as

$$z = \rho_{n+1}^{-1}(y) = \hat{\pi}_{n+1}^{-1}(y) + \rho_n^{-1}(y). \tag{7.45}$$

This reverse function is the sum of the two PCMD functions, $\hat{\pi}_{n+1}^{-1}(y)$ and $\rho_n^{-1}(y)$. Therefore, the optimal multiplier $\rho_{n+1}(y)$ is also PCMD.

Combining the above three cases, when $m = n+1$, the optimal multiplier is expressed as

$$\rho_{n+1}(z) = \min_{x\in[0,z]} \max\left\{\hat{\pi}_{n+1}(x), \rho_n(z-x)\right\}, \tag{7.46}$$

and it is a PCMD function.

By induction, the optimal multiplier $\rho_m(z)$ is PCMD when $1 \leq m \leq K$. It can be calculated with the recursive function in (7.46) and the maximum derivative functions given in (7.28). In practice, this indicates that the optimal water-level $L^* = 1/\rho^*$ monotonically increases with the total transmit power budget z.

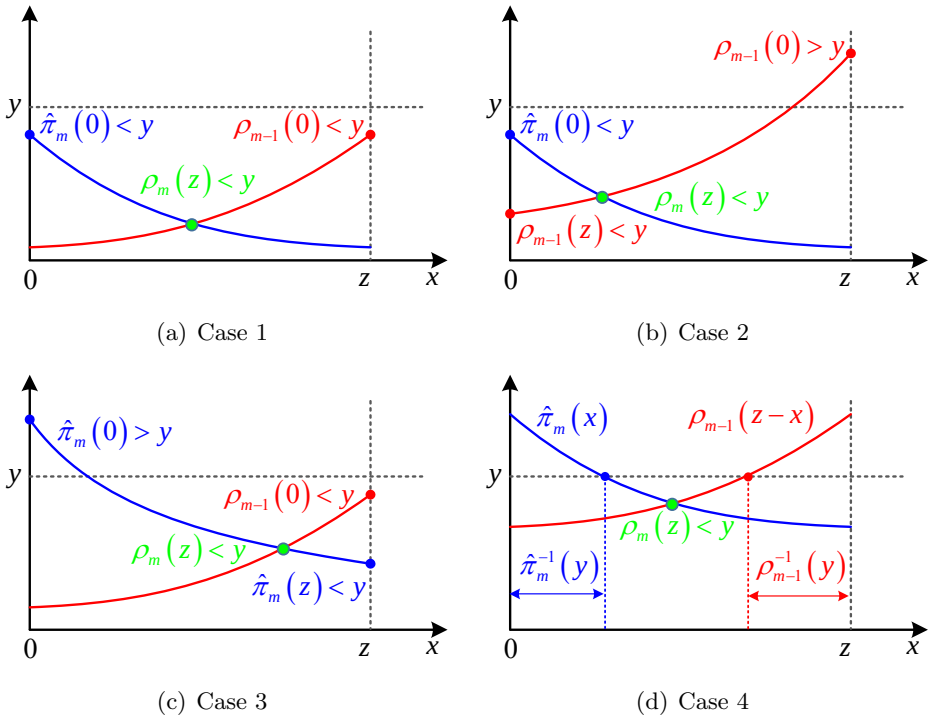
□

Under fluctuating channels, the optimal multiplier $\rho_m(z)$ is a random variable determined by the channel gains, the number of channels m and the power budget z. We define the conditional CDF of the optimal multiplier $\rho_m(z)$ given z as

$$G_{\rho_m}(y\,|z) = \Pr\left\{\rho_m(z) < y\right\}. \tag{7.47}$$

It is identical with the conditional CDF of the reverse function $\rho_m^{-1}(y)$, i.e.,

$$\begin{aligned} G_{\rho_m^{-1}}(z\,|y) &= \Pr\left\{\rho_m^{-1}(y) < z\right\} \\ &= \Pr\left\{\rho_m(z) < y\right\} \\ &= G_{\rho_m}(y\,|z). \end{aligned} \tag{7.48}$$



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Figure 7.4: Four cases of $\hat{\pi}_m(x)$, $\rho_{m-1}(z-x)$, and $\rho_m(z) < y$.

When $m = 1$, we have the special case that

$$G_{\rho_1}(y|z) = \Pr\{\rho_1(z) = \hat{\pi}_1(z) < y\} = F_{\hat{\pi}_1}(y|z). \tag{7.49}$$

To calculate $G_{\rho_m}(y|z)$ when $1 < m \leq K$, we utilize the recursive function in (7.46) and consider four cases of $\hat{\pi}_m(x)$, $\rho_{m-1}(z-x)$, and $\rho_m(z) < y$, as presented in Figure 7.4.

Case 1) When $\hat{\pi}_m(0) < y$ and $\rho_{m-1}(0) < y$, the optimal multiplier $\rho_m(z) < y$, as shown in Figure 7.4(a). The probability of this case is

$$P_{\mathrm{C1}} = F_{\hat{\pi}_m}(y|0)\, G_{\rho_{m-1}}(y|0). \tag{7.50}$$

Case 2) When $\hat{\pi}_m(0) < y$ and $\rho_{m-1}(z) < y < \rho_{m-1}(0)$, the optimal multiplier $\rho_m(z) < y$, as shown in Figure 7.4(b). The probability of this case is

$$P_{\mathrm{C2}} = \left[G_{\rho_{m-1}}(y|z) - G_{\rho_{m-1}}(y|0)\right] F_{\hat{\pi}_m}(y|0). \tag{7.51}$$

Case 3) When $\hat{\pi}_m(z) < y < \hat{\pi}_m(0)$ and $\rho_{m-1}(0) < y$, the optimal multiplier $\rho_m(z) < y$, as shown in Figure 7.4(c). The probability of this case is

$$P_{\mathrm{C3}} = \left[F_{\hat{\pi}_m}(y|z) - F_{\hat{\pi}_m}(y|0)\right] G_{\rho_{m-1}}(y|0). \tag{7.52}$$

Case 4) When $\hat{\pi}_m(z) < y < \hat{\pi}_m(0)$ and $\rho_{m-1}(z) < y < \rho_{m-1}(0)$, the two functions $\hat{\pi}_m(x)$ and $\rho_{m-1}(x)$ must have an intersection point in the range $x \in (0, z)$, as shown in Figure 7.4(d). By (7.46), the condition that optimal multiplier $\rho_m(z) < y$ is equivalent to

$$z < \rho_m^{-1}(y) = \hat{\pi}_m^{-1}(y) + \rho_{m-1}^{-1}(y). \tag{7.53}$$

Thus, the probability of $\rho_m(z) < y$ in this case is calculated as

$$\begin{aligned} P_{\mathrm{C4}} &= \Pr\left\{\hat{\pi}_m^{-1}(y) + \rho_{m-1}^{-1}(y) < z\right\} \\ &= \int_0^z \int_0^{z-x} \frac{\partial F_{\hat{\pi}_m^{-1}}(x\,|y)}{\partial x} \frac{\partial G_{\rho_{m-1}^{-1}}(t\,|y)}{\partial t} dt dx \\ &= \int_0^z \frac{\partial F_{\hat{\pi}_m}(y\,|x)}{\partial x} \left[G_{\rho_{m-1}^{-1}}(z-x\,|y) - G_{\rho_{m-1}^{-1}}(0\,|y)\right] dx \\ &= \int_0^z \frac{\partial F_{\hat{\pi}_m}(y\,|x)}{\partial x} G_{\rho_{m-1}}(y\,|z-x)\, dx \\ &\quad - F_{\hat{\pi}_m}(y\,|z)\, G_{\rho_{m-1}}(y\,|0) + F_{\hat{\pi}_m}(y\,|0)\, G_{\rho_{m-1}}(y\,|0). \end{aligned} \tag{7.54}$$

Combining all of the four cases analyzed above, we calculate the conditional CDF of the optimal multiplier $\rho_m(z)$ given z as follows,

$$\begin{aligned} G_{\rho_m}(y\,|z) &= P_{\mathrm{C1}} + P_{\mathrm{C2}} + P_{\mathrm{C3}} + P_{\mathrm{C4}} \\ &= \int_0^z \frac{\partial F_{\hat{\pi}_m}(y\,|x)}{\partial x} G_{\rho_{m-1}}(y\,|z-x)\, dx + F_{\hat{\pi}_m}(y\,|0)\, G_{\rho_{m-1}}(y\,|z). \end{aligned} \tag{7.55}$$

It is calculated based on $F_{\hat{\pi}_m}(y\,|x)$ and $G_{\rho_{m-1}}(y\,|x)$. Specially, when $m = 0$, the conditional CDF of $\rho_0(z)$ defined as

$$G_{\rho_0}(y\,|z) = 1 \tag{7.56}$$

Substituting (7.56) into (7.55), we obtain the identical conditional CDF of the optimal multiplier $\rho_1(z)$ as in (7.49), i.e.,

$$\begin{aligned} G_{\rho_1}(y\,|z) &= \int_0^z \frac{\partial F_{\hat{\pi}_1}(y\,|x)}{\partial x} dx + F_{\hat{\pi}_1}(y\,|0) \\ &= F_{\hat{\pi}_1}(y\,|x). \end{aligned} \tag{7.57}$$

When $1 < m \leq K$, $G_{\rho_m}(y\,|z)$ can be calculated with the recursive function in (7.55) and the conditional CDF $F_{\hat{\pi}_k}(y\,|x)$ in (7.41).

7.4.3 Ergodic User Data Rate

In particular, we calculate $G_{\rho_{K-1}}(y|z)$ with $m = K-1$ and use it to derive the expected value of the PF scheduling factor for user u on subchannel K as follows,

$$\overline{\omega}_{K,u} = B_{sc} \int_0^K \int_0^{\frac{1}{R_u x}} \underbrace{y f_{\pi_{K,u}}(y|x)}_{\text{(I)}} \underbrace{\prod_{v \in (\mathbf{U}/u)} F_{\pi_{K,v}}(y|x)}_{\text{(II)}} \underbrace{G_{\rho_{K-1}}(y|K-x)}_{\text{(III)}} dydx \quad (7.58)$$

where item (I) is the density of the scheduling factor with its conditional PDF given in (7.40), item (II) is the probability that user u has its derivative function larger than the other users within the same subchannel K at point x, and item (III) is the probability of $\hat{\pi}_K(x) > \rho_{K-1}(K-x)$, i.e., the derivative function of user u on subchannel K at point x is larger than the optimal multiplier of the other $(K-1)$ subchannels at point $(K-x)$. Specially, when $K = 1$ in the SC-NOMA system, item (III) is a constant value, i.e., $G_{\rho_0}(y|z) = 1$. Therefore, it results in the same formula as in (6.62). To solve the equation (7.58), we utilize the following theorem.

Theorem 7.1. *In the multi-channel transmission system with PFS, when $\tau \gg 1$, there exists the approximation that*

$$\sum_{k \in \mathbf{K}} \overline{r}_{k,u} \approx R_u. \quad (7.59)$$

Proof. According to the definition in (7.16), the expectation of the EMA data rate is expressed as

$$\mathbb{E}[R_u(t+1)] = \left(1 - \frac{1}{\tau}\right) \mathbb{E}[R_u(t)] + \frac{1}{\tau} \sum_{k \in \mathbf{K}} \mathbb{E}[r_{k,u}(t)]. \quad (7.60)$$

Assuming ergodicity for $R_u(t)$ for stable PFS, it holds that

$$\mathbb{E}[R_u(t+1)] = \mathbb{E}[R_u(t)]. \quad (7.61)$$

Substituting (7.61) into (7.60), we have

$$\mathbb{E}[R_u(t)] = \sum_{k \in \mathbf{K}} \mathbb{E}[r_{k,u}(t)] = \sum_{k \in \mathbf{K}} \overline{r}_{k,u}. \quad (7.62)$$

When the averaging coefficient $\tau \gg 1$, we have the approximation as

$$\frac{1}{\tau^T} \approx 0, \quad T \geq 2. \quad (7.63)$$

Considering again the ergodicity of $R_u(t)$, we assume that for a certain frame t there exists

$$R_u(t+T) = R_u(t), \quad (7.64)$$

which means that the status $R_u(t)$ repeats after a long enough period, i.e, T frames after frame t. Then, we can deduce $R_u(t)$ as follows,

$$R_u(t+T) = \left(1-\frac{1}{\tau}\right)^T R_u(t) + \sum_{n=0}^{T-1}\frac{1}{\tau}\left(1-\frac{1}{\tau}\right)^{T-n-1}\sum_{k\in\mathbf{K}} r_{k,u}(t+n)$$
$$\overset{(7.63)}{\approx} \left(1-\frac{T}{\tau}\right) R_u(t) + \sum_{n=0}^{T-1}\frac{1}{\tau}\sum_{k\in\mathbf{K}} r_{k,u}(t+n), \tag{7.65}$$

By (7.64), when T is large enough, we have

$$R_u(t) \approx \frac{1}{T}\sum_{n=0}^{T-1}\sum_{k\in\mathbf{K}} r_{k,u}(t+n) \approx \sum_{k\in\mathbf{K}} \overline{r}_{k,u}. \tag{7.66}$$

Combining (7.62) and (7.66), it is proved that

$$R_u(t) \approx \mathbb{E}[R_u(t)] = \sum_{k\in\mathbf{K}} \overline{r}_{k,u}, \quad \forall t. \tag{7.67}$$

□

Due to the i.i.d. channel states, the ergodic data rate of a user on every subchannel is identical, i.e.,

$$\overline{r}_{k,u} = \overline{r}_{K,u}, \quad \forall k \in K. \tag{7.68}$$

Thus, we have $R_u \approx K\overline{r}_{K,u}$. By Theorem 7.1, the expectation of the PF scheduling factor on subchannel K is

$$\overline{\omega}_{K,u} = \mathbb{E}\left[\frac{r_{K,u}}{R_u}\right] \approx \frac{\overline{r}_{K,u}}{K\overline{r}_{K,u}} = \frac{1}{K}. \tag{7.69}$$

Substituting (7.69) and the approximation $R_u \approx K\overline{r}_{K,u}$ into (7.58), we have the equations of the ergodic user data rates and can solve them by numerical methods [114]. Therefore, the sum data rate of a user over the K subchannels is calculated as

$$\overline{r}_u = K\overline{r}_{K,u}. \tag{7.70}$$

In addition, the average power ratio allocated to user u can be calculated with the estimated user data rates as follows.

$$\overline{p}_u = \mathbb{E}[p_u] = \sum_{k\in\mathbf{K}} \mathbb{E}[p_{k,u}]$$
$$= K\int_0^K \int_0^{\frac{1}{x\overline{r}_u}} f_{\pi_{K,u}}(y\,|x\,) \prod_{v\in(\mathbf{U}/u)} F_{\pi_{K,v}}(y\,|x\,) G_{\rho_{K-1}}(y\,|K-x\,)\,dydx \tag{7.71}$$

On a given subchannel, when the derivative function of user u is the maximum at

a power point $x \in [0, K]$, it obtains the corresponding power at this point. Thus, $\mathbb{E}[p_{k,u}]$ in (7.71) is the integral of the expected power density in the range of $x \in [0, K]$. The total obtained power is the sum on the K subchannels.

7.4.4 Allocated Power per Subchannel

The power allocated on each subchannel changes continuously due to the fluctuation of channel states and dynamic inter-channel power allocation. Under the i.i.d. channel states, the probability distribution of the allocated power per subchannel is identical. Without loss of generality, we calculate the CDF of the optimal power allocated to subchannel K. It is defined as

$$F_{p_K^*}(x) = \Pr\{p_K^* < x\}, \quad x \in (0, K]. \tag{7.72}$$

We derive it based on the three cases described as follows.

Case 1) When $\rho_{K-1}(K) \geq \hat{\pi}_K(0)$, there exists $\rho_{K-1}(K-x) > \hat{\pi}_K(x)$, $x \in (0, K]$, due to the monotonicity. Therefore, no power is allocated to subchannel K, i.e., $p_K^* = 0$. The total transmit power is allocated to the first $(K-1)$ subchannels. In this case, we have

$$\hat{\pi}_K(p_K^*) \leq \rho_{K-1}(K - p_K^*) \tag{7.73}$$

Thus, given any $x \in (0, K]$, it holds that $p_K^* = 0 < x$.

Case 2) When $\rho_{K-1}(0) \leq \hat{\pi}_K(K)$, there exists $\rho_{K-1}(K-x) < \hat{\pi}_K(x)$, $x \in [0, K)$, due to the monotonicity. Therefore, all of the transmit power is allocated to subchannel K, i.e., $p_K^* = K$. In this case, we have

$$\hat{\pi}_K(p_K^*) \geq \rho_{K-1}(K - p_K^*). \tag{7.74}$$

Thus, given any $x \in (0, K]$, it holds that $p_K^* = K \geq x$.

Case 3) When $\rho_{K-1}(K) < \hat{\pi}_K(0)$ and $\rho_{K-1}(0) > \hat{\pi}_K(K)$, the optimal power allocated to subchannel K satisfies

$$\hat{\pi}_K(p_K^*) = \rho_{K-1}(K - p_K^*), \tag{7.75}$$

and it is in the range $p_K^* \in (0, K)$.

Combining the three cases above, if and only if $\hat{\pi}_K(p_K^*) \leq \rho_{K-1}(K - p_K^*)$ as in case 1 and 3, there exists $x \in (0, K]$ such that $p_K^* < x$. Since both of $\hat{\pi}_K(x)$ and $\rho_{K-1}(x)$ are PCMD functions, we have

$$\hat{\pi}_K(x) < \hat{\pi}_K(p_K^*), \tag{7.76}$$

$$\rho_{K-1}(K - p_K^*) < \rho_{K-1}(K - x). \tag{7.77}$$

Thus, the following equivalence holds,

$$p_K^* < x \;\Leftrightarrow\; \hat{\pi}_K(x) < \hat{\pi}_K(p_K^*) \leq \rho_{K-1}(K - p_K^*) < \rho_{K-1}(K - x) \tag{7.78}$$

Therefore, the CDF of the optimal power allocated to subchannel K is calculated as

$$\begin{aligned} F_{p_K^*}(x) &= \Pr\{\hat{\pi}_K(x) < \rho_{K-1}(K - x)\} \\ &= \int_0^\infty \frac{\partial G_{\rho_{K-1}}(y\,|K - x)}{\partial y} F_{\hat{\pi}_K}(y\,|x)\, dy. \end{aligned} \tag{7.79}$$

In particular, the probability that subchannel K has no allocated power as in case 1 is

$$\begin{aligned} \Pr\{p_K^* = 0\} &= \Pr\{\rho_{K-1}(K) \geq \hat{\pi}_K(0)\} \\ &= \int_0^\infty \frac{\partial G_{\rho_{K-1}}(y\,|K)}{\partial y} F_{\hat{\pi}_K}(y\,|0)\, dy. \end{aligned} \tag{7.80}$$

It is a special case of (7.79) when $x \to 0$, i.e.,

$$\Pr\{p_K^* = 0\} = \lim_{x \to 0} F_{p_K^*}(x). \tag{7.81}$$

The probability that subchannel K obtains the full transmit power K as in case 2 is

$$\begin{aligned} \Pr\{p_K^* = K\} &= \Pr\{\rho_{K-1}(0) \leq \hat{\pi}_K(K)\} \\ &= \int_0^\infty \frac{\partial F_{\hat{\pi}_K}(y\,|K)}{\partial y} G_{\rho_{K-1}}(y\,|0)\, dy. \end{aligned} \tag{7.82}$$

It is the probability of the complementary event of $p_K^* < K$. This is verified as follows.

$$\begin{aligned} &\Pr\{p_K^* = K\} + \Pr\{p_K^* < K\} \\ =&\Pr\{p_K^* = K\} + F_{p_K^*}(K) \\ =&\int_0^\infty \frac{\partial F_{\hat{\pi}_K}(y\,|K)}{\partial y} G_{\rho_{K-1}}(y\,|0)\, dy + \int_0^\infty \frac{\partial G_{\rho_{K-1}}(y\,|0)}{\partial y} F_{\hat{\pi}_K}(y\,|K)\, dy \\ =&\left[G_{\rho_{K-1}}(y\,|0)\, F_{\hat{\pi}_K}(y\,|K)\right]_{y=0}^{\infty} \\ =&1 \end{aligned} \tag{7.83}$$

Table 7.2: Simulation Parameters of the MC-NOMA System

Parameter	Value
Inter-site distance	500 m
Minimum link distance	35 m
Bandwidth per subchannel (B_{sc})	200 kHz @ 2.0 GHz
Transmit power budget per subchannel	29 dBm
Transmit antenna gain	15 dBi
Path loss	128.1+37.6lg(d[km])
Standard deviation of shadow fading	8 dB
Fast fading	Rayleigh model
Noise power density	-174 dBm/Hz
Noise figure	5 dB
Frame duration	1 ms
Averaging coefficient (τ)	1000
Number of discrete power levels (M)	200 per subchannel
Maximum number of reported RSRPs (I_R)	8

7.5 Simulations and Numerical Results

In this section, the data rate performance of various DCPA schemes for MC-NOMA is evaluated by system-level simulations in Matlab [131]. Our proposed UP-based DCPA scheme is compared to the upper bound performance obtained by the optimal solution to $P7.1$ as well as the optimal DP-based DCPA scheme proposed in [13]. Then, we compute the analytical results of the upper bound and use them to estimate the ergodic user data rates obtained by simulations in practical MC-NOMA systems. The simulation parameters are configured according to a downlink multi-cell network in [136] and are listed in Table 7.2. A downlink cellular network with 37 cells is deployed in a hexagonal grid pattern. To avoid the edge effect, only the performance of the central cell is computed while the other 36 neighbor BSs act as interferers. User terminals are uniformly randomly distributed in the cell.

In Figure 7.5, we present the change of EMA data rate per user over time and compare it to the ergodic user data rate. The UP-based DCPA scheme is utilized for the MC-NOMA system with 5 users and 16 subchannels. The EMA data rate is initialized according to the expectation of the sum spectrum efficiency calculated with the mean CQIs on the K subchannels, i.e.,

$$R_u\left(t=0\right)=\frac{K\ln\left(1+\phi_{k,u}\right)}{U}, \tag{7.84}$$

where $\phi_{k,u}$ is the mean CQI of user u on subchannel k and it is identical on the K subchannels. As shown in Figure 7.5, the EMA data rates become relatively steady after about 4,000 frames. Therefore, the simulations are carried out for 25,000 frames, including 5,000 frames for initialization and 20,000 frames for computing

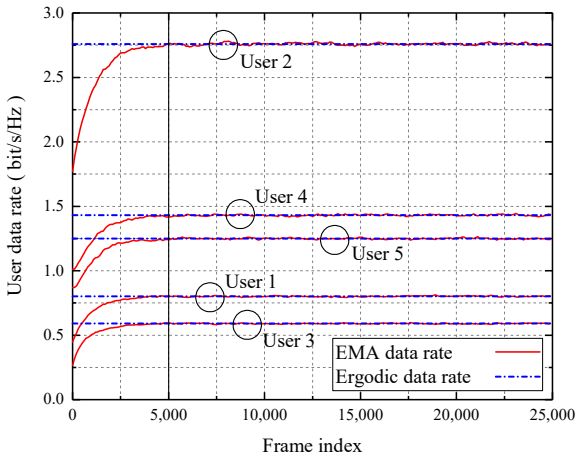


Figure 7.5: Comparison of the EMA and ergodic user data rates (UP-baesd DCPA, $U = 5$, $K = 16$, and $S = 2$).

the statistic performance.

As we proved in Theorem 7.1, the EMA data rate approximately equals to the ergodic data rate although it is time-varying. In comparison to the results in the SC-NOMA system as shown in Figure 6.3, the EMA data rates are less fluctuating over time in the MC-NOMA system. This is because the EMA data rate of a user is the sum results on all subchannels and is more steady than the link capacity of a single channel.

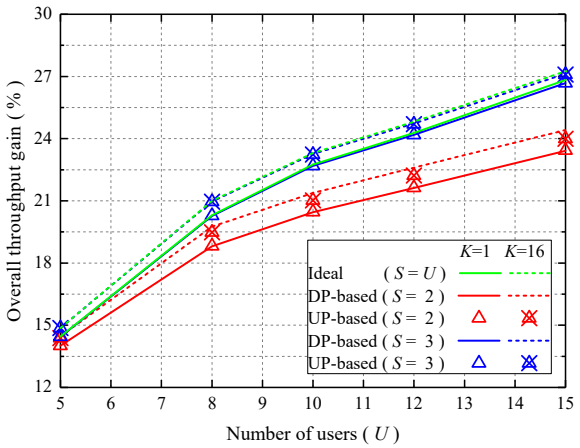
7.5.1 Simulation Results

In order to draw a fair comparison, we use the throughput obtained in the SC-OMA system as a benchmark and calculate the performance gains of various NOMA systems over it. The overall and cell-edge performance gains are present in Figure 7.6. The cell-edge throughput is defined as the sum data rate of the lowest 5% users. Compared to the optimal DP-based scheme, our proposed UP-based DCPA scheme obtains close-to-optimal performance in practical MC-NOMA systems, as shown in Figure 7.6. The performance gains increase with U and S due to the higher multi-user diversity gain. The number of candidate user sets for multiplexing on each subchannel is

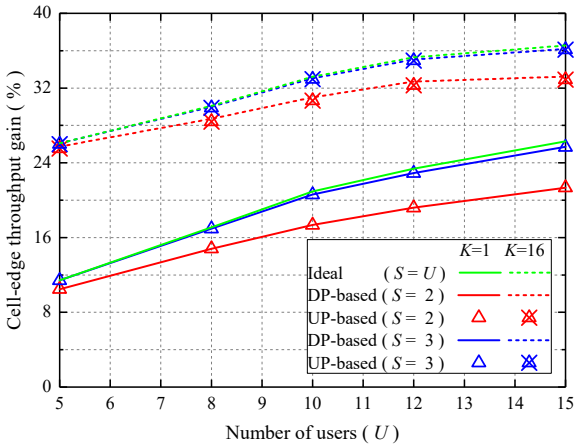
$$\sum_{i=1}^{S} C_U^i, \quad S \leq U, \tag{7.85}$$

Therefore, more candidate user sets are available and consequently enhance the multi-user diversity gain with larger U and S. Particularly, the number is maximum when $S = U$, resulting in the highest multi-user diversity gain and an upper bound performance in the ideal MC-NOMA system.

The results of SC- and MC-NOMA systems are compared by setting the number of subchannels to $K = 1$ and 16 as shown in Figure 7.6. The MC-NOMA systems



(a) Overall throughput gains



(b) Cell-edge throughput gains

Figure 7.6: Performance gains of NOMA systems over the SC-OMA system under different number of users.

outperform the SC-NOMA counterpart due to the multi-channel diversity gain brought by dynamic inter-channel power allocation. This performance improvement is more significant in the aspect of the cell-edge throughput than the overall throughput. According to (7.21), if a user has a very high CQI, i.e., $\Phi_{k,u} \gg 1$, the derivative function of its scheduling factor is approximately $\pi_{k,u}(x) \approx 1/R_u x$, which is identical on all subchannels. However, it is $\pi_{k,u}(x) \approx \Phi_{k,u}/R_u$ for the users with very low CQIs. Therefore, the time-varying subchannels have a significant effect on the fluctuating scheduling factors of the low-CQI users rather than the high-CQI ones. Thus, the cell-edge users benefit more from the multi-channel diversity gain.

To further investigate the multi-channel diversity gain in MC-NOMA systems, we present the cell-edge throughput gains with different numbers of subchannels in Figure. 7.7. Although it increases with K due to the enlarged multi-channel diversity gain, the additional improvement is limited when $K \geq 8$. Therefore, it

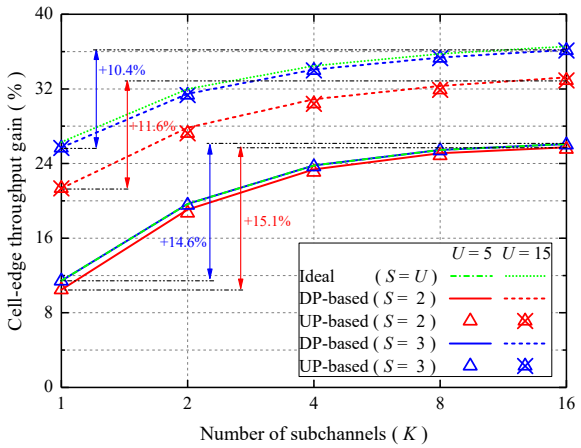


Figure 7.7: Cell-edge throughput gains of NOMA systems over the SC-OMA system with different number of subchannels.

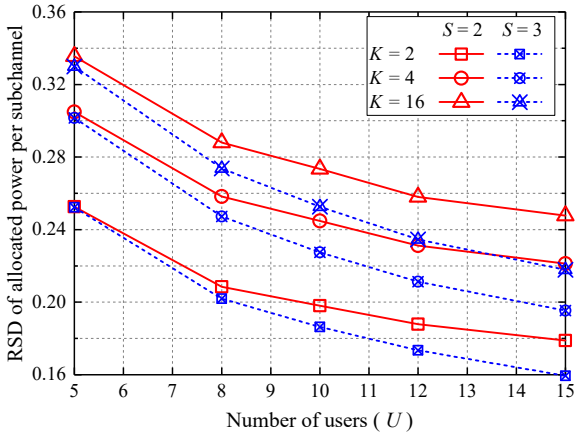


Figure 7.8: Relative standard deviation of the allocated power per subchannel in MC-NOMA systems (UP-based DCPA).

is reasonable to use a moderate number of subchannels for inter-channel power allocation in practice while considering the additional computational complexity for DCPA. Moreover, we calculate the differences between the performance gains in the MC-NOMA ($K = 16$) and SC-NOMA ($K = 1$) systems, as marked in the Figure 7.7. The multi-channel diversity gain is larger when U and S are smaller, e.g., +15.1% when $U = 5$ and $S = 2$.

To explain this, we present the relative standard deviation (RSD) of the allocated power per subchannel in Figure 7.8. As K increases in MC-NOMA system, the RSD raises due to the enlarged multi-channel diversity. However, it decreases as U and S increase. This is because a large multi-user diversity improves the stability of each subchannel. Specifically, when there are more candidate user sets in the MC-NOMA system, the probability that all of them have poor channel conditions on a subchannel is smaller. However, this results in less fluctuating link capacities on the multiple subchannels and reduces the multi-channel diversity. Therefore,

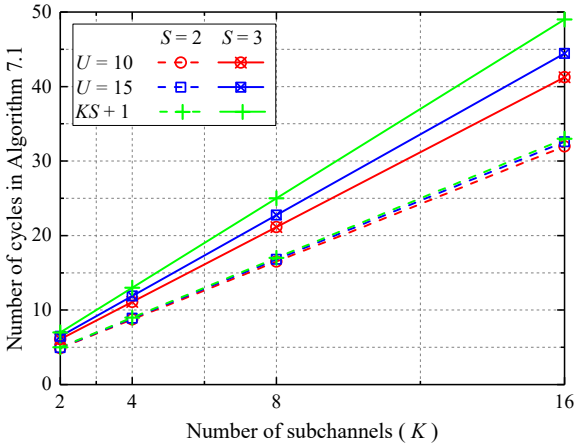


Figure 7.9: Computational complexity of Algorithm 7.1.

inter-channel power allocation improves the cell-edge throughput more significantly for the scenario where the multi-user diversity is lower, as shown in Figure 7.7.

We compute the average number of iterations in Algorithm 7.1, as shown in Figure 7.9. It is linear in the number of subchannels and increases with the number of users. Consistent with our analysis in Section 7.3, it is bounded by $(KS+1)$ such that the computational complexity of Algorithm 7.1 is $O\left(K^2S\right)$. In addition, the computational complexity of the PSU scheme is verified to be linear in the number of users in Section 6.5. Therefore, it is $O\left(KU\right)$ for the user preselection over the K subchannels. Thus, our proposed UP-based DCPA scheme costs extremely low computational complexity, i.e., $O\left(KU\right)$ and $O\left(K^2S\right)$ for the two stages, respectively. In contrast, the DP-based scheme costs much higher complexity that is $O(USK^3M^2)$ [13].

7.5.2 Data Rate Estimation

In this part, we utilize the analytical results of the upper bound performance to estimate the user data rates obtained by the UP-based DCPA scheme in practical MC-NOMA systems. The relative differences of the estimation are computed and their CDFs are presented in Figure 7.10, in which the number of subchannels is set to 16. The relative differences of the estimated user data rates increase with the number of users and most of the estimated user data rates are larger than the actual ones because of the superiority of the upper bound performance as we analyzed in Section 7.3. In addition, the throughput performance is higher in the practical MC-NOMA systems and close to the upper bound when $S = 3$ than $S = 2$. Therefore, the estimation results are more accurate in the former case.

In Figure 7.11, we present the relative differences of the estimated user data rates under different numbers of subchannels in the MC-NOMA systems. The number of users is set to $U = 15$. Similar to the results in Figure 7.10, most of the estimated user data rates have slightly positive biases that are larger when $S = 2$. While using

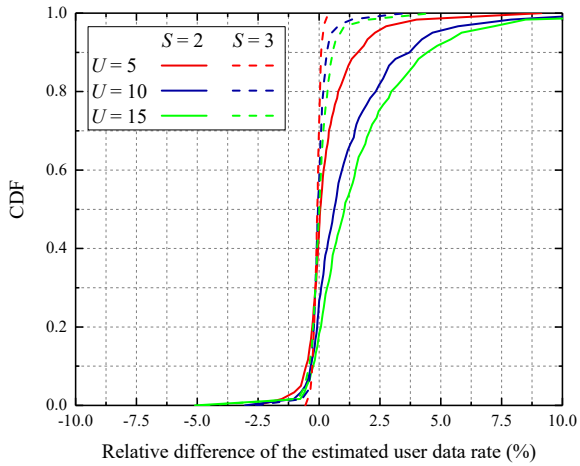


Figure 7.10: Relative differences of the estimated user data rates under different numbers of users ($K = 16$).

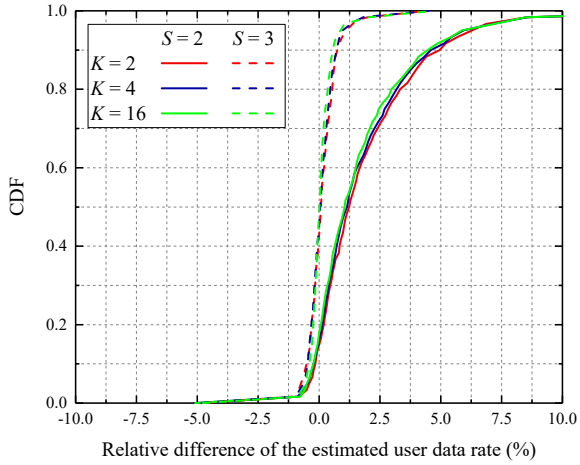


Figure 7.11: Relative differences of the estimated user data rates with different numbers of subchannels ($U = 15$).

different numbers of subchannels in the MC-NOMA system, the estimation errors are very similar. This indicates that the data rate estimation is insensitive to the change in the number of subchannels.

In Figure 7.12, we present the relative difference of the estimated overall throughput. The estimation results are very accurate, e.g., their relative differences are lower than 2% when $S = 2$ and lower than 0.2% when $S = 3$. The estimation error is mainly caused by the gap between the multi-user diversity gains in the ideal and practical MC-NOMA systems as we have explained. When there are fewer subchannels in the MC-NOMA system, the effect of inter-channel power allocation is limited. Under this condition, the multi-user diversity plays a more important role in keeping high transmission stability and efficiency on each subchannel. Therefore, the gap between the multi-user diversity gains in the ideal and practical systems

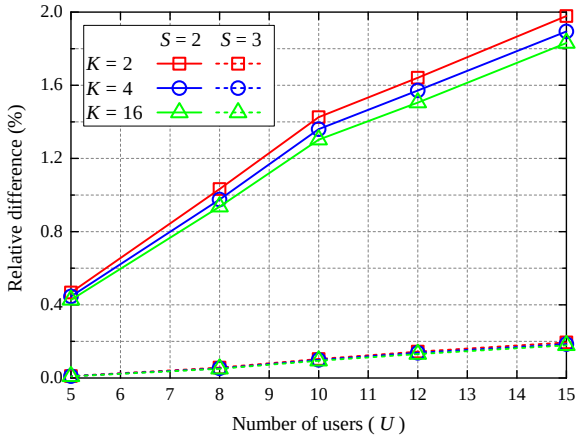


Figure 7.12: Relative difference of the estimated overall throughput with different system configurations.

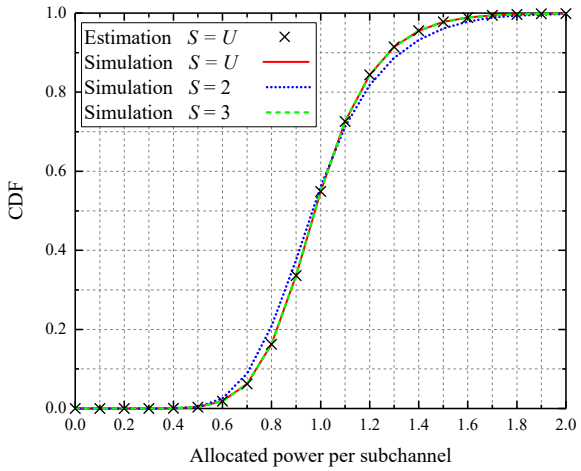


Figure 7.13: The CDF of the optimal allocated power per subchannel in the MC-NOMA system ($U = 15$, $K = 16$).

is larger as K decreases, resulting in very slightly enlarged estimation errors when there are fewer subchannels, as shown in Figure 7.12.

We compute the CDF of the optimal allocated power per subchannel in the MC-NOMA system, as shown in Figure 7.13. While $S = U$, our analytical solution derived in (7.79) is identical with the corresponding simulation results. As we explained in Section 7.5.1, the multi-user diversity increases with the number of multiplexed users per subchannel, which reduces the multi-channel diversity. Therefore, the variation of the allocated power per subchannel is larger when $S = 2$ than that in the cases of $S = 3$ and $S = U$.

7.6 Summary

In this chapter, we studied the DCPA problem for the MC-NOMA system. We first used the basic decomposition method to solve the optimal DCPA for the ideal MC-NOMA system. We designed a low-complexity algorithm for inter-channel power allocation. Its performance serves as an upper bound for the practical MC-NOMA systems and has been analyzed based on the stochastic channel models.

In order to apply this DCPA scheme to practical MC-NOMA systems, we proposed a UP-based DCPA scheme to limit the number of candidate users for multiplexing on each subchannel. Simulation results indicate that our proposed UP-based DCPA scheme achieves close-to-optimal performance with extremely low computational complexity. While using the analytical results of the upper bound performance for data rate estimation in the 2-user and 3-user practical MC-NOMA systems, the estimation results are very accurate and insensitive to the change in the number of subchannels.

The performance comparison among different system configurations and scenarios reveals that the multi-user diversity gain increases with both of the user density and the number of multiplexed users per subchannel. Moreover, dynamic inter-channel power allocation improves the throughput for cell-edge users, especially in the scenario where the multi-user diversity is low. In addition, it is necessary to consider the tradeoff between the multi-channel diversity gain and power allocation complexity in practice and implement dynamic inter-channel power allocation with an appropriate number of subchannels in MC-NOMA systems.

8 Conclusions

In the continuous evolution process of mobile networks, radio access technology always plays a significant role in the efficient exploitation of limited radio resources. Due to the fluctuating and fading nature of wireless channels, dynamic resource allocation is necessary for providing reliable and high-speed data links for multi-user access and consequently attracts a lot of research interests. Therefore, in this thesis, we were motivated to study the dynamic resource allocation problems in the OMA systems and the NOMA systems for future 5G networks. The main goals of this thesis are to develop accurate analytical solutions to the performance of dynamic resource allocation schemes in OMA systems, such as MSR, MMR, and PFS, and to address the DCPA problems for practical SC- and MC-NOMA systems, including designing low-complexity DCPA schemes and their performance analysis. The design and analytical works in this thesis provide methodology and guidelines for the optimization and analysis of future radio access networks. In this final chapter, we summarize the main results of this thesis in Section 8.1, followed by the discussion on the potential directions for the future work in Section 8.2.

8.1 Summary of Main Results

To meet the requirement of high-bitrate mobile services, the wireless cellular networks are trending towards densely deployed and reuse the limited spectrum bandwidth for improving transmission efficiency. This makes the inter-cell interference an inevitable factor that influences the transmission performance in cellular networks. Therefore, the research on the performance of dynamic resource allocation schemes under interference-limited channels is significant for optimizing system operations and resource management. As an important factor and prerequisite for the performance analysis, we first addressed the stochastic channel modeling problem for the multi-cell networks. We derived the probability distribution of the instantaneous user SINR under Rayleigh fading channels. Moreover, its upper and lower bounds were developed and extended into a weighted sum SINR model based on partial CSI. With the stochastic channel models, we analyzed the performance of three classic scheduling schemes in OMA systems, namely, MSR, MMR, and PFS, and use the analytical results for user data rate estimation. In comparison to the existing works, our estimation results have been verified to be more accurate and robust in the multi-cell networks.

The accuracy of the data rate estimation is influenced by several factors. Firstly, various dynamic resource allocation schemes depend on different system parameters for their scheduling. Some of them rely much on the instantaneous channel qualities and consequently have a higher requirement on the accuracy of the channel models for their data rate estimation, such as the MSR scheduling scheme. In contrast,

the MMR scheduler uses no CSI and is much less sensitive to the deviation of the channel models. In addition, the results of user data rate estimation in the urban and suburban environments have been compared. In the suburban area, the inter-cell interference has a larger effect on the user signals due to a smaller path loss exponent. Moreover, the average channel quality in the suburb is lower than that in the urban area, leading to higher data rate estimation errors because of the low-SINR effect. However, the data rate estimation based on our analytical results effectively overcomes these shortcomings and achieves significantly improved accuracy even with limited CSI feedback.

The estimated user data rates can be utilized as the performance prediction for potential system operations, e.g., assisting user association in multi-cell networks. In HCNs, various cells have different sizes and are deployment unevenly. Therefore, user association is an important issue for balancing the traffic load in the network and improving transmission performance. We applied our data rate estimation results to designing online user handover schemes with three different objectives, i.e., MSR, MMR, and PF. It has been verified by simulation results that our proposed schemes can effectively balance the traffic load in HCNs and outperform the conventional Max-BRP scheme in terms of both overall and cell-edge throughput. The MSR-based scheme obtains the maximum aggregated throughput while the MMR-based scheme significantly improves user fairness. A good balance between them is achieved by the PF-based scheme. Besides, the simulation results indicate that the performance of HCNs depends much on the distributions of base stations and user terminals. Therefore, targeted deployment of various types of cells and the dynamic control of user traffic load are necessary for the performance optimization in HCNs.

Then, we extended our performance analysis of dynamic resource allocation to the case of bursty on-off traffic flows that are used for modeling the wireless streaming services at the session level. Under this condition, the active user set is dynamically changing due to the random on-off processes of the user sessions. We proposed a semi-static approximation method for long-term performance analysis and derived the ergodic user data rates under the RR, MMR, and MSR scheduling schemes. We also designed a hybrid approximation method to estimate the user data rates under PFS. It combines the GA method and our MIA-based results under saturated traffic flows. The simulation results verified the high accuracy of the data rate estimation based on our analytical results and revealed that the scheduling behavior and performance were influenced by different traffic loads. Under a high traffic load, the system status is nearly saturated. Thus, the scheduling results of different resource allocation schemes under saturated traffic flows can be utilized to approximate their performance. In contrast, there are fewer active users to be scheduled as the traffic load decreases. In particular, when the traffic load is very low, all of the schedulers obtains the same results as RR since it is rare that multiple users are active and scheduled in parallel.

As a promising candidate radio access technology for the upcoming 5G networks, NOMA introduces a new dimension of user multiplexing, namely, in the power domain. However, it also brings challenges to the design of dynamic resource allocation schemes. Apart from the channel allocation problem as in the OMA

systems, the problem of inter-user power allocation is necessary to be addressed. Therefore, we focused on the DCPA problems for NOMA systems in the second part of our thesis. Consistent with the target of NOMA, i.e., improving both transmission efficiency and user fairness, PF was adopted as the optimization objective. The contributions of our work on NOMA include two main aspects: designing low-complexity DCPA schemes for practical applications and developing the analytical models for the performance analysis and user data rate estimation in NOMA systems.

We first addressed the DCPA problems for the SC-NOMA system. To simplify the problem, we assumed an ideal SC-NOMA system model in which the capability of SIC decoding was not limited. In contrast to the practical NOMA systems, the limitation on the number of multiplexed users was relaxed, resulting in an upper bound of the transmission performance. We derived the optimal solution to the power allocation problem in a closed form and designed a low-complexity DCPA scheme based on it for the ideal SC-NOMA system. In the practical SC-NONA system, the DCPA problem is more difficult to solve due to the additional SIC limitation. It has been decoupled into two stages, namely, the power allocation for each candidate user set and the optimal user set selection. The former problem has been solved by our optimal power allocation for the ideal SC-NOMA. In order to reduce the computational complexity for USS in the second stage, we proposed an optimal TSU scheme and a suboptimal PSU scheme. The performance of our proposed schemes as well as the ones in the existing works has been evaluated and compared via simulations. The TSU scheme obtains the optimal performance while the performance of PSU is extremely close to the optimal one. Moreover, in comparison to the other existing schemes, our PSU scheme significantly reduces the USS complexity, which is nearly linear in the number of users.

The SC-NOMA system is verified to be superior to the OMA one in terms of both overall and cell-edge throughput. Its performance increases with the number of users that are allowed to be multiplexed by NOMA as well as the total user amount. This is attributed to the higher multi-user diversity gain brought by the opportunistic scheduling. The simulation results indicate that the performance of the 3-user SC-NOMA system is close to the upper bound obtained in the ideal SC-NOMA system. Therefore, more users multiplexed by NOMA provide less additional benefit, considering the extra cost and signaling overhead for the high-order SIC decoding.

In addition, we derived the ergodic user data rates based on the ideal SC-NOMA assumption and our stochastic channel models developed in the first part. We utilized this analytical result of the upper bound performance for data rate estimation in the practical SC-NOMA systems. The simulation results verified that our estimation results were very accurate even with partial CSI feedback. However, imperfect CSI causes higher estimation errors inevitably thus needs to be avoided. Besides, the 3-user SC-NOMA system obtains higher throughput performance than the 2-user one, which is closer to the upper bound. Hence, the results of our data rate estimation in the 3-user SC-NOMA system are relatively more accurate.

Finally, we extended our research on the DCPA problems for the SC-NOMA systems

to the MC-NOMA case. The DCPA problem is more complex in the MC-NOMA system because of the joint design of intra- and inter-channel power allocation. Again, we addressed the DCPA problems for the ideal MC-NOMA system at first. The number of multiplexed users per subchannel was assumed to be unlimited. Under this condition, we utilized the basic decomposition method to solve the DCPA problem in two stages, namely, intra-channel power allocation and inter-channel power allocation. We designed a low-complexity water-filling algorithm for the optimal inter-channel power allocation in the second stage. Based on this optimal DCPA scheme, we further proposed a UP-based scheme for the practical MC-NOMA systems with a limited number of multiplexed users per subchannel. It obtains close-to-optimal performance with extremely low computational complexity.

Due to the time-varying channel gain on each subchannel, the multi-channel diversity gain is achievable in the MC-NOMA system. It effectively improves the data rates of the low-SINR users and consequently enhances the cell-edge throughput as the number of subchannels increases. On the other hand, the multi-user diversity gain is larger when there are more candidate users for multiplexing. This increases the stability of each subchannel and consequently reduces the multi-channel diversity. Thus, inter-channel power allocation is more beneficial in the cases where the multi-user diversity is low. In the scenario with a high user density, the inter-channel power allocation can be implemented with fewer subchannels for the sake of reducing its computational complexity. Based on the optimal solution to the DCPA problem for the ideal MC-NOMA system, we analyzed the upper bound performance and applied the analytical results to the data rate estimation in the practical MC-NOMA system. By simulations, the estimation accuracy has been verified to be very good and insensitive to the change in the number of subchannels.

8.2 Directions for Future Work

The research work in this thesis can be potentially extended in a number of directions. Firstly, there are some unexplored fields that are related to our data rate estimation and its applications. Our stochastic channel models built for data rate analysis is based on the Rayleigh fast fading channel model, which is suitable for the non-line-of-sight signal propagation in the heavy build-up environments. If there exists a dominant line-of-sight received signal, the Rician or Nakagami models can be used for modeling the fast fading channels. The stochastic channel modeling based on them is significant for both theoretical performance analysis and practical data rate estimations, especially in short-range transmissions and indoor environments. Under this channel condition, the instantaneous channel gain of the received useful signal is less fluctuating than that in the case of Rayleigh fading. Therefore, the time variation of the channel quality is caused mainly by the interference signals that have longer propagation distances and thus normally undergo Rayleigh fading. On the other hand, a further study on the dynamic resource allocation schemes and their performance under different types of traffic flows is desired. In a multi-service wireless network, users with different types of mobile services can be classified and

scheduled in groups. The analytical data rate performance can be applied to resource management for improving the quality of services in various groups.

To implement the DCPA schemes in practice, some technical issues are necessary to be taken into account for their design and performance analysis. The multi-user detection with SIC relies on the accurate and timely reported channel state information. In this thesis, the SIC receivers have been assumed to be able to decode the multiplexed user signals successfully. However, the detection error and feedback delay of CSI may lead to decoding errors. Different from the OMA system, the error in one stage of the SIC decoding process can influence the subsequent decoding stages. Therefore, it is necessary to design the DCPA schemes with the avoidance of error propagation in SIC receivers. In addition, the optimization of superposition coding is currently an open topic under study. Various modulation and coding schemes determine the obtainable link capacity in NOMA systems. Therefore, the DCPA scheme based on realistic MCSs is a potential direction to extend our work. On the other hand, the evolution from legacy wireless networks to 5G is required to be gradual and backward compatible. Thus, the NOMA systems should be able to provide services to both NOMA and OMA devices. In particular, the multiplexed user with the lowest CQI can use the legacy detection techniques since it needs no SIC. Hence, the coexistence of OMA and NOMA devices is necessary to be taken into account in the design of DCPA schemes and its impact on the system performance needs to be investigated.

Apart from NOMA, several other advanced technologies have been proposed for future 5G networks, such as mmWave communications and massive MIMO. In the mmWave communication systems, the path loss factor is large since spectrum bands with extremely high frequencies are exploited. Under this condition, the difference of the channel qualities between the near and far users is increased, resulting in a high multi-user diversity gain in NOMA systems. Therefore, NOMA is expected to enhance the transmission performance more significantly for mmWave communications, which needs to be further verified in future research. In addition, the massive MIMO provides an additional degree of freedom to the NOMA system and can further improve the link capacity. However, it is necessary to address the dynamic precoding and power allocation problems jointly. This brings increasing challenges in the design of DCPA schemes but is an interesting and potential direction in future work.

Acronyms

BPR Biased-Received-Power

BS Base Station

CD-NOMA Code-Domain NOMA

CDF Cumulative Density Function

CDMA Code Division Multiple Access

CP Cumulative Power

CQI Channel Quality Indicator

CSI Channel State Information

DCPA Dynamic Channel and Power Allocation

DP Dynamic Programming

EMA Exponential-Moving-Average

FDMA Frequency Division Multiple Access

FPA Fixed Power Allocation

FSPA Full Searching-Based Power Allocation

FTPA Fractional Transmit Power Allocation

FUSC Full User Set Comparison

GA Gaussian Approximation

HA Hybrid Approximation

HCN Heterogeneous Cellular Network

i.i.d. Independent and identically distributed

IaN Interference as Noise

IRSRP Interference Reference Signal Received Power

LDS Low-Density Spreading

MC-NOMA Multi-Channel NOMA

MCS Modulation and Coding Scheme

MIA Multi-Interference Analysis

MIMO Multiple-Input-Multiple-Output

MMR Max-Min Rate

MSR Max-Sum Rate

NOMA Non-Orthogonal Multiple Access

OFDMA Orthogonal Frequency Division Multiple Access

OMA Orthogonal Multiple Access

PD-NOMA Power-Domain NOMA

PDF Probability Density Function

PF Proportional Fairness

PFS Proportional Fair Scheduling

PPP Poisson Point Process

PSU Preselection-Based USS

QoS Quality of Service

RA Random Allocation

RAT Radio Access Technology

RB Resource Block

RR Round-Robin

RS Reference Signal

RSD Relative Standard Deviation

RSRP Reference Signal Received Power

RSSI Received Signal Strength Indicator

SC-NOMA Single-Channel NOMA

SIC Successive Interference Cancellation

SINR Signal-to-Interference-plus-Noise Ratio

SNR Signal-to-Noise Ratio

TSU Tree-Searching-Based USS

TTPA Tree-search based Transmit Power Allocation

UP User-Preselection

USS User Set Selection

WSR Weighted-Sum Rate

List of Figures

List of Tables

Bibliography

[1] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[2] A. Asadi and V. Mancuso, "A survey on opportunistic scheduling in wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1671–1688, 2013.

[3] O. Østerbø, "Scheduling and capacity estimation in LTE," in *Proceedings of 2011 23rd International Teletraffic Congress (ITC)*, 2011, pp. 63–70.

[4] N. Bui, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," in *Proceedings of 20th European Wireless Conference*, 2014, pp. 1–6.

[5] F. Naghibi and J. Gross, "How bad is interference in IEEE 802.16e system?" in *Proceedings of 2010 European Wireless Conference*, 2010, pp. 865–872.

[6] D. Parruca, M. Grysla, S. Gortzen, and J. Gross, "Analytical model of proportional fair scheduling in interference-limited OFDMA/LTE networks," in *Proceedings of 2013 IEEE Vehicular Technology Conference (VTC Fall)*, 2013, pp. 1–7.

[7] Y.-D. Yao and A. U. Sheikh, "Investigations into cochannel interference in microcellular mobile radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 2, pp. 114–123, 1992.

[8] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017.

[9] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. E98.B, no. 3, pp. 403–414, 2015.

[10] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," in *Proceedings of 2012 International Symposium on Wireless Communication Systems (ISWCS)*, 2012, pp. 476–480.

[11] A. Li, A. Harada, and H. Kayama, "A novel low computational complexity power allocation method for non-orthogonal multiple access systems," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E97-A, no. 1, pp. 57–67, 2014.

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

[13] 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 Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8580–8594, 2016.

[14] E. Okamoto, "An improved proportional fair scheduling in downlink non-orthogonal multiple access system," in *Proceedings of 82nd IEEE Vehicular Technology Conference (VTC Fall)*, 2015, pp. 1–5.

[15] T. Seyama, T. Dateki, and H. Seki, "Efficient selection of user sets for downlink non-orthogonal multiple access," in *Proceedings of 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2015, pp. 1241–1245.

[16] F. Liu, J. Riihijärvi, and M. Petrova, "Robust data rate estimation with stochastic SINR modeling in multi-interference OFDMA networks," in *Proceedings of 12th Annual International Conference on Sensing, Communication, and Networking (SECON)*, 2015, pp. 211–219.

[17] F. Liu, P. Mähönen, and M. Petrova, "A handover scheme towards downlink traffic load balance in heterogeneous cellular networks," in *Proceedings of International Conference on Communications (ICC)*, 2014, pp. 4875–4880.

[18] F. Liu and M. Petrova, "Traffic load balancing based on user data rate estimation in heterogeneous cellular networks," in *Proceedings of 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2014, pp. 1514–1519.

[19] F. Liu, J. Riihijärvi, and M. Petrova, "Analysis of proportional fair scheduling under bursty on-off traffic," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1175–1178, 2017.

[20] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *Proceedings of 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2015, pp. 1127–1131.

[21] ——, "Proportional fairness-based power allocation and user set selection for downlink NOMA systems," in *Proceedings of International Conference on Communications (ICC)*, 2016, pp. 1–6.

[22] F. Liu and M. Petrova, "Proportional fair scheduling for downlink single-carrier NOMA systems," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2017, pp. 1–7.

[23] ——, "Performance of proportional fair scheduling for downlink PD-NOMA networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 7027–7039, 2018.

[24] ——, "Dynamic power and channel allocation for downlink multi-channel NOMA systems," in *Proceedings of 29th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2018, pp. 1–7.

[25] ——, "Dynamic power allocation for downlink multi-carrier NOMA systems," *IEEE Communications Letters*, vol. 22, no. 9, pp. 1930–1933, 2018.

[26] V. K. Garg, *Wireless Communications and Networking.* Morgan Kaufmann, San Francisco, 2007.

[27] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proceedings of 2000 IEEE Vehicular Technology Conference (VTC Spring)*, 2000, pp. 1854–1858.

[28] F. P. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.

[29] F. Liu, W. Xiang, Y. Zhang, K. Zheng, and H. Zhao, "A novel QoE-based carrier scheduling scheme in LTE-advanced networks with multi-service," in *Proceedings of 2012 IEEE Vehicular Technology Conference (VTC Fall)*, 2012, pp. 1–5.

[30] P. Brooks and B. Hestnes, "User measures of quality of experience: Why being objective and quantitative is important," *IEEE Network*, vol. 24, no. 2, pp. 8–13, 2016.

[31] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, New York, 2008.

[32] J. Fan, Q. Yin, G. Y. Li, B. Peng, and X. Zhu, "MCS selection for throughput improvement in downlink LTE systems," in *Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, 2011, pp. 1–5.

[33] H. A. Suraweera, J. T. Y. Ho, T. Sivahumaran, and J. Armstrong, "An approximated Gaussian analysis and results on the capacity distribution for MIMO-OFDM," in *Proceedings of 16th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2005, pp. 211–215.

[34] P. J. Smith and M. Shafi, "On a Gaussian approximation to the capacity of wireless MIMO systems," in *Proceedings of International Conference on Communications (ICC)*, 2002, pp. 406–410.

[35] K. Zheng, F. Liu, L. Lei, C. Lin, and Y. Jiang, "Stochastic performance analysis of a wireless finite-state Markov channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 782–793, 2013.

[36] E. Liu and K. K. Leung, "Proportional fair scheduling: Analytical insight under Rayleigh fading environment," in *Proceedings of Wireless Communications and Networking Conference (WCNC)*, 2008, pp. 1883–1888.

[37] J. Leinonen, J. Hämäläinen, and M. Juntti, "Performance analysis of downlink OFDMA resource allocation with limited feedback," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2927–2937, 2009.

[38] M. Torabi, D. Haccoun, and W. Ajib, "Performance analysis of scheduling schemes for rate-adaptive MIMO OSFBC-OFDM systems," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2363–2379, 2009.

[39] S. N. Donthi and N. B. Mehta, "An accurate model for EESM and its application to analysis of CQI feedback schemes and scheduling in LTE," *IEEE Transactions on Wireless Communications*, vol. 10, no. 10, pp. 3436–3448, 2011.

[40] R. K. Almatarneh, M. H. Ahmed, and O. A. Dobre, "Performance analysis of proportional fair scheduling in OFDMA wireless systems," in *Proceedings of 72nd IEEE Vehicular Technology Conference (VTC Fall)*, 2010, pp. 1–5.

[41] J.-G. Choi and S. Bahk, "Cell-throughput analysis of the proportional fair scheduler in the single-cell environment," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 766–778, 2007.

[42] E. Liu and K. K. Leung, "Expected throughput of the proportional fair scheduling over Rayleigh fading channels," *IEEE Communications Letters*, vol. 14, no. 6, pp. 515–517, 2010.

[43] ——, "Fair resource allocation under Rayleigh and/or Rician fading environments," in *Proceedings of 19th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2008, pp. 1–5.

[44] J. Francis and N. B. Mehtaa, "Characterizing the impact of feedback delays on wideband rate adaptation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 960–971, 2014.

[45] K. Zheng, F. Liu, W. Xiang, and X. Xin, "Dynamic downlink aggregation carrier scheduling scheme for wireless networks," *IET Communications*, vol. 8, no. 1, pp. 114–123, 2014.

[46] L. Lei, C. Lin, J. Cai, and X. Shen, "Flow-level performance of opportunistic OFDM-TDMA and OFDMA networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5461–5472, 2008.

[47] M. H. Ahmed, O. A. Dobre, and R. K. Almatarneh, "Analytical evaluation of the performance of proportional fair scheduling in OFDMA-based wireless systems," *Journal of Electrical and Computer Engineering*, vol. 2012, no. 680318, pp. 1–12, 2012.

[48] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 36–43, 2014.

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

[50] P. Marsch, I. D. Silva, O. Bulakci, M. Tesanovic, S. E. E. Ayoubi, T. Rosowski, A. Kaloxylos, and M. Boldi, "5G radio access network architecture: Design guidelines and key considerations," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 24–32, 2016.

[51] H. Lee, S. Kim, and J.-H. Lim, "Multiuser superposition transmission (MUST) for LTE-A systems," in *Proceedings of International Conference on Communications (ICC)*, 2016, pp. 1–6.

[52] TR 36.859, "Study on downlink multiuser superposition transmission (MUST) for LTE," 3GPP, Tech. Rep., 2016.

[53] P. Wang, J. Xiao, and L. Ping, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 4–11, 2006.

[54] Y. Yuan, Z. Yuan, G. Yu, C.-H. Hwang, P.-K. Liao, A. Li, and K. Takeda, "Non-orthogonal transmission technology in LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, 2016.

[55] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2016.

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

[57] R. Razavi, R. Hoshyar, M. A. Imran, and Y. Wang, "Information theoretic analysis of LDS scheme," *IEEE Communications Letters*, vol. 15, no. 8, pp. 798–800, 2011.

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

[59] M. A.-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proceedings of 11th International Symposium on Wireless Communications Systems (ISWCS)*, 2014, pp. 781–785.

[60] M. A.-Imari, M. A. Imran, and R. Tafazolli, "Low density spreading for next generation multicarrier cellular systems," in *Proceedings of International Conference on Future Communication Networks (ICFCN)*, 2012, pp. 52–57.

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

[62] H. Nikopour, E. Yi, A. Bayesteh, K. Au, M. Hawryluck, H. Baligh, and J. Ma, "SCMA for downlink multiple access of 5G wireless networks," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2014, pp. 3940–3945.

[63] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access—A novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3185–3196, 2016.

[64] J. Huang, K. Peng, C. Pan, F. Yang, and H. Jin, "Scalable video broadcasting using bit division multiplexing," *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 701–706, 2014.

[65] J. Zhang, S. Chen, X. Mu, and L. Hanzo, "Turbo multi-user detection for OFDM/SDMA systems relying on differential evolution aided iterative channel estimation," *IEEE Transactions on Communications*, vol. 60, no. 6, pp. 1621–1633, 2012.

[66] ——, "Evolutionary-algorithm-assisted joint channel estimation and Turbo multiuser detection/decoding for OFDM/SDMA," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1204–1222, 2013.

[67] F. Brannstrom, T. M. Aulin, and L. K. Rasmussen, "Iterative detectors for trellis-code multiple-access," *IEEE Transactions on Communications*, vol. 50, no. 9, pp. 1478–1485, 2002.

[68] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 938–947, 2006.

[69] K. Kusume, G. Bauch, and W. Utschick, "IDMA vs. CDMA: Analysis and comparison of two multiple access schemes," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 78–87, 2011.

[70] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proceedings of International Symposium*

on Intelligent Signal Processing and Communication Systems, 2013, pp. 770–774.

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

[72] A. Benjebbovu, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proceedings of Globecom Workshops*, 2013, pp. 66–77.

[73] N. Nonaka, Y. Kishiyama, and K. Higuchi, "Non-orthogonal multiple access using intra-beam superposition coding and SIC in base station cooperative MIMO cellular downlink," in *Proceedings of 80th Vehicular Technology Conference (VTC Fall)*, 2014, pp. 1–5.

[74] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proceedings of 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2013, pp. 611–615.

[75] Z. Q. A.-Abbasi and D. K. C. So, "Power allocation for sum rate maximization in non-orthogonal multiple access system," in *Proceedings of 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2015, pp. 1839–1843.

[76] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3226–3237, 2016.

[77] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 76–88, 2015.

[78] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, 2015.

[79] J. Mei, L. Yao, H. Long, and K. Zheng, "Joint user pairing and power allocation for downlink non-orthogonal multiple access systems," in *Proceedings of International Conference on Communications (ICC)*, 2016, pp. 1–6.

[80] S. N. Datta and S. Kalyanasundaram, "Optimal power allocation and user selection in non-orthogonal multiple access systems," in *Proceedings of Wireless Communications and Networking Conference (WCNC)*, 2016, pp. 1–6.

[81] 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.

[82] P. Parida and S. S. Das, "Power allocation in OFDM based NOMA systems: A DC programming approach," in *Proceedings of Globecom Workshops*, 2014, pp. 1026–1031.

[83] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for MC-NOMA systems," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[84] ——, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1077–1091, 2017.

[85] M.-R. Hojeij, C. A. Nour, J. Farah, and C. Douillard, "Waterfilling-based proportional fairness scheduler for downlink non-orthogonal multiple access," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 230–233, 2017.

[86] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5825–5837, 2017.

[87] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.

[88] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2744–2757, 2017.

[89] J. Zhu, J. Wang, Y. Huang, S. He, and X. You, "Multichannel resource allocation for downlink non-orthogonal multiple access systems," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2017, pp. 1–6.

[90] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3722–3732, 2016.

[91] Z. Wei, D. W. K. Ng, J. Yuan, and H.-M. Wang, "Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3944–3961, 2017.

[92] P. Xu and K. Cumanan, "Optimal power allocation scheme for NOMA with adaptive rates and alpha-fairness," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2017, pp. 1–6.

[93] ——, "Optimal power allocation scheme for non-orthogonal multiple access with α-fairness," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2357–2369, 2017.

[94] X. Chen, A. Benjebboui, Y. Lan, A. Li, and H. Jiang, "Evaluations of downlink non-orthogonal multiple access (NOMA) combined with SU-MIMO," in *Proceedings of 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2014, pp. 1887–1891.

[95] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Joint subchannel and power allocation for NOMA enhanced D2D communications," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 5081–5094, 2017.

[96] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 654–667, 2016.

[97] J. A. Oviedo and H. R. Sadjadpour, "A new NOMA approach for fair power allocation," in *Proceedings of Conference on Computer Communications (INFOCOM) Workshops*, 2016, pp. 1–5.

[98] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "A new evaluation criterion for non-orthogonal multiple access in 5G software defined networks," *IEEE Access*, vol. 3, pp. 1633–1639, 2015.

[99] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1501–1505, 2014.

[100] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 405–408, 2015.

[101] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, 2016.

[102] Z. Yang, Z. Ding, P. Fan, and Z. Ma, "Outage performance for dynamic power allocation in hybrid non-orthogonal multiple access systems," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1695–1698, 2016.

[103] Z. Yang, Z. Ding, P. Fan, and N. A.-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7244–7257, 2016.

[104] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, 2016.

[105] J.-B. Kim and I.-H. Lee, "Capacity analysis of cooperative relaying systems using non-orthogonal multiple access," *IEEE Communications Letters*, vol. 19, no. 11, pp. 1949–1952, 2015.

[106] J. Men, J. Ge, and C. Zhang, "Performance analysis of nonorthogonal multiple access for relaying networks over Nakagami-m fading channels," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 2, pp. 1200–1208, 2017.

[107] J. Men and J. Ge, "Performance analysis of non-orthogonal multiple access in downlink cooperative network," *IET Communications*, vol. 9, no. 18, pp. 2267–2273, 2015.

[108] P. Sedtheetorn and K. Panyim, "Accurate uplink spectral efficiency for non-orthogonal multiple access in Nakagami fading," in *Proceedings of 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2016, pp. 1–4.

[109] T. Chulajata and P. Sedtheetorn, "Theoretical analysis on spectral efficiency of non-orthogonal multiple access in Nakagami fading," in *Proceedings of International Conference on Control System, Computing and Engineering (ICCSCE)*, 2015, pp. 146–149.

[110] TS 36.331, "Evolved universal terrestrial radio access (E-UTRA); Radio resource control (RRC)," 3GPP, Tech. Rep., 2016.

[111] TS 36.133, "E-UTRA requirements for support of radio resource management," 3GPP, Tech. Rep., 2014.

[112] M. Akkouchi, "On the convolution of exponential distributions," *Journal of the Chungcheong Mathematical Society*, vol. 21, no. 4, pp. 501–509, 2008.

[113] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proceedings of 25th International Conference on Computer Communications*, 2006, pp. 1–12.

[114] K. E. Atkinson, "A survey of numerical methods for solving nonlinear integral equations," *Journal of Integral Equations and Applications*, vol. 4, no. 1, pp. 15–46, 1992.

[115] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous celluar networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[116] Unwiredlabs, "OpenCellid: Open database of cell towers," 2014, http://opencellid.org [Online accessed Oct. 2014].

[117] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54–64, 2012.

[118] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[119] Y. Bejerano and S. J. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 735–749, 2009.

[120] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3484–3495, 2012.

[121] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, 2013.

[122] Q. Ye, B. Rong, Y. Chen, M. A.-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, 2013.

[123] D. Fooladivanda, A. Al Daoud, and C. Rosenberg, "Joint channel allocation and user association for heterogeneous wireless cellular networks," in *Proceedings of 22nd Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2012, pp. 384–390.

[124] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in *Proceedings of Conference on Computer Communications(INFOCOM)*, 2008, pp. 1678–1686.

[125] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 636–647, 2005.

[126] Y. Bejerano, S. J. Han, and L. Li, "Fairness and load balancing in wireless LANs using association control," *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, pp. 560–573, 2007.

[127] H. Kim, G. D. Veciana, X. Yang, M. Venkatachalam, "Alpha-optimal user association and cell load balancing in wireless networks," in *Proceedings of Conference on Computer Communications(INFOCOM)*, 2010, pp. 1–5.

[128] H. Kim, G. D. Beciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, 2012.

[129] Q. Ye, B. Rong, Y. Chen, C. Caramanis, and J. G. Andrews, "Towards an optimal user association in heterogeneous cellular networks," in *Proceedings of Global Communications Conference (GLOBECOM)*, 2012, pp. 4143–4147.

[130] P. Hande, S. Patil, and H. G. Myung, "Distributed load-balancing in a multi-carrier wireless system," in *Proceedings of Wireless Communications and Networking Conference (WCNC)*, 2009, pp. 1–6.

[131] MathWorks, "MATLAB R2013a," 2013, https://www.mathworks.com/products/matlab.html [Online accessed Apr. 2014].

[132] S. Mukherjee, *Analytical Modeling of Heterogeneous Cellular Networks: Geometry, Coverage, and Capacity.* Cambridge University Press, New York, 2014.

[133] WiGLE, "Cell tower and WiFi access point location data," 2014, http://wigle.net [Online accessed Apr. 2014].

[134] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, and M. Arlitt, "A comparative analysis of web and peer-to-peer traffic," in *Proceedings of 17th International Conference on World Wide Web*, 2008, pp. 287–296.

[135] K. J. Devlin, *Fundamentals of Contemporary Set Theory.* Springer-Verlag, New York, 1979.

[136] TS 36.839, "Evolved universal terrestrial radio access (E-UTRA); Mobility enhancements in heterogeneous networks," 3GPP, Tech. Rep., 2015.

[137] S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley, "Notes on decomposition methods," 2017, https://web.stanford.edu/class/ee364b/lectures/decomposition_notes.pdf [Online accessed Oct. 2018].

[138] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 686–695, 2005.

List of Publications

Journal papers relevant to this thesis

[J1] **F. Liu**, J. Riihijärvi, and M. Petrova. Analysis of proportional fair scheduling under bursty on-off traffic. *IEEE Communications Letters*, 21(5):1175–1178, 2017

[J2] **F. Liu** and M. Petrova. Dynamic power allocation for downlink multi-carrier NOMA systems. *IEEE Communications Letters*, 22(9):1930–1933, 2018

[J3] **F. Liu** and M. Petrova. Performance of proportional fair scheduling for downlink PD-NOMA networks. *IEEE Transactions on Wireless Communications*, 17(10):7027–7039, 2018

[J4] **F. Liu** and M. Petrova. Performance of dynamic power and channel allocation for downlink MC-NOMA systems. *Submitted to IEEE Transactions on Wireless Communications*, 2019

Conference contributions relevant to this thesis

[C1] **F. Liu**, P. Mähönen, and M. Petrova. A handover scheme towards downlink traffic load balance in heterogeneous cellular networks. In *Proceedings of International Conference on Communications (ICC)*, pages 4875–4880, 2014

[C2] **F. Liu** and M. Petrova. Traffic load balancing based on user data rate estimation in heterogeneous cellular networks. In *Proceedings of 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pages 1514–1519, 2014

[C3] **F. Liu**, J. Riihijärvi, and M. Petrova. Robust data rate estimation with stochastic SINR modeling in multi-interference OFDMA networks. In *Proceedings of 12th Annual International Conference on Sensing, Communication, and Networking (SECON)*, pages 211–219, 2015

[C4] **F. Liu**, P. Mähönen, and M. Petrova. Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access. In *Proceedings of 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1127–1131, 2015

[C5] **F. Liu**, P. Mähönen, and M. Petrova. Proportional fairness-based power allocation and user set selection for downlink NOMA systems. In *Proceedings of International Conference on Communications (ICC)*, pages 1–6, 2016

[C6] **F. Liu** and M. Petrova. Proportional fair scheduling for downlink single-carrier NOMA systems. In *Proceedings of Global Communications Conference (GLOBECOM)*, pages 1–7, 2017

[C7] **F. Liu** and M. Petrova. Dynamic power and channel allocation for downlink multi-channel NOMA systems. In *Proceedings of 29th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pages 1–7, 2018

Curriculum Vitae

Fei Liu

Personal Information

Date of birth	10. March 1988
Place of birth	Henan, China
Nationality	Chinese

Education

09/2010-03/2013	Master of Engineering in Communication and Information Systems, Beijing University of Posts and Telecommunications (BUPT), China
09/2006-06/2010	Bachelor of Engineering in Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), China
09/2003-06/2006	Zhengzhou No.11 Middle School, Henan, China

Professional Experience

05/2013-05/2019	Research Assistant at the Institute for Networked Systems (iNETS), RWTH Aachen University, Germany
10/2009-03/2013	Student Assistant at Wireless Signal Processing and Network Laboratory (WSPN), Beijing University of Posts and Telecommunications (BUPT), China
07/2008-09/2009	Student Assistant at State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT), China