

Wireless Communications and Mobile Computing

Nonorthogonal Multiple Access for 5G and Beyond

Lead Guest Editor: Oğuz Kucur

Guest Editors: Güneş K. Kurt, Muhammad Z. Shakir, and Imran S. Ansari






Nonorthogonal Multiple Access for 5G and Beyond

Wireless Communications and Mobile Computing

Nonorthogonal Multiple Access for 5G and Beyond

Lead Guest Editor: Oğuz Kucur

Guest Editors: Güneş K. Kurt, Muhammad Z. Shakir,
and Imran S. Ansari



Copyright © 2018 Hindawi. All rights reserved.

This is a special issue published in “Wireless Communications and Mobile Computing.” All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board

- Javier Aguiar, Spain
Wessam Ajib, Canada
Muhammad Alam, China
Eva Antonino-Daviu, Spain
Shlomi Arnon, Israel
Leyre Azpilicueta, Mexico
Paolo Barsocchi, Italy
Alessandro Bazzi, Italy
Zdenek Becvar, Czech Republic
Francesco Benedetto, Italy
Olivier Berder, France
Ana M. Bernardos, Spain
Mauro Biagi, Italy
Dario Bruneo, Italy
Jun Cai, Canada
Zhipeng Cai, USA
Claudia Campolo, Italy
Gerardo Canfora, Italy
Rolando Carrasco, UK
Vicente Casares-Giner, Spain
Dajana Cassioli, Italy
Luis Castedo, Spain
Ioannis Chatzigiannakis, Greece
Lin Chen, France
Yu Chen, USA
Hui Cheng, UK
Luca Chiaraviglio, Italy
Ernestina Cianca, Italy
Riccardo Colella, Italy
Mario Collotta, Italy
Massimo Condoluci, Sweden
Bernard Cousin, France
Telmo Reis Cunha, Portugal
Igor Curcio, Finland
Laurie Cuthbert, Macau
Donatella Darsena, Italy
Pham Tien Dat, Japan
André de Almeida, Brazil
Antonio De Domenico, France
Antonio de la Oliva, Spain
Gianluca De Marco, Italy
Luca De Nardis, Italy
Alessandra De Paola, Italy
Liang Dong, USA
- Mohammed El-Hajjar, UK
Oscar Esparza, Spain
Maria Fazio, Italy
Mauro Femminella, Italy
Manuel Fernandez-Veiga, Spain
Gianluigi Ferrari, Italy
Ilario Filippini, Italy
Jesus Fontecha, Spain
Luca Foschini, Italy
A. G. Fragkiadakis, Greece
Sabrina Gaito, Italy
Óscar García, Spain
Manuel García Sánchez, Spain
L. J. García Villalba, Spain
José A. García-Naya, Spain
Miguel Garcia-Pineda, Spain
A.-J. García-Sánchez, Spain
Piedad Garrido, Spain
Vincent Gauthier, France
Carlo Giannelli, Italy
Carles Gomez, Spain
Juan A. Gomez-Pulido, Spain
Ke Guan, China
Daojing He, China
Paul Honeine, France
Sergio Ilarri, Spain
Antonio Jara, Switzerland
Xiaohong Jiang, Japan
Minho Jo, Republic of Korea
Shigeru Kashihara, Japan
Dimitrios Katsaros, Greece
Minseok Kim, Japan
Mario Kolberg, UK
Nikos Komninos, UK
Juan A. L. Riquelme, Spain
Pavlos I. Lazaridis, UK
Tuan Anh Le, UK
Xianfu Lei, China
Hoa Le-Minh, UK
Jaime Lloret, Spain
Miguel López-Benítez, UK
Martín López-Nores, Spain
Javier D. S. Lorente, Spain
Tony T. Luo, Singapore
- Maode Ma, Singapore
Imadeldin Mahgoub, USA
Pietro Manzoni, Spain
Álvaro Marco, Spain
Gustavo Marfia, Italy
Francisco J. Martinez, Spain
Davide Mattera, Italy
Michael McGuire, Canada
Nathalie Mitton, France
Klaus Moessner, UK
Antonella Molinaro, Italy
Simone Morosi, Italy
Kumudu S. Munasinghe, Australia
Enrico Natalizio, France
Keivan Navaie, UK
Thomas Newe, Ireland
Wing Kwan Ng, Australia
Tuan M. Nguyen, Vietnam
Petros Nicolitidis, Greece
Giovanni Pau, Italy
Rafael Pérez-Jiménez, Spain
Matteo Petracca, Italy
Nada Y. Philip, UK
Marco Picone, Italy
Daniele Pinchera, Italy
Giuseppe Piro, Italy
Vicent Pla, Spain
Javier Prieto, Spain
Rüdiger C. Prys, Germany
Junaid Qadir, Pakistan
Sujan Rajbhandari, UK
Rajib Rana, Australia
Luca Reggiani, Italy
Daniel G. Reina, Spain
Abusayed Saifullah, USA
Jose Santa, Spain
Stefano Savazzi, Italy
Hans Schotten, Germany
Patrick Seeling, USA
Muhammad Z. Shakir, UK
Mohammad Shojafar, Italy
Giovanni Stea, Italy
Enrique Stevens-Navarro, Mexico
Zhou Su, Japan



Luis Suarez, Russia
Ville Syrjälä, Finland
Hwee Pink Tan, Singapore
Pierre-Martin Tardif, Canada
Mauro Tortonesi, Italy

Federico Tramarin, Italy
Reza Monir Vaghefi, USA
Juan F. Valenzuela-Valdés, Spain
Aline C. Viana, France
Enrico M. Vitucci, Italy

Honggang Wang, USA
Jie Yang, USA
Sherali Zeadally, USA
Jie Zhang, UK
Meiling Zhu, UK

Contents

Nonorthogonal Multiple Access for 5G and Beyond

Oğuz Kucur , Güneş Karabulut Kurt , Muhammad Zeeshan Shakir, and Imran Shafique Ansari
Editorial (2 pages), Article ID 1907506, Volume 2018 (2018)

Multiway Physical-Layer Network Coding via Uniquely Decodable Codes

Michel Kulhandjian , Claude D'Amours, and Hovannes Kulhandjian 
Research Article (8 pages), Article ID 2034870, Volume 2018 (2018)



A Tutorial on Nonorthogonal Multiple Access for 5G and Beyond

Mahmoud Aldababsa, Mesut Toka, Selahattin Gökçeli , Güneş Karabulut Kurt, and Oğuz Kucur 
Review Article (24 pages), Article ID 9713450, Volume 2018 (2018)

Weighted Proportional Fair Scheduling for Downlink Nonorthogonal Multiple Access

Marie-Rita Hojeij , Charbel Abdel Nour , Joumana Farah , and Catherine Douillard 
Research Article (12 pages), Article ID 5642765, Volume 2018 (2018)

NOMA for Multinumerology OFDM Systems

Ayman T. Abusabah  and Huseyin Arslan 
Research Article (9 pages), Article ID 8514314, Volume 2018 (2018)


Nonuniform Code Multiple Access

Cheng Yan , Ningbo Zhang , and Guixia Kang 
Research Article (11 pages), Article ID 7603797, Volume 2018 (2018)

On the Performance of Security-Based Nonorthogonal Multiple Access in Coordinated Multipoint Networks



Yue Tian , Xianling Wang , and Zhanwei Wang 
Research Article (6 pages), Article ID 8921895, Volume 2018 (2018)

An Efficient SCMA Codebook Optimization Algorithm Based on Mutual Information Maximization

Chao Dong , Guili Gao, Kai Niu, and Jiaru Lin
Research Article (13 pages), Article ID 8910907, Volume 2018 (2018)

Editorial

Nonorthogonal Multiple Access for 5G and Beyond

Oğuz Kucur ¹, **Güneş Karabulut Kurt** ²,
Muhammad Zeeshan Shakir,³ and **Imran Shafique Ansari**⁴

¹*Gebze Technical University, Kocaeli, Turkey*

²*Istanbul Technical University, Istanbul, Turkey*

³*University of the West of Scotland, Paisley, UK*

⁴*Global College of Engineering and Technology in Partnership with UWE Bristol, Muscat, Oman*

Correspondence should be addressed to Oğuz Kucur; okucur@gtu.edu.tr

Received 3 May 2018; Accepted 3 May 2018; Published 5 July 2018

Copyright © 2018 Oğuz Kucur et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Over the past decades, rapid developments and evolving demands of wireless communications changed the selected multiple access (MA) technique in each generation. Today, increasing demands of high spectral/energy efficiency, high connectivity, and low latency of future generations such as 5G and beyond can be satisfied by Non-orthogonal Multiple Access (NOMA). Different from conventional MA techniques such as frequency division MA, time division MA, code division MA, and orthogonal frequency division MA of previous generations, which are based on orthogonal resources, the key idea in NOMA is to allocate non-orthogonal resources to serve multiple users, yielding a high spectral efficiency while allowing some degree of interference at receivers. Recently proposed NOMA techniques can be mainly categorized into two groups: power domain and code domain. In power domain NOMA, multiple users are superposed by different power levels opposite to the channel conditions such that successive interference cancellation is applied at receivers, providing a good tradeoff between system throughput and user fairness. Other NOMA approaches include low density spreading (LDS), sparse code multiple access (SCMA), multi-user shared access (MUSA), and interleave division multiple access (IDMA), and so on.

Despite having several advantages for 5G and beyond, challenges and obstacles exist in the efficient deployment of NOMA. The motivation behind this special issue has been to address such challenging issues of NOMA. Following a rigorous review process (including a second review round), 7 outstanding papers have been finally selected for inclusion in the special issue. The accepted papers cover a wide range

of research subjects in the broader area of NOMA to meet the increasing demands of 5G and beyond.

The paper entitled “A Tutorial on Non-Orthogonal Multiple Access (NOMA) for 5G and Beyond” by M. Aldababsa et al. provides a unified model for NOMA, including uplink and downlink transmissions, along with the extensions to multiple input multiple output and cooperative communication scenarios. The authors compare the performances of orthogonal multiple access (OMA) and NOMA networks through numerical examples and also provide discussions about implementation aspects and open issues.

The paper “NOMA for Multinumerology OFDM Systems” by A. T. Abusabah et al. proposes an orthogonal frequency division multiplexing (OFDM) based NOMA scheme, which utilizes the multi-numerology concept, i.e., different subcarrier spacings to reduce the constraints associated with the multi-user detection. By this scheme, which is less spectrally efficient than conventional NOMA schemes, but more spectrally efficient than OMA schemes, the authors improve both user fairness and error performance.

A new member of NOMA family, named as non-uniform code multiple access (NCMA), is introduced in “Nonuniform Code Multiple Access”, by C. Yan et al. In NCMA, different transmitted layers can be generated from different complex multi-dimensional constellations. Benefiting from the proposed codebook design, the minimum intra-partition distance can be increased. Simulation results demonstrate performance improvement against SCMA.

The paper “Weighted Proportional Fair Scheduling for Downlink Nonorthogonal Multiple Access” by M.-R. Hojeij

et al. has presented a novel weighted proportional fair scheduling scheme for NOMA. The proposed design can adapt the weights and maximize the capacity while improving the long term fairness amongst the users. The study has been supported by the set of the simulation results and comparative evaluation with the benchmarks such as OMA and classic NOMA-based proportional fair scheduler. The concept has potential to be exploited further for multi-user and multi-antenna scenarios with critical focus on the complexity of the design.

The paper, “Multiway Physical-Layer Network Coding via Uniquely Decodable Codes” by M. Kulhandjian et al. has proposed a novel but simple network coding by exploiting Uniquely Decodable (UD) Codes to allow users to uniquely recover the information bits from the noise channel. The proposed decoder has been simulated and compared with the traditional maximum likelihood (ML) decoders in terms of the bit error rate and sum rate. It has been shown that the proposed scheme utilizing UD codes achieve near-ML performance with less complex design and improve the sum rates almost 7 and 16 times compared to traditional physical layer network coding scheme.

The paper “An Efficient SCMA Codebook Optimization Algorithm Based on Mutual Information Maximization” by C. Dong et al. proposes an efficient SCMA codebook optimization algorithm to maximize mutual information between the discrete input and continuous output. Initially, SCMA signal model is described according to superposition modulation structure wherein the channel matrix is column extended that can well represent the relationship between the codebook matrix and received signal. The superposition model can well describe the relationship between the codebook matrix and received signal. Based on this superposition model, an iterative codebook optimization algorithm is proposed wherein the linear search method is applied in order to find locally optimal codebooks thereby maximizing mutual information between discrete input and continuous output. This algorithm can efficiently adapt to multi-user channels with arbitrary channel coefficients. The simulation results demonstrate that the proposed algorithm approaches Gaussian capacity upper bound in low and medium signal-to-noise ratio regimes. Moreover, the performance loss in non-additive white Gaussian noise channel is relatively small when compared with the upper bound. In addition, message passing algorithm works well with the codebook optimized with the proposed algorithm.

The paper “On the Performance of Security-Based Nonorthogonal Multiple Access in Coordinated Multipoint Networks” by Y. Tian et al. focuses on the security-based NOMA (S-NOMA) systems that aim to improve the physical layer security issues of conventional NOMA systems in the coordinated multi-point (CoMP) networks. The authors analyze the secrecy performance of S-NOMA in CoMP, i.e., the secrecy sum-rate and the secrecy outage probability, and demonstrate that the proposed S-NOMA outperforms conventional NOMA in terms of the secrecy outage probability and security-based effective sum-rate, especially when the target transmission data rate is high.

Acknowledgments

We would like to thank all the authors who submitted their excellent research articles to this special issue and all the reviewers for providing their valuable and timely feedback through the review process, which helped to improve the quality of this special issue.

*Oğuz Kucur
Güneş Karabulut Kurt
Muhammad Zeeshan Shakir
Imran Shafique Ansari*

Research Article

Multiway Physical-Layer Network Coding via Uniquely Decodable Codes

Michel Kulhandjian ¹, Claude D'Amours,¹ and Hovannes Kulhandjian ²

¹School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada K1N 6N5

²Department of Electrical and Computer Engineering, California State University, Fresno, Fresno, CA 93740, USA

Correspondence should be addressed to Hovannes Kulhandjian; hkulhandjian@csufresno.edu

Received 27 November 2017; Revised 30 March 2018; Accepted 4 April 2018; Published 28 June 2018

Academic Editor: Muhammad Z. Shakir

Copyright © 2018 Michel Kulhandjian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We focus on a multiway relay channel (MWRC) network where two or more users simultaneously exchange information with each other through the help of a relay node. We propose for the first time to apply ternary uniquely decodable (UD) code sets that we have developed to allow each user to uniquely recover the information bits from the noisy channel environment. One of the key features of the proposed scheme is that it utilizes a very simple decoding algorithm, which requires only a few logical comparisons. Simulation results in terms of bit error rate (BER) demonstrate that the performance of the proposed decoder is almost as good as the maximum-likelihood (ML) decoder. In addition to that through simulations, we show that the proposed scheme can significantly improve the sum-rate capacity, which in turn can potentially improve overall throughput, as it needs only two time slots (TSs) to exchange information compared to the conventional methods.

1. Introduction

Network coding (NC) has attracted a lot of attention in the research community due to its capability to enhance the throughput of lossless wireline networks in a multicast environment [1]. In a two-way bidirectional relay channel, NC requires three time slots (TSs) for information exchange while the conventional scheme requires four. Therefore, NC can potentially improve the channel throughput by 4/3 times in bidirectional channels.

A lot of work has been done to investigate some of the NC issues in wireless environments [2–5]. The conventional NC data forwarding schemes based on decode-and-forward (DF) relaying are not able to fully utilize the wireless channel. More recently, denoise-and-forward (DNF) relaying adopting the physical-layer network coding (PNC) proposed by Zhang et al. [6] can potentially provide further throughput improvements of the wireless channel. The PNC has several attractive features; in particular, it is relatively simple to implement compared to the conventional NC method as mentioned in [7]. Unlike the conventional two-way relay channel (TWRC), which requires four time slots, PNC requires only two time

slots; hence, it can potentially double the throughput. A large volume of research on PNC, mainly focusing on the TWRC, has been reported in the literature, which outperform the conventional NC technique in terms of throughput and achievable data rates. In [7, 8], the authors present a PNC scheme based on frequency-shift keying (FSK).

More recently, a code-division multiple-access (CDMA) based analog network coding (ANC) scheme that utilizes amplify-and-forward (AF) relaying has been introduced in [9, 10] for multipath and asynchronous underwater acoustic sensor networks (UW-ASNs). The authors developed an adaptive RAKE receiver that equalizes the received signal and then jointly estimates the two multipath faded channels. The relay node cancels the interference before decoding the information of interest. The simulation and experiment results demonstrate that their proposed scheme can significantly improve the channel utilization by up to 50% for unidirectional and 100% for bidirectional networks compared to the conventional DS-CDMA scheme.

One of the main constraints of PNC and ANC techniques is the limitation on the number of users that can simultaneously transmit to the relay node. The concept of

a multiway relay channel (MWRC), which was introduced in [11], is a generalization of a TWRC, where $K > 2$ users simultaneously aim to achieve full information exchange with the help of a single relay node. A major application of MWRC is in satellite communication in which multiple users around the world simultaneously exchange data through a satellite. Several works studied different multiple access schemes to achieve MWRC. For example, in [12, 13] authors present a transmission method that uses binary phase-shift keying (BPSK), where multiuser detection (MUD) is achieved at the expense of an increase of channel use by identifying “minority” nodes. In [14], MUD relies on iterative multiuser detection of users at the receiver. Nonorthogonal multiple access schemes can be considered to be potential candidate for MWRC.

A well-known example of nonorthogonal multiple access is CDMA. The existence of uniquely decodable (UD) codes for overloaded synchronous CDMA where the number of multiplexed signals K is greater than the signature length L over the antipodal alphabet $\{\pm 1\}$ with linear MUD in noiseless channels is reported in [15]. However, in a noisy channel the high computation cost involved in MUD, which may increase in an exponential order with the number of users, has limited use in practical applications.

An interesting technique referred to as interleave-division multiple-access (IDMA) was introduced in [16]. This technique is attractive because of its low-complexity receiver design. A similar technique referred to as sparse code multiple access (SCMA) that possesses a low-complexity receiver is also presented in [17]. Authors in [18] utilize the UD code set over the ternary alphabet $\{\pm 1, 0\}$, which as mentioned before has linear MUD in noiseless environment only. In addition to that the authors present a decision decoding scheme at the relay for the $K = 3$ user case only, with $8 = 2^3$ decision regions in a noisy channel.

In this paper, we focus on a MWRC network, where $K \geq 2$ users exchange information with each other simultaneously via the help of a relay node. For MWRC networks utilizing PNC, it is proved in [19] that $2(K - 1)$ TSs are required. Hence, we propose for the first time to apply ternary UD code sets, which will allow each user to uniquely recover the information bits from a noisy channel. One of the attractive features of the proposed scheme is that it utilizes a very simple decoding algorithm, which requires only a few logical comparisons. Simulation results in terms of bit error rate (BER) demonstrate that the performance of the proposed decoder is very close to the maximum-likelihood (ML) decoder. Moreover, through simulations, we show that the proposed scheme can significantly improve the sum-rate capacity, which in turn can potentially improve overall throughput, as it needs only two TSs to exchange information compared to the conventional methods.

The rest of the paper is organized as follows. In Section 2, we present the system model. In Section 3, we discuss the construction of the uniquely decodable code sets followed by decoding algorithm in Section 4. The complexity analysis is performed in Section 5. After illustrating simulation results in Section 6, a few conclusions are drawn in Section 7.

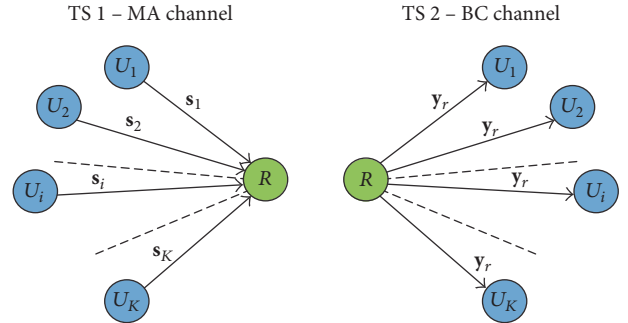


FIGURE 1: PNC with AF criteria.

The following notations are used in this paper. All bold-face lower case letters indicate column vectors and upper case letters indicate matrices, $()^T$ denotes transpose operation, \mathbb{C} denotes the set of all complex numbers, $*$ denotes conjugate, sgn denotes the sign function, $|\cdot|$ denotes complex amplitude, and $\lfloor \cdot \rfloor$, $\lceil \cdot \rceil$ are the floor and ceiling functions, respectively.

2. System Model

We consider a MWRC network with $K \geq 2$ users where each user intends to transmit binary data to all the other user nodes; that is, full data exchange is desired. Direct communication is assumed infeasible, and thus, users can only exchange information through the help of a relay node, as shown in Figure 1. We assume that perfect synchronization is available during the whole transmission, and all channel state information (CSI) is known at all the user nodes and the relay node [20]. Unlike the conventional PNC technique in which DNF is utilized, in this paper, we explore PNC with the AF criteria. Notice that the proposed scheme is still coined PNC and not ANC because of synchronization assumption and the fact that the relay node can potentially detect all of the users’ information even without the knowledge of users’ previous binary data. The communication takes place in two phases, that is, multiple access (MA) and broadcast (BC) phases. In each TS_i, that is, the MA phase, each user encodes the data sequence and then simultaneously transmits their signals to the relay through an additive white Gaussian noise (AWGN) channels. The received signal at the relay node can be expressed as

$$\mathbf{y}_r = \sum_{k=1}^K \alpha_k h_k \mathbf{c}_k d_k + \mathbf{n} = \sum_{k=1}^K h_k \mathbf{s}_k + \mathbf{n}, \quad (1)$$

where $\mathbf{c}_k \in \{\pm 1, 0\}^{L \times 1}$ is the ternary spreading code assigned to the user k from UD code set $\mathbf{C}_{L \times K}$, h_k stands for the complex channel gain between the k -th user and the relay node, $d_k \in \{\pm 1\}$ is BPSK modulated data bits, and the channel noise \mathbf{n} is assumed to be AWGN. Since each user has perfect CSI at the transmitter it is reasonable to utilize transmit diversity scheme, where it precodes the signal by $\alpha_k = h_k^* / |h_k|^2$ before the transmission to mitigate the channel effects [20]. For each user, the encoder combines its data sequence d_k into $\mathbf{s}_k = \alpha_k \mathbf{c}_k d_k$, which is then transmitted through the MA channel, as shown in Figure 2.

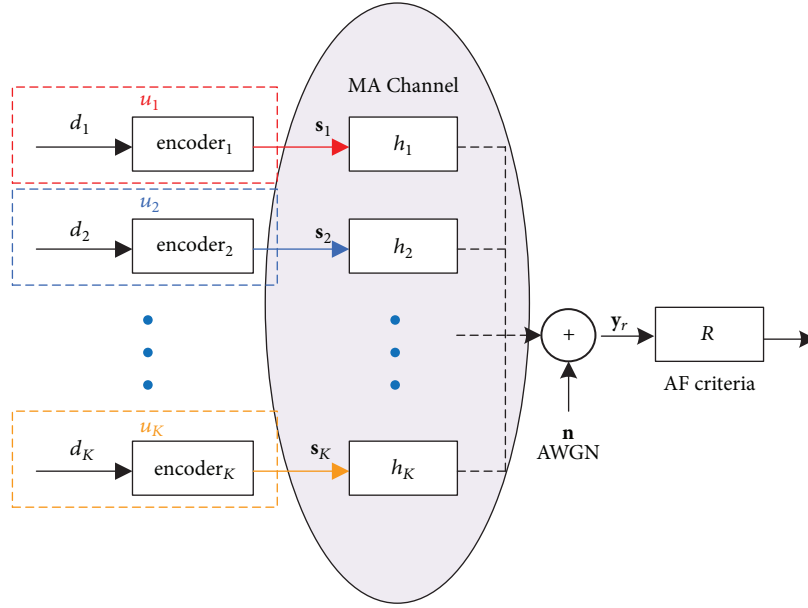


FIGURE 2: MA phase (TS1).

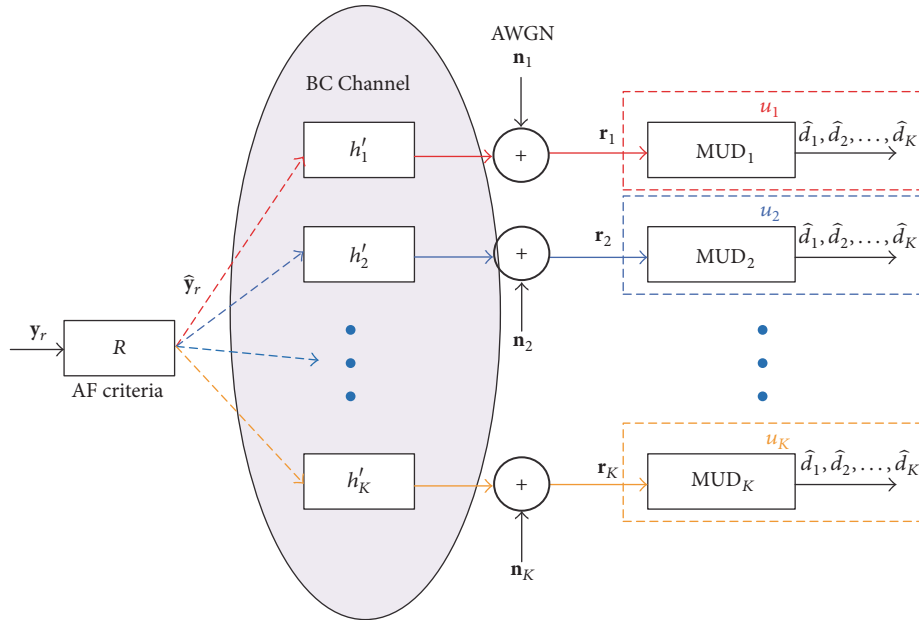


FIGURE 3: BC phase (TS2).

The BC phase is shown in Figure 3. In TS2, the relay node may apply the AF criteria and amplify the received vector \mathbf{y}_r to obtain $\hat{\mathbf{y}}_r$ depending on the channel conditions. The relay node broadcasts the combined sum-signals to all the user nodes. Each end user's received signal is given by

$$\mathbf{r}_k = h'_k \hat{\mathbf{y}}_r + \mathbf{n}_k, \quad (2)$$

where h'_k is the complex channel gain between the relay node and the k -th user and \mathbf{n}_k is AWGN. Note that to get rid of the channel effects each user can multiply the received signal

by $\alpha'_k = h_k^* / |h'_k|^2$ before applying the MUD. In the following section, we will present the construction of the proposed UD code sets $\mathcal{C}_{L \times K}$ and their linear MUD decoder.

The beauty behind the proposed scheme is that any user node can uniquely decode every user's binary data from the received sum-signal without using any table, as discussed in [18], or previous decoded data, which is the case for the ANC. Compared to the conventional PNC, sum data rate is much greater than 1. At the relay, we apply AF criteria instead of performing DNF.

3. Iterative Construction

In the proposed MWRC system, we utilize the UD code set that allows K -users to exchange information with the help of the relay node in only 2 TSs. These codes along with the proposed linear MUD decoder make a perfect candidate for MWRC systems.

We recall that a ternary code set $\mathbf{C} \in \{0, \pm 1\}^{L \times K}$ is uniquely decodable over signals $\mathbf{x} \in \{\pm 1\}^{K \times 1}$ or $\mathbf{x} \in \{0, 1\}^{K \times 1}$, if and only if, for any $\mathbf{x}_1 \neq \mathbf{x}_2$, $\mathbf{C}\mathbf{x}_1 \neq \mathbf{C}\mathbf{x}_2$ or, equivalently, $\mathbf{C}(\mathbf{x}_1 - \mathbf{x}_2) \neq \mathbf{0}_{L \times 1}$ [15]. We can rewrite the unique decodability necessary and sufficient condition as $\text{Null}(\mathbf{C}) \cap \{0, \pm 2\}^{K \times 1} = \{0\}^{K \times 1}$ or in an equivalent manner as

$$\text{Null}(\mathbf{C}) \cap \{0, \pm 1\}^{K \times 1} = \{0\}^{K \times 1}. \quad (3)$$

Let $f_t(L)$ represent the maximum number of columns (signals) that matrix can have for a given L and still be uniquely decodable. For the ternary code matrix with codes of length $L = 2$, $f_t(2)$ is simple and can be found by looking at the total number of possible columns $3^2 = 9$. Excluding the $[0, 0]^T$ column, half of the remaining is the negative of the other half, which makes it a total of 4 distinct columns that can be chosen to be $[0, 1]^T$, $[1, 0]^T$, $[1, -1]^T$, and $[1, 1]^T$. We conclude that no possible distinct combinations of these 4 columns satisfy uniquely decodability criteria (3). Out of all the possible combinations there are only few matrices with number of columns 3 that satisfy (3); therefore, $f_t(2) = 3$. Every possible ternary matrix of dimension 2×3 that has uniquely decodable property can be reduced to

$$\mathbf{C}_{2 \times 3}^1 = \begin{bmatrix} +1 & +1 & +1 \\ +1 & 0 & -1 \end{bmatrix}, \quad (4)$$

by applying operations such as multiplying columns by negative one, permuting rows, and columns. For the case of $L = 3$ and $L = 4$ it can be shown with an exhaustive search that $f_t(3) = 5$ and $f_t(4) = 8$, respectively.

In the preparation of general construction of matrices having $L = 2^i$, where $i \geq 2$, we carefully choose our seed matrix $\mathbf{C}_{4 \times 8}^2$ from distinct uniquely decodable matrices, which are found by exhaustive search,

$$\mathbf{C}_{4 \times 8}^2 = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & 0 & -1 & -1 & -1 \\ +1 & +1 & 0 & -1 & 0 & +1 & 0 & -1 \\ +1 & 0 & 0 & -1 & 0 & -1 & 0 & +1 \end{bmatrix}. \quad (5)$$

Next, we are ready to propose a general $L_i \times K_i$ code set design when $L_i = 2^i$ with $K_i = 2^{i+1} + 2^{i-2} - 1$, $i = 3, 4, \dots$. Starting from $\mathbf{C}_{4 \times 8}^2$ the following recursive relation defines

a sequence of matrices. The i th recursive matrix $\mathbf{C}_{L_i \times K_i}^i$ is formed as follows:

$$\mathbf{C}_{L_i \times K_i}^i = \begin{bmatrix} +1 & \cdots & +1 & +1 & +1 & \cdots & +1 \\ +1 & \cdots & +1 & 0 & -1 & \cdots & -1 \\ & & & 0 & & & \\ & & \widehat{\mathbf{C}}^{i-1} & 0 & & \mathbf{0} & \\ & & & \vdots & & & \\ & & \mathbf{0} & 0 & & \widehat{\mathbf{C}}^{i-1} & \\ & & & 0 & & & \end{bmatrix}, \quad (6)$$

where $L_i = 2L_{i-1}$, $K_i = 2K_{i-1} + 1$, and $\widehat{\mathbf{C}}^{i-1}$ is derived by eliminating the first row of $\mathbf{C}_{L_{i-1} \times K_{i-1}}^{i-1}$. The above code sequences $\mathbf{C}_{L_i \times K_i}^i$ preserve the uniquely decodability property, given that $\widehat{\mathbf{C}}^{i-1}$ is a UD matrix.

4. The Proposed Fast Decoder

In this section, we present our proposed fast decoder algorithm (FDA). In the system with signature matrix $\mathbf{C} \in \{\pm 1, 0\}^{L \times K}$, where the columns are the user spreading codes. At k 's user, the received vector \mathbf{r}_k after multiplication by α_k' is expressed by

$$\mathbf{y} = \sum_{k=1}^K \mathbf{c}_k x_k + \mathbf{n} = \mathbf{C}\mathbf{x} + \mathbf{n}, \quad (7)$$

where $\mathbf{c}_j \in \{\pm 1, 0\}^{L \times 1}$ are signatures for $1 \leq j \leq K$, $\mathbf{x} \in \{\pm 1\}^{K \times 1}$ is user data, and \mathbf{n} is AWGN noise. The objective of the receiver is the following: given the received vector \mathbf{y} and \mathbf{C} recover the user data $\widehat{\mathbf{x}}$ such that the mean square error $E\{\|\mathbf{x} - \widehat{\mathbf{x}}\|^2\}$ is minimized. It is known that obtaining the ML solution is generally NP-hard [21].

For our detection problem, where the overloaded signature matrix has a UD structure, can be solved efficiently if there is a function that maps $\mathbf{y} \mapsto \widehat{\mathbf{y}} \in \Lambda$, where Λ is a \mathbb{Z} -module with rank L . It is equivalent to finding the closest point in a lattice Λ , such that

$$\widehat{\mathbf{y}} = \arg \min_{\mathbf{y}' \in \Lambda} \|\mathbf{y} - \mathbf{y}'\|^2. \quad (8)$$

Gaining the knowledge of $\widehat{\mathbf{y}}$, one of the points in Λ generated by \mathbf{C} , we can obtain $\widehat{\mathbf{x}}$ uniquely, since \mathbf{C} satisfies the uniquely decodability criteria (3). However, there is no known polynomial algorithm that can obtain $\widehat{\mathbf{y}}$ from \mathbf{y} .

We first present the general form of the proposed FDA for the $\mathbf{C}_{L_i \times K_i}^i$, $i \geq 2$ case, where the vector $\mathbf{1}$ is defined as $\mathbf{1} \in 1^{K \times 1}$ and the quantizer $Q: \mathbb{R} \mapsto \mathcal{N}$, $z_1 = Q(y, -K, K)$ is a mapping of $y \in \mathbb{R}$ to the constellation of $\{\pm K, \pm(K-2), \dots\}$. The output of the quantizer z_1 shows the number of -1 s in $\widehat{\mathbf{x}}$. Furthermore, let $m_1, m_2, m_3, m_{11}, k_1, k_2$, and k_3 represent the number of -1 's at (1, 2), 3, 4, 1, 6, 7, and 8 locations of $\widehat{\mathbf{x}}$, respectively. Note that when $z_1 = K$ or $z_1 = -K$

Input: \mathbf{y}

- (1) $z_1 \leftarrow Q(y_1, -K, K)$
- (2) **if** $z_1 = |K|$, $\hat{\mathbf{x}} \leftarrow \text{sgn}(z_1)\mathbf{1}$
- (3) **else**
- (4) $n \leftarrow (K - z_1)/2$
- (5) $z_2 \leftarrow Q(y_2, -(K - |z_1|), K - |z_1|)$
- (6) $n_l \leftarrow (2n - z_2)/4$, $n_r \leftarrow n - n_l$
- (7) $n_l \leftarrow \lfloor n_l \rfloor$, $n_r \leftarrow \lfloor n_r \rfloor$
- (8) **if** $K = 8$, $\hat{\mathbf{x}} \leftarrow \text{subDecoder}(\mathbf{y}, n_l, n_r)$
- (9) **else**
- (10) $\hat{\mathbf{y}}_l \leftarrow [(2^i + 2^{i-3} - 1 - 2n_l), y_3, \dots, y_{2^{i-1}+1}]^T$
- (11) $\hat{\mathbf{y}}_r \leftarrow [(2^i + 2^{i-3} - 1 - 2n_r), y_{2^{i-1}+2}, \dots, y_{2^i}]^T$
- (12) $\hat{\mathbf{x}}_l \leftarrow \text{decoder}(\hat{\mathbf{y}}_l)$, $\hat{\mathbf{x}}_r \leftarrow \text{decoder}(\hat{\mathbf{y}}_r)$
- (13) $x_m \leftarrow z_1 - (\hat{\mathbf{x}}_l^T \mathbf{1} + \hat{\mathbf{x}}_r^T \mathbf{1})$, $\hat{\mathbf{x}} \leftarrow [\hat{\mathbf{x}}_l^T, x_m, \hat{\mathbf{x}}_r^T]^T$

Output: $\hat{\mathbf{x}}$

ALGORITHM 1: Fast decoder algorithm (FDA).

Input: \mathbf{y}, n, n_l, n_r

- (1) **if** $n_l = 0$, $[m_1, m_2, m_3, m_{11}] \leftarrow [0, 0, 0, 0]$, $S_l \leftarrow 1$
- (2) **elseif** $n_l = 4$, $[m_1, m_2, m_3, m_{11}] \leftarrow [2, 1, 1, 1]$, $S_l \leftarrow 1$
- (3) **if** $n_r = 0$, $[k_1, k_2, k_3] \leftarrow [0, 0, 0]$, $S_r \leftarrow 1$
- (4) **elseif** $n_r = 3$, $[k_1, k_2, k_3] \leftarrow [1, 1, 1]$, $S_r \leftarrow 1$
- (5) **if** $S_l = 1$ AND $S_r = 0$,
- (6) $[k_1, k_2, k_3] \leftarrow \text{rightDecoder}(\mathbf{y}, m_1, m_2, m_3, m_{11})$
- (7) **if** $S_l = 0$ AND $S_r = 1$,
- (8) $[m_1, m_2, m_3, m_{11}] \leftarrow \text{leftDecoder}(\mathbf{y}, k_1, k_2, k_3)$
- (9) **else**, $S_l = 0$ AND $S_r = 0$
- (10) $[m_1, m_2, m_3, m_{11}, k_1, k_2, k_3] \leftarrow \text{lrDecoder}(\mathbf{y})$
- (11) $[\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4] \leftarrow -2[m_{11}, (m_1 - m_{11}), m_3, m_2] + 1$
- (12) $\hat{x}_5 \leftarrow -2(n - n_l - n_r) + 1$
- (13) $[\hat{x}_6, \hat{x}_7, \hat{x}_8] \leftarrow -2[k_1, k_2, k_3] + 1$

Output: $\hat{\mathbf{x}}$

ALGORITHM 2: subDecoder algorithm.

only one comparison is required. The algorithm proceeds by computing n , n_l , and n_r , which denote the number of -1 's in $\hat{\mathbf{x}}$, $[\hat{x}_1, \dots, \hat{x}_{(K-1)/2}]$, and $[\hat{x}_{(K-1)/2+1}, \dots, \hat{x}_K]$, respectively.

For the case of $\mathbf{C}_{2 \times 3}^1$ the decoding is trivial and will not be covered in this article, instead we start with the nonsymmetric case of $\mathbf{C}_{4 \times 8}^2$. The FDA, shown in Algorithm 1, calls the *subDecoder* (Algorithm 2) at line (8) with $[y_1, \dots, y_4]^T$, n_l , and n_r parameters. This algorithm will proceed in four different paths depending on n_l and n_r . If n_l is 0 or 4 then the *leftDecoder* (Algorithm 4) will never be called and will assign $[m_1, m_2, m_3, m_{11}] = [0, 0, 0, 0]$ or $[m_1, m_2, m_3, m_{11}] = [2, 1, 1, 1]$, respectively. Similarly, if n_r is 0 or 3 then the *rightDecoder* (Algorithm 3) will never be called and will assign $[k_1, k_2, k_3] = [0, 0, 0]$ or $[k_1, k_2, k_3] = [1, 1, 1]$, respectively. Therefore, the trivial case is when both the *leftDecoder* and the *rightDecoder* are not required; other scenarios are that the *rightDecoder* is called when the *leftDecoder* is not required, the *leftDecoder* is called when the *rightDecoder* is not required, and the last case is when left and right decoder, *lrDecoder*, is called (Algorithm 5).

Input: $\mathbf{y}, n_r, m_1, m_2$

- (1) $y_{3m} \leftarrow (y_3 - 1)/2 - m_2 + m_1$
- (2) $z_{3m} \leftarrow Q(y_{3m}, -1, +1)$
- (3) $k_2 \leftarrow \lfloor (z_{3m} + n_r)/2 \rfloor$
- (4) $k_3 \leftarrow z_{3m} + n_r - 2k_2$
- (5) $k_1 \leftarrow n_r - k_2 - k_3$

Output: $[k_1, k_2, k_3]$

ALGORITHM 3: rightDecoder algorithm.

Input: $\mathbf{y}, n_l, k_1, k_2$

- (1) $y_{3k} \leftarrow (y_3 - 1)/2$
- (2) $z_{3k} \leftarrow Q(y_{3k}, -k_1 + k_2 + \delta_{\min}, -k_1 + k_2 + \delta_{\max})$
- (3) $m_2 \leftarrow \lfloor (z_{3k} - k_2 + k_1 + n_l)/2 \rfloor$
- (4) $m_3 \leftarrow z_{3k} - k_2 + k_1 + n_l - 2m_2$
- (5) $m_1 \leftarrow n_l - m_2 - m_3$
- (6) **if** $m_1 = 2$, $m_{11} \leftarrow 1$
- (7) **elseif** $m_1 = 0$, $m_{11} \leftarrow 0$
- (8) **elseif** $y_4/2 - k_1 - m_2 + k_2 \geq -0.5$, $m_{11} \leftarrow 0$
- (9) **else**, $m_{11} \leftarrow 1$

Output: $[m_1, m_2, m_3, m_{11}]$

ALGORITHM 4: leftDecoder algorithm.

Input: \mathbf{y}, n_l, n_r

- (1) $y_{3n} \leftarrow (y_3 - 1)/2$, $d_3 \leftarrow e^{10}$
- (2) $z_{3n} \leftarrow Q(y_{3n}, -\delta_{\min} - 1, \delta_{\max} + 1)$
- (3) **for** $\delta_3 \in \{-1 + \gamma_{\min}, \dots, -1 + \gamma_{\max}\}$
- (4) $m'_2 \leftarrow \lfloor (z_{3n} - \delta_3 + n_l)/2 \rfloor$
- (5) $m'_3 \leftarrow z_{3n} - \delta_3 + n_l - 2m'_2$
- (6) $m'_1 \leftarrow n_l - m'_2 - m'_3$, $k'_2 \leftarrow \lfloor (\delta_3 + n_r)/2 \rfloor$
- (7) $k'_3 \leftarrow n_r + \delta_3 - 2k'_2$, $k'_1 \leftarrow n_r - k'_2 - k'_3$
- (8) **if** $m'_1 = 2$, $m'_{11} \leftarrow 1$
- (9) **elseif** $m'_1 = 0$, $m'_{11} \leftarrow 0$
- (10) **elseif** $y_4/2 - k'_1 - m'_2 + k'_2 \geq -0.5$, $m'_{11} \leftarrow 0$
- (11) **else** $m'_{11} \leftarrow 1$
- (12) **if** $d'_3 \leftarrow |y_4/2 + m'_{11} - m'_2 - k'_1 + k'_2| < d_3$
- (13) $[m_1, m_2, m_3, m_{11}] \leftarrow [m'_1, m'_2, m'_3, m'_{11}]$
- (14) $[k_1, k_2, k_3] \leftarrow [k'_1, k'_2, k'_3]$
- (15) $d_3 \leftarrow d'_3$

Output: $[m_1, m_2, m_3, m_{11}, k_1, k_2, k_3]$

ALGORITHM 5: lrDecoder algorithm.

The *rightDecoder* and the *leftDecoder* decoders are straightforward, having the knowledge of $(\mathbf{y}, n_r, m_1, m_2)$ the *rightDecoder* computes (k_1, k_2, k_3) , and similarly, having the knowledge of $(\mathbf{y}, n_l, k_1, k_2)$, the *leftDecoder* computes (m_1, m_2, m_3, m_{11}) (Algorithms 3 and 4).

Note that the parameters in the *leftDecoder* and the *lrDecoder* are computed as such $\delta_{\min} = -\text{rnd}(3(n_l + 1)/5)$, $\delta_{\max} = \text{mod}(\text{rnd}(3n_l/5), 2)$, $\gamma_{\min} = (\text{sgn}(\eta - 1/10) + 1)\eta/2$, and $\gamma_{\max} = \lambda(\zeta - 3)/2 - 1$, where $\eta = \zeta + \delta_{\min} - \delta_{\max} - 1$,

TABLE 1: Complexity of the proposed ternary codes.

Decoder	Complexity	(4 × 8)	(8 × 17)	(16 × 35)
Proposed	Comparisons	5.86	17.98	50.24
ML	Comparisons	2 ⁸	2 ¹⁷	2 ³⁵

$\lambda = \text{sgn}(31/10 - \zeta) + 1$ and ζ is the index of the constellation returned by $Q(\cdot)$ function (Algorithms 4 and 5).

Having all the required information now the *subDecoder* assigns $[\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_6, \hat{x}_7, \hat{x}_8] = -2[m_{11}, (m_1 - m_{11}), m_3, m_2, k_1, k_2, k_3] + 1$ and $\hat{x}_5 = -2(n - n_l - n_r) + 1$. Now we completed the case when $K = 8$; the rest of the algorithm in FDA proceeds by applying the general *decoder* algorithm with the inputs of \hat{y}_l and \hat{y}_r to obtain \hat{x}_l and \hat{x}_r , respectively, to find the middle element $x_m = z_l - (\hat{x}_l^T \mathbf{1} + \hat{x}_r^T \mathbf{1})$. The decoded data is $\hat{\mathbf{x}} = [\hat{x}_l^T, x_m, \hat{x}_r^T]^T$. In the following section, we study the complexity of the proposed fast decoder analytically.

5. Complexity Analysis

The proposed decoder, discussed in Section 4, deciphers all the users' data at the receiver side in a recursive manner. In this section, we demonstrate the computational complexity analytically. It is important to state that the proposed FDA neither requires any multiplications nor additions; instead, only a few comparisons are performed in the $Q(\cdot)$ function. First, we will look at the average number of comparisons required for the $\mathbf{C}_{4 \times 8}^2$ case, whose decoding algorithm is presented in the *subDecoder* algorithm. Since, our proposed $\mathbf{C}_{4 \times 8}^2$ matrix is nonsymmetric, we will analyze the complexity of decoding all the 2⁸ possible input vectors. By closely analyzing FDA algorithm the comparison required for $n = 0, 1, 2, 3, 4, 5, 6, 7, 8$ is 1, 25, 144, 289, 488, 369, 155, 28, 1, respectively, and there are $\binom{8}{n}$ of input vectors per n . There are a total of 1500 comparisons; hence, the average computational complexity is $T_2 = 1500/256 = 5.86$. The recursive structure of our proposed matrices for $i \geq 3$ possesses symmetries that enable us to present the general case. In order to express the relationship for T_i , where $i \geq 3$, we will first introduce a few definitions. Let us define

$$G_i = \sum_{j=0}^{2^i + 2^{(i-3)} - 1} \binom{2^{(i+1)} + 2^{(i-2)} - 1}{j} (j+1), \quad (9)$$

$$H_i = \sum_{j=1}^{2^i + 2^{(i-3)} - 1} \left\{ \binom{2^i + 2^{(i-3)} - 1}{\lfloor \frac{j-1}{2} \rfloor} \right\} (j+1) + 2 \sum_{k=0}^{\lfloor (j-1)/2 \rfloor} \binom{2^{(i-3)} - 1}{k} \binom{2^i + 2^{(i-3)} - 1}{j-k} (2k+1) + 2 \sum_{k=0}^{\lfloor (j-2)/2 \rfloor} \binom{2^i + 2^{(i-3)} - 1}{k} \binom{2^i + 2^{(i-3)} - 1}{j-k-1} (2k+2) \Big\}, \quad (10)$$

$$U_i = 4 \left(2^{2^i - 1} - 2 \right) + 2 \sum_{j=2}^{2^i + 2^{(i-3)} - 1} \left\{ \binom{2^i + 2^{(i-3)} - 1}{\lfloor \frac{j-1}{2} \rfloor} \right\}^2 + 2 \sum_{k=1}^{\lfloor (j-1)/2 \rfloor} \binom{2^i + 2^{(i-3)} - 1}{k} \binom{2^i + 2^{(i-3)} - 1}{j-k} + 2 \sum_{k=1}^{\lfloor (j-2)/2 \rfloor} \binom{2^i + 2^{(i-3)} - 1}{k} \binom{2^i + 2^{(i-3)} - 1}{j-k-1} \Big\}, \quad (11)$$

where G_i is the number of comparisons that are required in the first call of the $Q(\cdot)$ function. If the input vector contains j number of -1 's, in $Q(\cdot)$ function it needs $(j+1)$ comparisons, as shown in (9). Note that, due to symmetry, we do not consider all the input vectors $\mathbf{x} \in \{\pm 1\}^K$, instead, only half of them, that is, $2^i + 2^{(i-3)} - 1$. The H_i is related to the number of comparisons required in the second call of the $Q(\cdot)$ function, while the last term U_i shows how many times left and/or right subdecoders are called. The general relation for $i \geq 3$ can be expressed as

$$T_i = \frac{1}{2^{2^{(i+1)} + 2^{(i-2)} - 2}} [G_i + H_i + U_i \times \hat{T}_{i-1}], \quad (12)$$

where

$$\hat{T}_{i-1} = \frac{1}{2^{2^i + 2^{(i-3)} - 2} - 1} [2^{2^i + 2^{(i-3)} - 2} T_{i-1} - G_{i-1}], \quad (13)$$

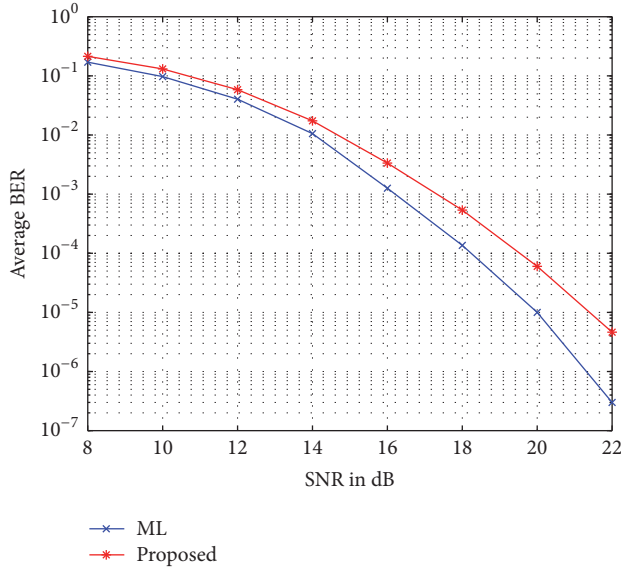
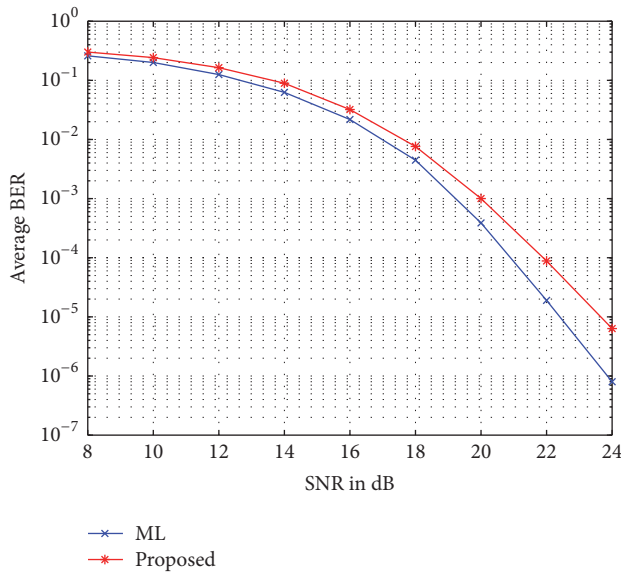
is the altered version of the T_{i-1} by excluding the number of comparisons in the first call of the $Q(\cdot)$ calculations.

In Table 1, we show the complexity results for (4 × 8), (8 × 17), (16 × 35) using the proposed FDA and ML algorithms. As we can see, the complexity of ML decoder increases exponentially, while the proposed decoder has fairly small complexity even for a relatively large matrix size (16 × 35).

6. Simulation Results

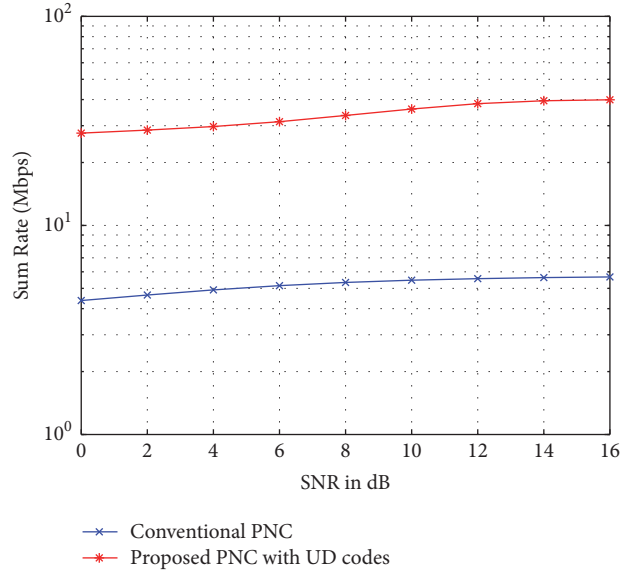
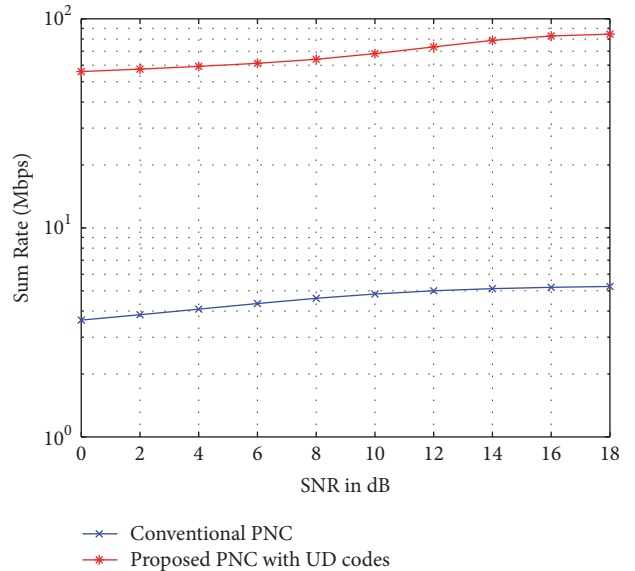
In this section, we evaluate the performance of the MWRC network by employing our proposed ternary uniquely decodable codes at the physical layer. All the simulations at the physical layer of the proposed scheme are performed in Matlab. We consider wireless transmission between $K = 8$ and $K = 17$ users. In the MA phase, each user k spreads its data $d_k \in \{\pm 1\}$, using BPSK modulation and the proposed ternary code \mathbf{c}_k , and then transmits to the relay node through a slow Rayleigh fading channel. The relay node in BC phase transmits the combined sum-signals to all the user nodes.

We assume that the transmission is synchronous and all CSI is known at all the user as well as relay nodes. Each user gets rid of the complex channel effects and then


 FIGURE 4: Average BER versus SNR for the UD codes, $C_{4 \times 8}^2$.

 FIGURE 5: Average BER versus SNR for the UD codes, $C_{8 \times 17}^3$.

applies the proposed FDA to decode its information. For comparison purposes, we compare FDA algorithm with the optimum ML decoder. In Figures 4 and 5, we plot the BER performance averaged over all the different users for the UD code sets of $C_{4 \times 8}^2$ and $C_{8 \times 17}^3$, respectively. For a BER of 10^{-3} the performance of FDA is only about 1 dB worse than the ML. In other words, our proposed FDA achieves near-ML performance without having an exponentially complex algorithm.

In order to show the advantage of PNC with UD codes compared to the conventional PNC, we first define the sum rate to be the number of correctly transmitted bits per unit time. Suppose that the bit rate of each node is 10 Mbps. In the case of conventional PNC $2(K-1)$ TSs are required compared


 FIGURE 6: Sum rate versus SNR for $K = 8$ user nodes using UD codes, $C_{4 \times 8}^2$.

 FIGURE 7: Sum rate versus SNR for $K = 17$ user nodes using UD codes, $C_{8 \times 17}^3$.

to only 2 TSs for the PNC with UD code. Thus, the sum rates are $8/14 \cdot (1 - P_{e,\text{PNC}}) \cdot 10$ Mbps and $17/32 \cdot (1 - P_{e,\text{PNC}}) \cdot 10$ Mbps, where $P_{e,\text{PNC}}$ is the error rate of conventional PNC, with $K = 8$ and $K = 17$ user nodes, respectively. Using the conventional PNC 80 Mbps and 170 Mbps are exchanged during 14 and 32 TSs. Meanwhile, the sum rates of the proposed PNC with UD code sets are $8/2 \cdot (1 - P_e) \cdot 10$ Mbps and $17/2 \cdot (1 - P_e) \cdot 10$ Mbps, where P_e is the error rate of the detector of the UD codes.

In Figures 6 and 7, we plot the sum rates for the cases of $K = 8$ and $K = 17$ user nodes, respectively. We can see from Figures 6 and 7 that the proposed scheme utilizing UD codes improves the sum rates by almost 7 and 16 times, respectively,

compared to conventional PNC scheme. When the signal-to-noise ratio (SNR) is high enough, the sum rate is nearly enhanced by about $K - 1$ times.

7. Conclusion

In this paper, we have proposed to apply a ternary uniquely decodable (UD) code sets to a multiway relay channel (MWRC) network where two or more users are able to exchange information with each other simultaneously through the help of a relay node. A key feature of the proposed scheme is that it utilizes a novel UD code sets in a transmission, which requires only two time slots (TSs) as opposed to $2(K - 1)$ TSs for the conventional physical-layer network coding (PNC) scheme. We developed for the proposed UD code set a very simple decoding algorithm, which requires only a few logical comparisons. Simulation results in terms of bit error rate (BER) and sum rates demonstrate that the proposed decoder outperforms the conventional methods. The performance of the proposed low computational cost decoder is almost as good as the maximum-likelihood (ML) decoder.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] R. Ahlswede, N. Cai, S. R. Li, and R. W. Yeung, "Network information flow," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.
- [2] T. Matsuda, T. Noguchi, and T. Takine, "Survey of network coding and its applications," *IEICE Transactions on Communications*, vol. 94, no. 3, pp. 698–717, 2011.
- [3] M. Medard and A. Sprintson, *Network Coding: Fundamentals and Applications*, Academic Press, 2011.
- [4] C. Fragouli and E. Soljanin, "Network coding fundamentals," *Foundations and Trends in Networking*, vol. 2, no. 1, pp. 1–133, 2007.
- [5] T. Ho and D. S. Lun, *Network Coding: An Introduction*, Cambridge University Press, New York, NY, USA, 2008.
- [6] S. Zhang, S. Liew, and P. Lam, "Hot topic: physical-layer network coding," in *Proceedings of the ACM 12th Annual International Conference of Mobile Computing and Networks (MobiCom '06)*, pp. 358–365, Los Angeles, USA, September 2006.
- [7] Q.-Y. Yu, D.-Y. Zhang, H.-H. Chen, and W.-X. Meng, "Physical-layer network coding systems with MFSK modulation," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 1, pp. 204–213, 2016.
- [8] J. H. Sørensen, R. Krigslund, P. Popovski, T. K. Akino, and T. Larsen, "Physical layer network coding for FSK systems," *IEEE Communications Letters*, vol. 13, no. 8, pp. 597–599, 2009.
- [9] H. Kulhandjian, T. Melodia, and D. Koutsonikolas, "Securing underwater acoustic communications through analog network coding," in *Proceedings of the 2014 11th Annual IEEE International Conference on Sensing, Communication, and Networking, SECON 2014*, pp. 266–274, Singapore, July 2014.
- [10] H. Kulhandjian, T. Melodia, and D. Koutsonikolas, "CDMA-Based Analog Network Coding for Underwater Acoustic Sensor Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6495–6507, 2015.
- [11] D. Gunduz, A. Yener, A. Goldsmith, and H. V. Poor, "The multiway relay channel," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 59, no. 1, pp. 51–63, 2013.
- [12] S. Sharifian and T. A. Gulliver, "Performance of physical-layer network coded multi-way relay channels with binary signaling," in *Proceedings of the 14th IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing, PACRIM 2013*, pp. 292–295, Victoria, BC, Canada, August 2013.
- [13] R. Y. Chang, S. Lin, and W. Chung, "Transmission Protocol Design for Binary Physical Network Coded Multi-Way Relay Networks," in *Proceedings of the 2014 IEEE Vehicular Technology Conference (VTC'14)*, pp. 1–5, Seoul, South Korea, May 2014.
- [14] Z. A. Almaalie, X. Tang, Z. Ghassemlooy, I. E. Lee, and A. A. Al-Rubaie, "Iterative multiuser detection with physical layer network coding for multi-pair communications," in *Proceedings of the 2016 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, pp. 1–6, Prague, Czech Republic, July 2016.
- [15] M. Kulhandjian and D. A. Pados, "Uniquely decodable code-division via augmented Sylvester-Hadamard matrices," in *Proceedings of the 2012 IEEE Wireless Communications and Networking Conference, WCNC 2012*, pp. 359–363, Paris, France, April 2012.
- [16] 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.
- [17] M. Kulhandjian and C. D'Amours, "Design of permutation-based sparse code multiple access system," in *Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, Montreal, QC, October 2017.
- [18] Y. Li, Q. Yu, K. He, and W. Xiang, "Apply Uniquely-Decodable Codes to Multiuser Physical-Layer Network Coding Based on Amplify-and-Forward Criterion," in *Proceedings of the 2015 IEEE Global Telecommunications Conference 2015*, pp. 1–6, San Diego, CA, USA, December 2015.
- [19] D.-Y. Zhang, Q.-Y. Yu, W.-X. Meng, and C. Li, "2FSK modulation for multiuser physical-layer network coding network," in *Proceedings of the 2014 1st IEEE International Conference on Communications, ICC 2014*, pp. 514–519, Australia, June 2014.
- [20] T. Koike-Akino, P. Popovski, and V. Tarokh, "Adaptive modulation and network coding with optimized precoding in two-way relaying," in *Proceedings of the 2009 IEEE Global Telecommunications Conference, GLOBECOM 2009*, pp. 1–6, Honolulu, HI, USA, December 2009.
- [21] R. Lupas and S. Verdú, "Linear multiuser detectors for synchronous code-division multiple-access channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 35, no. 1, pp. 123–136, 1989.

Review Article

A Tutorial on Nonorthogonal Multiple Access for 5G and Beyond

Mahmoud Aldababsa,¹ Mesut Toka,^{1,2} Selahattin Gökçeli ,³
Güneş Karabulut Kurt,³ and Oğuz Kucur ¹

¹Electronics Engineering Department, Gebze Technical University, Gebze, 41400 Kocaeli, Turkey

²Electrical and Electronics Engineering Department, Ömer Halisdemir University, 51240 Niğde, Turkey

³Department of Communications and Electronics Engineering, Istanbul Technical University, 34469 Istanbul, Turkey

Correspondence should be addressed to Oğuz Kucur; okucur@gtu.edu.tr

Received 23 November 2017; Accepted 5 February 2018; Published 28 June 2018

Academic Editor: Nathalie Mitton

Copyright © 2018 Mahmoud Aldababsa et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Today's wireless networks allocate radio resources to users based on the orthogonal multiple access (OMA) principle. However, as the number of users increases, OMA based approaches may not meet the stringent emerging requirements including very high spectral efficiency, very low latency, and massive device connectivity. Nonorthogonal multiple access (NOMA) principle emerges as a solution to improve the spectral efficiency while allowing some degree of multiple access interference at receivers. In this tutorial style paper, we target providing a unified model for NOMA, including uplink and downlink transmissions, along with the extensions to multiple input multiple output and cooperative communication scenarios. Through numerical examples, we compare the performances of OMA and NOMA networks. Implementation aspects and open issues are also detailed.

1. Introduction

Wireless mobile communication systems became an indispensable part of modern lives. However, the number and the variety of devices increase significantly and the same radio spectrum is required to be reused several times by different applications and/or users. Additionally, the demand for the Internet of Things (IoT) introduces the necessity to connect every person and every object [1]. However, current communication systems have strict limitations, restricting any modifications and improvements on the systems to meet these demands. Recently, researchers have been working on developing suitable techniques that may be integrated in next generation wireless communication systems in order to fundamentally fulfill the emerging requirements, including very high spectral efficiency, very low latency, massive device connectivity, very high achievable data rate, ultrahigh reliability, excellent user fairness, high throughput, supporting diverse quality of services (QoS), energy efficiency, and a dramatic reduction in the cost [2]. Some potential technologies have

been proposed by the academia and the industry in order to satisfy the aforementioned tight requirements and to address the challenges of future generations. For example, millimeter wave (mmWave) technology was suggested to enlarge the transmission bandwidth for very high speed communications [3], massive multiple input multiple output (MIMO) concept was presented to improve capacity and energy efficiency [4], and ultradense networks were introduced to increase the throughput and to reduce the energy consumption through using a large number of small cells [5].

Besides the aforementioned techniques, a new radio access technology is also developed by researchers to be used in communication networks due to its capability in increasing the system capacity. Recently, nonorthogonality based system designs are developed to be used in communication networks and have gained significant attention of researchers. Hence, multiple access (MA) techniques can now be fundamentally categorized as orthogonal multiple access (OMA) and nonorthogonal multiple access (NOMA). In OMA, each user can exploit orthogonal communication resources within

either a specific time slot, frequency band, or code in order to avoid multiple access interference. The previous generations of networks have employed OMA schemes, such as frequency division multiple access (FDMA) of first generation (1G), time division multiple access (TDMA) of 2G, code division multiple access (CDMA) of 3G, and orthogonal frequency division multiple access (OFDMA) of 4G. In NOMA, multiple users can utilize nonorthogonal resources concurrently by yielding a high spectral efficiency while allowing some degree of multiple access interference at receivers [6, 7].

In general, NOMA schemes can be classified into two types: power-domain multiplexing and code-domain multiplexing. In power-domain multiplexing, different users are allocated different power coefficients according to their channel conditions in order to achieve a high system performance. In particular, multiple users' information signals are superimposed at the transmitter side. At the receiver side successive interference cancellation (SIC) is applied for decoding the signals one by one until the desired user's signal is obtained [8], providing a good trade-off between the throughput of the system and the user fairness. In code-domain multiplexing, different users are allocated different codes and multiplexed over the same time-frequency resources, such as multiuser shared access (MUSA) [9], sparse code multiple access (SCMA) [10], and low-density spreading (LDS) [11]. In addition to power-domain multiplexing and code-domain multiplexing, there are other NOMA schemes such as pattern division multiple access (PDMA) [12] and bit division multiplexing (BDM) [13]. Although code-domain multiplexing has a potential to enhance spectral efficiency, it requires a high transmission bandwidth and is not easily applicable to the current systems. On the other hand, power-domain multiplexing has a simple implementation as considerable changes are not required on the existing networks. Also, it does not require additional bandwidth in order to improve spectral efficiency [14]. In this review/tutorial paper, we will focus on the power-domain NOMA.

Although OMA techniques can achieve a good system performance even with simple receivers because of no mutual interference among users in an ideal setting, they still do not have the ability to address the emerging challenges due to the increasing demands in 5G networks and beyond. For example, according to International Mobile Telecommunications (IMT) for 2020 and beyond [15], 5G technology should support three main categories of scenarios, such as enhanced mobile broadband (eMBB), massive machine type communication (mMTC), and ultrareliable and low-latency communication (URLLC). The main challenging requirements of eMBB scenario are 100 Mbps user perceived data rate and more than 3 times spectrum efficiency improvement over the former LTE releases to provide services including high definition video experience, virtual reality, and augmented reality. Since a large number of IoT devices will have access to the network, the main challenge of mMTC is to provide connection density of 1 million devices per square kilometer. In case of URLLC, the main requirements include 0.5 ms end-to-end latency and reliability above 99.999% [16–18]. By using NOMA scheme, for mMTC and URLLC applications, the number of user connections can be increased by 5 and

9 times, respectively [18]. Also, according to [19], NOMA has been shown to be more spectral-efficient by 30% for downlink and 100% for uplink in eMBB when compared to OMA. Therefore, NOMA has been recognized as a strong candidate among all MA techniques since it has essential features to overcome challenges in counterpart OMA and achieve the requirements of next mobile communication systems [20–22]. The superiority of NOMA over OMA can be remarked as follows:

- (i) Spectral efficiency and throughput: in OMA, such as in OFDMA, a specific frequency resource is assigned to each user even it experiences a good or bad channel condition; thus the overall system suffers from low spectral efficiency and throughput. In the contrary, in NOMA the same frequency resource is assigned to multiple mobile users, with good and bad channel conditions, at the same time. Hence, the resource assigned for the weak user is also used by the strong user, and the interference can be mitigated through SIC processes at users' receivers. Therefore, the probability of having improved spectral efficiency and a high throughput will be considerably increased as depicted in Figure 1.
- (ii) User fairness, low latency, and massive connectivity: in OMA, for example in OFDMA with scheduling, the user with a good channel condition has a higher priority to be served while the user with a bad channel condition has to wait for access, which leads to a fairness problem and high latency. This approach can not support massive connectivity. However, NOMA can serve multiple users with different channel conditions simultaneously; therefore, it can provide improved user fairness, lower latency, and higher massive connectivity [20].
- (iii) Compatibility: NOMA is also compatible with the current and future communication systems since it does not require significant modifications on the existing architecture. For example, NOMA has been included in third generation partnership project long-term evolution advanced (3GPP LTE Release 13) [23–29]. More detailed, in the standards, a downlink version of NOMA, multiuser superposition transmission (MUST), has been used [23]. MUST utilizes the superposition coding concept for a multiuser transmission in LTE-A systems. In 3GPP radio access network (RAN), while using MUST, the deployment scenarios, evaluation methodologies, and candidate NOMA scheme have been investigated in [24–26], respectively. Then, system level performance and link level performance of NOMA have been evaluated in [27, 28], respectively. Next, 3GPP LTE Release 14 has been proposed [29], in which intracell interference is eliminated and hence LTE can support downlink intracell multiuser superposition transmission. Also, NOMA, known as layered division multiplexing (LDM), is used in the future digital TV standard, ATSC 3.0 [30]. Moreover, the standardization study of NOMA schemes for 5G New Radio (NR) continues

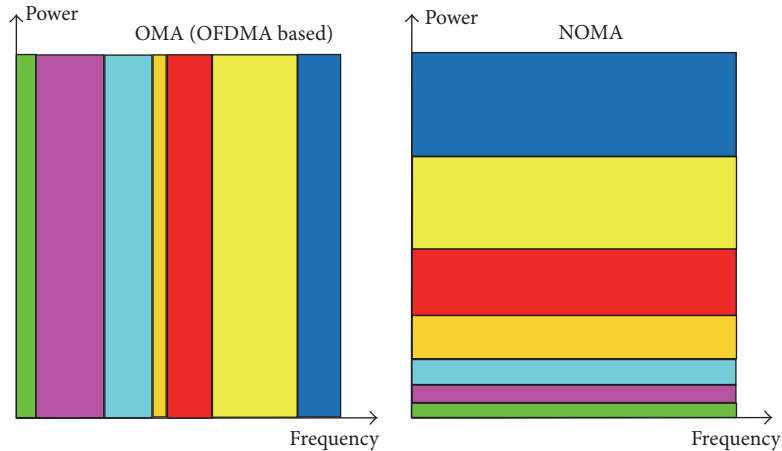


FIGURE 1: A pictorial comparison of OMA and NOMA.

within 3GPP LTE Release 15 [31]. Agreed objectives in Release 15 can be summarized as follows: (1) transmitter side signal processing schemes for NOMA, such as modulation and symbol level processing, coded bit level processing, and symbol to resource element mapping; (2) receivers for NOMA, such as minimum mean-square error (MMSE) receiver, SIC and/or parallel interference cancellation (PIC) receiver, joint detection type receivers, and complexity of the receivers; (3) NOMA procedures, such as uplink transmission detection, link adaptation MA, synchronous and asynchronous operation, and adaptation between OMA and NOMA; (4) link and system level performance evaluation or analysis for NOMA, such as traffic model and deployment scenarios of eMBB, mMTC and URLLC, coverage, latency, and signaling overhead.

In other words, the insufficient performance of OMA makes it inapplicable and unsuitable to provide the features needed to be met by the future generations of wireless communication systems. Consequently, researchers suggest NOMA as a strong candidate as an MA technique for next generations [32]. Although NOMA has many features that may support next generations, it has some limitations that should be addressed in order to exploit its full advantage set. Those limitations can be pointed out as follows. In NOMA, since each user requires to decode the signals of some users before decoding its own signal, the receiver computational complexity will be increased when compared to OMA, leading to a longer delay. Moreover, information of channel gains of all users should be fed back to the base station (BS), but this results in a significant channel state information (CSI) feedback overhead. Furthermore, if any errors occur during SIC processes at any user, then the error probability of successive decoding will be increased. As a result, the number of users should be reduced to avoid such error propagation. Another reason for restricting the number of users is that considerable channel gain differences among

users with different channel conditions are needed to have a better network performance.

This paper, written in a tutorial name, focuses on NOMA technique, along with its usage in MIMO and cooperative scenarios. Practice implementation aspects are also detailed. Besides, an overview about the standardizations of NOMA in 3GPP LTE and application in the 5G scenarios is provided. In addition, unlike previous studies, this paper includes performance analyses of MIMO-NOMA and cooperative NOMA scenarios to make the NOMA concept more understandable by researchers. The remainder of this paper is organized as follows. Basic concepts of NOMA, in both downlink and uplink networks, are given in Section 2. In Sections 3 and 4, MIMO-NOMA and cooperative NOMA are described, respectively. Practical implementation challenges of NOMA are detailed in Section 5. The paper is concluded in Section 6.

2. Basic Concepts of NOMA

In this section, an overview of NOMA in downlink and uplink networks is introduced through signal-to-interference-and-noise ratio (SINR) and sum rate analyses. Then, high signal-to-noise ratio (SNR) analysis has been conducted in order to compare the performances of OMA and NOMA techniques.

2.1. Downlink NOMA Network. At the transmitter side of downlink NOMA network, as shown in Figure 2, the BS transmits the combined signal, which is a superposition of the desired signals of multiple users with different allocated power coefficients, to all mobile users. At the receiver of each user, SIC process is assumed to be performed successively until user's signal is recovered. Power coefficients of users are allocated according to their channel conditions, in an inversely proportional manner. The user with a bad channel condition is allocated higher transmission power than the one which has a good channel condition. Thus, since the user with the highest transmission power considers the signals of other users as noise, it recovers its signal immediately

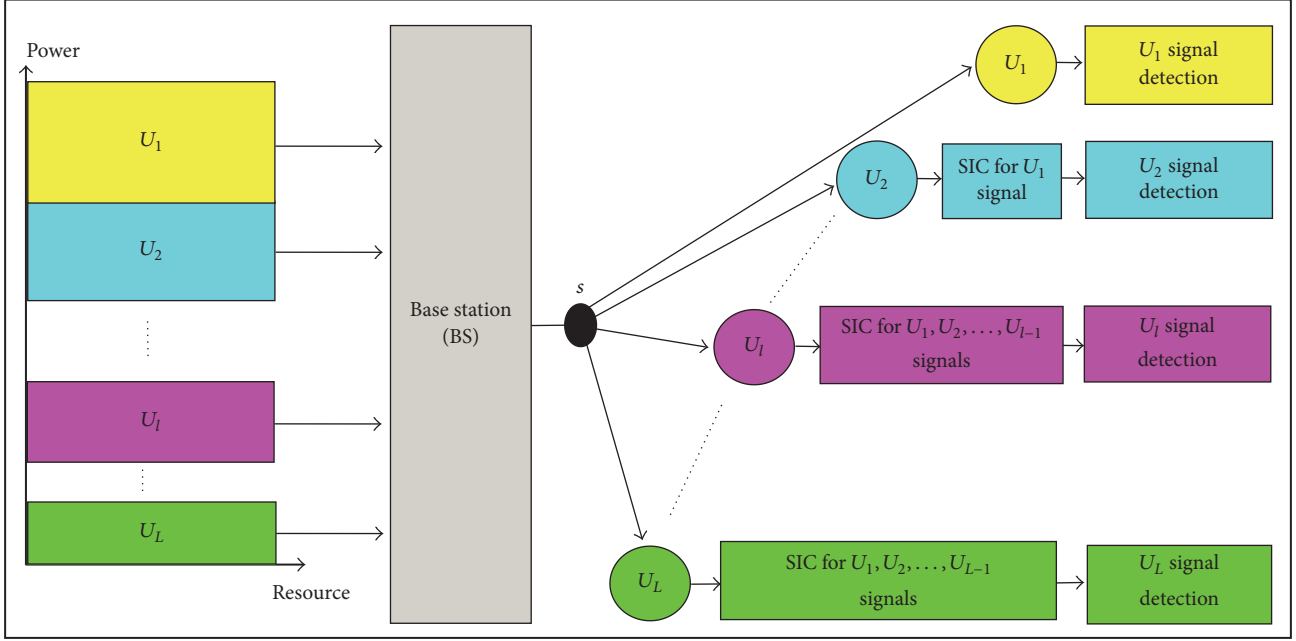


FIGURE 2: Downlink NOMA network.

without performing any SIC process. However, other users need to perform SIC processes. In SIC, each user's receiver first detects the signals that are stronger than its own desired signal. Next, those signals are subtracted from the received signal and this process continues until the related user's own signal is determined. Finally, each user decodes its own signal by treating other users with lower power coefficients as noise. The transmitted signal at the BS can be written as follows:

$$s = \sum_{i=1}^L \sqrt{a_i P_s} x_i, \quad (1)$$

where x_i is the information of user i (U_i) with unit energy. P_s is the transmission power at the BS and a_i is the power coefficient allocated for user i subjected to $\sum_{i=1}^L a_i = 1$ and $a_1 \geq a_2 \geq \dots \geq a_L$ since without loss of generality the channel gains are assumed to be ordered as $|h_1|^2 \leq |h_2|^2 \leq \dots \leq |h_L|^2$, where h_l is the channel coefficient of l th user, based on NOMA concept. The received signal at l th user can be expressed as follows:

$$y_l = h_l s + n_l = h_l \sum_{i=1}^L \sqrt{a_i P_s} x_i + n_l, \quad (2)$$

where n_l is zero mean complex additive Gaussian noise with a variance of σ^2 ; that is, $n_l \sim \text{CN}(0, \sigma^2)$.

2.1.1. SINR Analysis. By using (2), the instantaneous SINR of the l th user to detect the j th user, $j \leq l$, with $j \neq L$ can be written as follows:

$$\text{SINR}_{j \rightarrow l} = \frac{a_j \gamma |h_l|^2}{\gamma |h_l|^2 \sum_{i=j+1}^L a_i + 1}, \quad (3)$$

where $\gamma = P_s / \sigma^2$ denotes the SNR. In order to find the desired information of the l th user, SIC processes will be implemented for the signal of user $j \leq l$. Thus, the SINR of l th user can be given by

$$\text{SINR}_l = \frac{a_l \gamma |h_l|^2}{\gamma |h_l|^2 \sum_{i=l+1}^L a_i + 1}. \quad (4)$$

Then, the SINR of the L th user is expressed as

$$\text{SINR}_L = a_L \gamma |h_L|^2. \quad (5)$$

2.1.2. Sum Rate Analysis. After finding the SINR expressions of downlink NOMA, the sum rate analysis can easily be done. The downlink NOMA achievable data rate of l th user can be expressed as

$$\begin{aligned} R_l^{\text{NOMA-d}} &= \log_2(1 + \text{SINR}_l) \\ &= \log_2 \left(1 + \frac{a_l \gamma |h_l|^2}{\gamma |h_l|^2 \sum_{i=l+1}^L a_i + 1} \right). \end{aligned} \quad (6)$$

Therefore, the sum rate of downlink NOMA can be written as

$$\begin{aligned} R_{\text{sum}}^{\text{NOMA-d}} &= \sum_{l=1}^L \log_2(1 + \text{SINR}_l) \\ &= \sum_{l=1}^{L-1} \log_2 \left(1 + \frac{a_l \gamma |h_l|^2}{\gamma |h_l|^2 \sum_{i=l+1}^L a_i + 1} \right) \\ &\quad + \log_2(1 + a_L \gamma |h_L|^2) \end{aligned}$$

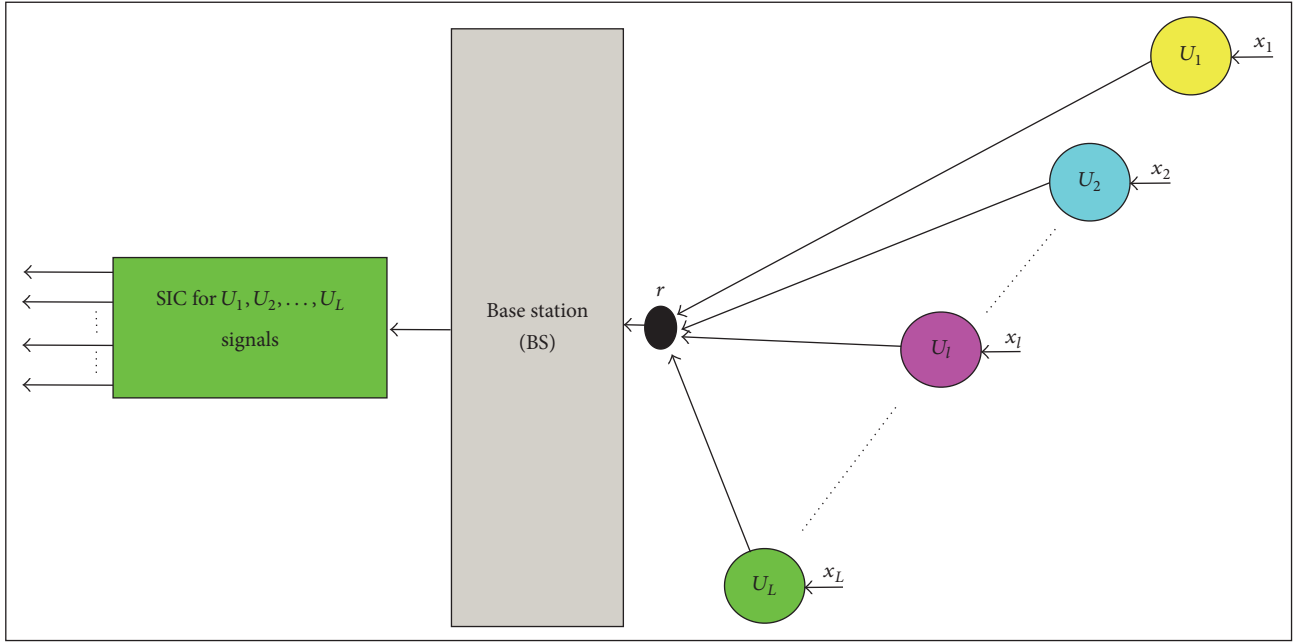


FIGURE 3: Uplink NOMA network.

$$\begin{aligned}
 &= \sum_{l=1}^{L-1} \log_2 \left(1 + \frac{a_l}{\sum_{i=l+1}^L a_i + 1/\gamma |h_l|^2} \right) \\
 &\quad + \log_2 (1 + a_L \gamma |h_L|^2).
 \end{aligned} \tag{7}$$

In order to figure out whether NOMA techniques outperform OMA techniques, we conduct a high SNR analysis. Thus, at high SNR, that is, $\gamma \rightarrow \infty$, the sum rate of downlink NOMA becomes

$$\begin{aligned}
 R_{\text{sum}}^{\text{NOMA-d}} &\approx \sum_{l=1}^{L-1} \log_2 \left(1 + \frac{a_l}{\sum_{i=l+1}^L a_i} \right) + \log_2 (\gamma |h_L|^2) \\
 &\approx \log_2 (\gamma |h_L|^2).
 \end{aligned} \tag{8}$$

2.2. Uplink NOMA Network. In uplink NOMA network, as depicted in Figure 3, each mobile user transmits its signal to the BS. At the BS, SIC iterations are carried out in order to detect the signals of mobile users. By assuming that downlink and uplink channels are reciprocal and the BS transmits power allocation coefficients to mobile users, the received signal at the BS for synchronous uplink NOMA can be expressed as

$$r = \sum_{i=1}^L h_i \sqrt{a_i P} x_i + n, \tag{9}$$

where h_i is the channel coefficient of the i th user, P is the maximum transmission power assumed to be common for all users, and n is zero mean complex additive Gaussian noise with a variance of σ^2 ; that is, $n \sim \text{CN}(0, \sigma^2)$.

2.2.1. SINR Analysis. The BS decodes the signals of users orderly according to power coefficients of users, and then the SINR for l th user $l \neq 1$ can be given by [33]

$$\text{SINR}_l = \frac{a_l \gamma |h_l|^2}{\gamma \sum_{i=1}^{l-1} a_i |h_i|^2 + 1}, \tag{10}$$

where $\gamma = P/\sigma^2$. Next, the SINR for the first user is expressed as

$$\text{SINR}_1 = a_1 \gamma |h_1|^2. \tag{11}$$

2.2.2. Sum Rate Analysis. The sum rate of uplink NOMA can be written as

$$\begin{aligned}
 R_{\text{sum}}^{\text{NOMA-u}} &= \sum_{l=1}^L \log_2 (1 + \text{SINR}_l) \\
 &= \log_2 (1 + a_1 \gamma |h_1|^2) \\
 &\quad + \sum_{l=2}^L \log_2 \left(1 + \frac{a_l \gamma |h_l|^2}{\gamma \sum_{i=1}^{l-1} a_i |h_i|^2 + 1} \right) \\
 &= \log_2 \left(1 + \gamma \sum_{l=1}^L a_l |h_l|^2 \right).
 \end{aligned} \tag{12}$$

When $\gamma \rightarrow \infty$, the sum rate of uplink NOMA becomes

$$R_{\text{sum}}^{\text{NOMA-u}} \approx \log_2 \left(\gamma \sum_{l=1}^L a_l |h_l|^2 \right). \tag{13}$$

2.3. *Comparing NOMA and OMA.* The achievable data rate of the l th user of OMA for both uplink and downlink can be expressed as [33]

$$R_l^{\text{OMA}} = \alpha_l \log_2 \left(1 + \frac{\beta_l \gamma |h_l|^2}{\alpha_l} \right), \quad (14)$$

where β_l and α_l are the power coefficient and the parameter related to the specific resource of U_l , respectively. And then, the sum rate of OMA is written as

$$R_{\text{sum}}^{\text{OMA}} = \sum_{l=1}^L \alpha_l \log_2 \left(1 + \frac{\beta_l \gamma |h_l|^2}{\alpha_l} \right). \quad (15)$$

For OMA, for example, FDMA, total bandwidth resource and power are shared among the users equally; then using $\alpha_l = \beta_l = 1/L$ the sum rate can be written as

$$R_{\text{sum}}^{\text{OMA}} = \sum_{l=1}^L \frac{1}{L} \log_2 (1 + \gamma |h_l|^2). \quad (16)$$

When $\gamma \rightarrow \infty$, the sum rate of OMA becomes

$$R_{\text{sum}}^{\text{OMA}} \approx \sum_{l=1}^L \frac{1}{L} \log_2 (\gamma |h_l|^2). \quad (17)$$

Using $|h_1|^2 \leq |h_2|^2 \leq \dots \leq |h_L|^2$,

$$\begin{aligned} R_{\text{sum}}^{\text{OMA}} &\approx \sum_{l=1}^L \frac{1}{L} \log_2 (\gamma |h_l|^2) \leq \sum_{l=1}^L \frac{1}{L} \log_2 (\gamma |h_L|^2) \\ &= \log_2 (\gamma |h_L|^2) \approx R_{\text{sum}}^{\text{NOMA-d}}. \end{aligned} \quad (18)$$

Hence, we conclude $R_{\text{sum}}^{\text{OMA}} \leq R_{\text{sum}}^{\text{NOMA-d}}$.

For the sake of simplicity, sum rates of uplink NOMA and OMA can be compared for two users. Then, using (13) and (17) the sum rate of uplink NOMA and OMA at high SNR can be expressed, respectively, as

$$R_{\text{sum}}^{\text{NOMA-u}} \approx \log_2 (\gamma |h_1|^2 + \gamma |h_2|^2), \quad (19)$$

$$\begin{aligned} R_{\text{sum}}^{\text{OMA}} &\approx \frac{1}{2} \log_2 (\gamma |h_1|^2) + \frac{1}{2} \log_2 (\gamma |h_2|^2) \\ &\leq \log_2 (\gamma |h_2|^2). \end{aligned} \quad (20)$$

From (19) and (20), we notice $R_{\text{sum}}^{\text{OMA}} \leq R_{\text{sum}}^{\text{NOMA-u}}$.

Figure 4 shows that NOMA outperforms OMA in terms of sum rate in both downlink and uplink of two user networks using (7), (12), and (16).

3. MIMO-NOMA

MIMO technologies have a significant capability of increasing capacity as well as improving error probability of wireless communication systems [34]. To take advantage of MIMO schemes, researchers have investigated the performance of NOMA over MIMO networks [35]. Many works have been

studying the superiority of MIMO-NOMA over MIMO-OMA in terms of sum rate and ergodic sum rate under different conditions and several constrictions [36–39]. Specifically, in [36], the maximization problem of ergodic sum rate for two-user MIMO-NOMA system over Rayleigh fading channels is discussed. With the need of partial CSI at the BS and under some limitations on both total transmission power and the minimum rate for the user with bad channel condition, the optimal power allocation algorithm with a lower complexity to maximize the ergodic capacity is proposed. However, in order to achieve a balance between the maximum number of mobile users and the optimal achievable sum rate in MIMO-NOMA systems, sum rate has been represented through two ways. The first approach targets the optimization of power partition among the user clusters [37]. Another approach is to group the users in different clusters such that each cluster can be allocated with orthogonal spectrum resources according to the selected user grouping algorithm [38]. Furthermore, in [37] performances of two users per cluster schemes have been studied for both MIMO-NOMA and MIMO-OMA over Rayleigh fading channels. In addition, in accordance with specified power split, the dominance of NOMA over OMA has been shown in terms of sum channel and ergodic capacities.

On the other side, the authors in [38] have examined the performance of MIMO-NOMA system, in which multiple users are arranged into a cluster. An analytical comparison has been provided between MIMO-NOMA and MIMO-OMA, and then it is shown that NOMA outperforms OMA in terms of sum channel and ergodic capacities in case of multiple antennas. Moreover, since the number of users per cluster is inversely proportional to the achievable sum rate and the trade-off between the number of admitted users and achieved sum rate has to be taken into account (which restricts the system performance), a user admission scheme, which maximizes the number of users per cluster based on their SINR thresholds, is proposed. Although the optimum performance is achieved in terms of the number of admitted users and the sum rate when the SINR thresholds of all users are equal, even when they are different good results are obtained. In addition, a low complexity of the proposed scheme is linearly proportional to the number of users per cluster. In [39], the performance of downlink MIMO-NOMA network for a simple case of two users, that is, one cluster, is introduced. In this case, MIMO-NOMA provides a better performance than MIMO-OMA in terms of both the sum rate and ergodic sum rate. Also, it is shown that for a more practical case of multiple users, with two users allocated into a cluster and sharing the same transmit beamforming vector, where ZF precoding and signal alignment are employed at the BS and the users of the same cluster, respectively, the same result still holds.

Antenna selection techniques have also been recognized as a powerful solution that can be applied to MIMO systems in order to avoid the adverse effects of using multiple antennas simultaneously. These effects include hardware complexity, redundant power consumption, and high cost. Meanwhile diversity advantages that can be achieved from MIMO systems are still maintained [40]. Several works apply

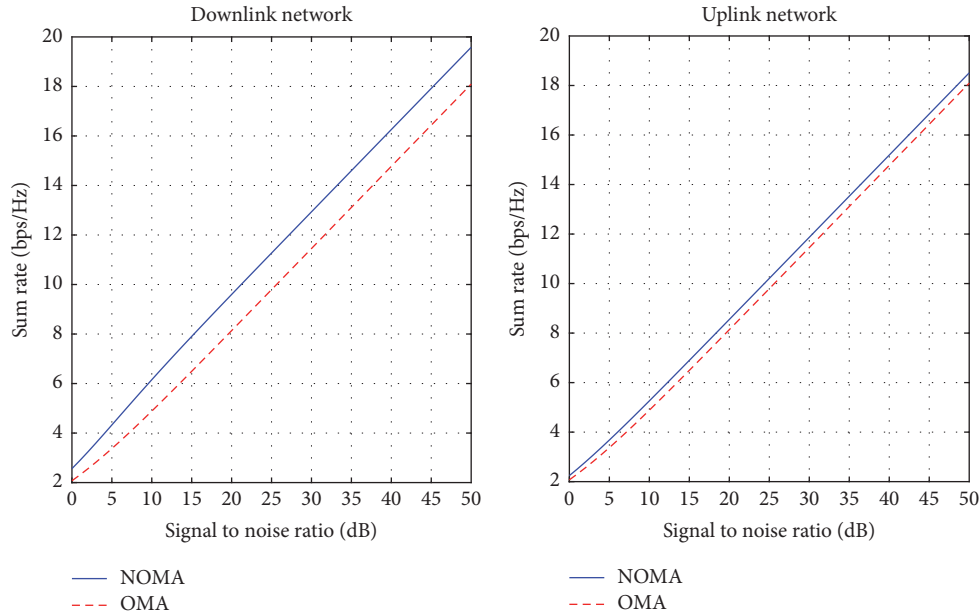


FIGURE 4: Sum rate of NOMA and OMA in both downlink and uplink networks with $a_1 = 0.6$, $a_2 = 0.4$, $|h_1|^2 = 0$ dB, and $|h_2|^2 = 20$ dB.

antenna selection techniques in MIMO-NOMA as they have already been developed for MIMO-OMA systems. But the gains can not be easily replicated since there is a heavy interuser interference in MIMO-NOMA networks, dissimilar from those in MIMO-OMA networks, in which information is transmitted in an interference-free manner. Consequently, there are a few works that challenged the antenna selection problem [41–43]. In [41], the sum rate performance for downlink multiple input single output- (MISO-) NOMA system is investigated with the help of transmit antenna selection (TAS) at the BS, where the transmitter of the BS and the receiver of each mobile user are equipped with multiantenna and single antenna, respectively. Basically, in TAS-OMA scheme, the best antenna at the BS offering the highest SINR is selected. However in the proposed TAS-NOMA scheme in [41], the best antenna at the BS providing the maximum sum rate is chosen. In addition to using an efficient TAS scheme, user scheduling algorithm is applied in two user massive MIMO-NOMA system in order to maximize the achievable sum rate in [42] for two scenarios, namely, the single-band two users and the multiband multiuser. In the first scenario, an efficient search algorithm is suggested. This algorithm aims to choose the antennas providing the highest channel gains in such a way that the desired antennas are only searched from specified finite candidate set, which are useful to the concerned users. On the other hand, in the second scenario, a joint user and antenna contribution algorithm is proposed. In particular, this algorithm manipulates the ratio of channel gain specified by a certain antenna-user pair to the total channel gain, and hence antenna-user pair offering the highest contribution to the total channel gain is selected. Moreover, an efficient search algorithm provides a better trade-off between system performance and complexity, rather than a joint antenna and user contribution algorithm. Unfortunately, neither the authors of [41] nor the authors

of [42] have studied the system performance analytically. In [43], the maximization of the average sum rate of two-user NOMA system, in which the BS and mobile users are equipped with multiantenna, is discussed through two computationally effective joint antenna selection algorithms; the max-min-max and the max-max-max algorithms. However, the instantaneous channel gain of the user with a bad channel condition is improved in max-min-max antenna selection scheme while max-max-max algorithm is the solution for the user with a good channel condition. Furthermore, asymptotic closed-form expressions of the average sum rates are evaluated for both proposed algorithms. Moreover, it is verified that better user fairness can be achieved by the max-min-max algorithm while larger sum rate can be obtained by the max-max-max algorithm.

Multicast beamforming can also be introduced as a technique that can be employed in MIMO schemes since it offers a better sum capacity performance even for multiple users. However, it can be applied in different ways. One approach is based on a single beam that can be used by all users; hence all users receive this common signal [44]. Another approach is to use multiple beams that can be utilized by many groups of users; that is, each group receives a different signal [45]. The following works have studied beamforming in MIMO-NOMA systems. In [46], multiuser beamforming in downlink MIMO-NOMA system is proposed. Particularly, a pair of users can share the same beam. Since the proposed beam can be only shared by two users with different channel qualities, it is probable to easily apply clustering and power allocation algorithms to maximize the sum capacity and to decrease the intercluster and interuser interferences. In [47], performance of multicast beamforming, when the beam is used to serve many users per cluster by sharing a common signal, is investigated with superposition coding for a downlink MISO-NOMA network in a simple scenario of two users.

Principally, the transmitter of the BS has multiantenna and its information stream is based on multiresolution broadcast concept, in which only low priority signal is sent to the user that is far away from the BS, that is, user with a bad channel quality. Both signals of high priority and low priority are transmitted to the user near to BS, that is, user with good channel quality. Furthermore, with superposition coding a minimum power beamforming problem has been developed in order to find the beamforming vectors and the powers for both users. Moreover, under the considered optimization condition and the given normalized beamforming vectors (which are founded by an iterative algorithm), the closed-form expression for optimal power allocation is easily obtained. In [48], random beamforming is carried out at the BS of a downlink MIMO-NOMA network. In the system model, each beam is assumed to be used by all the users in one cluster and all beams have similar transmission power allocations. Moreover, a spatial filter is suggested to be used in order to diminish the intercluster and interbeam interferences. Fractional frequency reuse concept, in which users with different channel conditions can accommodate many reuse factors, is proposed in order to improve the power allocation among multiple beams. In [49], interference minimization and capacity maximization for downlink multiuser MIMO-NOMA system are introduced, in which the number of receive antennas of mobile user is larger than the number of transmit antennas of the BS. Zero-forcing beamforming technique is suggested to reduce the intercluster interference, especially when distinctive channel quality users is assumed. In addition, dynamic power allocation and user-cluster algorithms have been proposed not only to achieve maximum throughput, but also to minimize the interference.

There are many research works investigating resource allocation problem in terms of maximization of the sum rate in case of perfect CSI [50–52]. Specifically, in [50] sum rate optimization problem of two-user MIMO-NOMA network, that is, two users in one cluster in which different precoders are implemented, has been introduced under the constraint of transmission power at the BS and the minimum transmission rate limitation of the user with bad channel condition. In [51], the sum rate maximization problem for downlink MISO-NOMA system is investigated. However, the transmitted signal for each mobile user is weighted with a complex vector. Moreover, for the sake of avoiding the high computational complexity related to nonconvex optimization problem, minorization-maximization method is suggested as an approximation. The key idea of minorization-maximization algorithm is to design the complex weighting vectors in such a way that the total throughput of the system is maximized, for a given order of users; that is, perfect CSI is assumed. In [52], a downlink MIMO-NOMA system, where perfect CSI available at all nodes is assumed and with different beams, BS broadcasts precoded signals to all mobile users; that is, each beam serves several users. However, there are three proposed algorithms combined in order to maximize the sum rate. The first one is where weighted sum rate maximization proposes to design a special beamforming matrix of each beam benefiting from all CSI at the BS. The second algorithm is where user scheduling

aims to have super SIC at the receiver of each mobile user. Thus, to take full benefits of SIC, differences in channel gains per cluster should be significant and the channel correlation between mobile users has to be large. The final one is where fixed power allocation targets optimization, offering not only a higher sum rate, but also convenient performance for the user with bad channel quality. In [53], the optimal power allocation method, in order to maximize the sum rate of two-user MIMO-NOMA with a layered transmission scheme under a maximum transmission power constraint for each mobile user, is investigated. Basically, by using the layered transmission, each mobile user performs sequence by sequence decoding signals throughout SIC, yielding much lower decoding complexity when compared to the case with nonlayered transmission. Moreover, the closed-form expression for the average sum rate and its bounds in both cases of perfect CSI and partial CSI are obtained. Also, it is shown that the average sum rate is linearly proportional to the number of antennas. In [54], a comprehensive resource allocation method for multiuser downlink MIMO-NOMA system including beamforming and user selection is proposed, yielding low computational complexity and high performance in cases of full and partial CSI. However, resource allocation has been expressed in terms of the maximum sum rate and the minimum of maximum outage probability (OP) for full CSI and partial CSI, respectively. Outage behavior for both downlink and uplink networks in MIMO-NOMA framework with integrated alignment principles is investigated in a single cell [55] and multicell [56, 57], respectively. Furthermore, an appropriate trade-off between fairness and throughput has been achieved by applying two strategies of power allocation methods. The fixed power allocation strategy realizes different QoS requirements. On the other hand cognitive radio inspired power allocation strategy verifies that QoS requirements of the user are achieved immediately. In addition, exact and asymptotic expressions of the system OP have been derived. In [58], the power minimization problem for downlink MIMO-NOMA networks under full CSI and channel distribution information scenarios are studied. In [59], linear beamformers, that is, precoders that provide a larger total sum throughput also improving throughput of the user with bad quality channel, are designed; meanwhile QoS specification requirements are satisfied. Also, it is shown that the maximum number of users per cluster that realizes a higher NOMA performance is achieved at larger distinctive channel gains.

Moreover, since massive MIMO technologies can ensure bountiful antenna diversity at a lower cost [4], many works have discussed performance of NOMA over massive MIMO. For instance, in [60], massive MIMO-NOMA system, where the number of the transmit antennas at the BS is significantly larger than the number of users, is studied with limited feedback. Also, the exact expressions of the OP and the diversity order are obtained for the scenarios of perfect order of users and one bit feedback, respectively. In [61], the scheme based on interleave division multiple access and iterative data-aided channel estimation is presented in order to solve the reliability problem of multiuser massive MIMO-NOMA system with

imperfect CSI available at the BS. In [62], the achievable rate in massive MIMO-NOMA systems and iterative data-aided channel estimation receiver, in which partially decoded information is required to get a better channel estimation, are investigated through applying two pilot schemes: orthogonal pilot and superimposed pilot. However, pilots in the orthogonal pilot scheme occupy time/frequency slots while they are superimposed with information in superimposed pilot one. Moreover, it is shown that the greatest part of pilot power in superimposed pilot scheme seems to be zero in the case when Gaussian signal prohibits overhead power and rate loss that may be resulted through using pilot. Consequently, with code maximization superimposed scheme has a superior performance over orthogonal one under higher mobility and larger number of mobile users. Different from massive MIMO, in [63] performance of massive access MIMO systems, in which number of users is larger than the number of antennas employed at the BS, is studied. Low-complexity Gaussian message specially passing iterative detection algorithm is used and both its mean and variance precisely converge with high speed to those concerned with the minimum mean square error multiuser detection in [64].

In addition, NOMA has been proposed as a candidate MA scheme integrated with beamspace MIMO in mmWave communication systems, satisfying massive connectivity, where the number of mobile users is much greater than the number of radio frequency chains, and obtaining a better performance in terms of spectrum and energy efficiency [65]. Furthermore, a precoding scheme designed on zero-forcing (ZF) concept has been suggested in order to reduce the interbeam interference. Moreover, iterative optimization algorithm with dynamic power allocation scheme is proposed to obtain a higher sum rate and lower complexity. In [66], the optimization problem of energy efficiency for MIMO-NOMA systems with imperfect CSI at the BS over Rayleigh fading channels is studied under specified limitations on total transmission power and minimum sum rate of the user of bad channel condition. However, two-user scheduling schemes and power allocation scheme are presented in [67] in order to maximize the energy efficiency. The user scheduling schemes depend on the signal space alignment; while one of them effectively deals with the multiple interference, the other one maximizes the multicollinearity among users. On the other hand, power allocation scheme uses a sequential convex approximation that roughly equalizes the nonconvex problem by a set of convex problems iteratively, that is, in each iteration nonconvex constraints are modified into their approximations in inner convex. Also, it is shown that higher energy efficiency is obtained when lower power is transmitted and a higher sum rate of center users is obtained when maximum multicollinearity scheme is employed.

Many other problems have been investigated in MIMO-NOMA systems. For example, in [68, 69], QoS optimization problem is proposed for two-user MISO-NOMA system. In particular, closed-form expressions of optimal precoding vectors over flat fading channels, are achieved by applying the Lagrange duality and an iterative method in [68] and [69], respectively.

As mentioned before, NOMA promises to satisfy the need of IoT, in which many users require to be served rapidly for small packet transmissions. Consequently, the literature tends to study performance of MIMO-NOMA for IoT. For instance, in [70] a MIMO-NOMA downlink network where one transmitter sending information to two users is considered. However, one user has a low data rate, that is, small packet transmission, while the second user has a higher rate. Particularly, outage performance in case of using precoding and power allocation method is investigated. Also, it is shown that the potential of NOMA is apparent even when channel qualities of users are similar.

Most current works of MIMO-NOMA focus on sum rate and capacity optimization problems. However, performance of symbol error rate (SER) for wireless communication systems is also very substantial. In [71], SER performance using the minimum Euclidean distance precoding scheme in MIMO-NOMA networks is studied. For simple transmission case, two-user 2×2 MIMO-NOMA is investigated. However, to facilitate realization of practical case of multiuser MIMO-NOMA network, two-user pairing algorithms are applied.

In order to demonstrate the significant performance of MIMO-NOMA systems in terms of both OP and sum rate, as well as its superiority over MIMO-OMA, a special case, performance of single input multiple output- (SIMO-) NOMA network based on maximal ratio combining (MRC) diversity technique in terms of both OP and ergodic sum rate is investigated in the following section. Moreover, closed-form expression of OP and bounds of ergodic sum rate are derived.

3.1. Performance Analysis of SIMO-NOMA. This network includes a BS and L mobile users as shown in Figure 5. The transmitter of BS is equipped with a single antenna and the receiver of each mobile user is equipped with N_r antennas. The received signal at the l th user after applying MRC can be written as follows:

$$r_l = \|\mathbf{h}_l\| \sum_{i=1}^L \sqrt{a_i P_s} x_i + \frac{\mathbf{h}_l^H}{\|\mathbf{h}_l\|} \mathbf{n}_l, \quad (21)$$

where \mathbf{h}_l is $N_r \times 1$ fading channel coefficient vector between the BS and l th user and without loss of generality and due to NOMA concept they are sorted in ascending way; that is, $\|\mathbf{h}_1\|^2 \leq \|\mathbf{h}_2\|^2 \leq \dots \leq \|\mathbf{h}_L\|^2$, and \mathbf{n}_l is $N_r \times 1$ zero mean complex additive Gaussian noise with $E[\mathbf{n}_l \mathbf{n}_l^H] = \mathbf{I}_{N_r} \sigma_l^2$ at the l th user, where $E[\cdot]$, $(\cdot)^H$, and \mathbf{I}_r denote the expectation operator, Hermitian transpose, and identity matrix of order r , respectively, and $\sigma_l^2 = \sigma^2$ is the variance of \mathbf{n}_l per dimension. From (21), instantaneous SINR for l th user to detect j th user, $j \leq l$, with $j \neq L$ can be expressed as follows:

$$\text{SINR}_{j \rightarrow l} = \frac{a_j \gamma \|\mathbf{h}_l\|^2}{\gamma \|\mathbf{h}_l\|^2 \sum_{i=j+1}^L a_i + 1}. \quad (22)$$

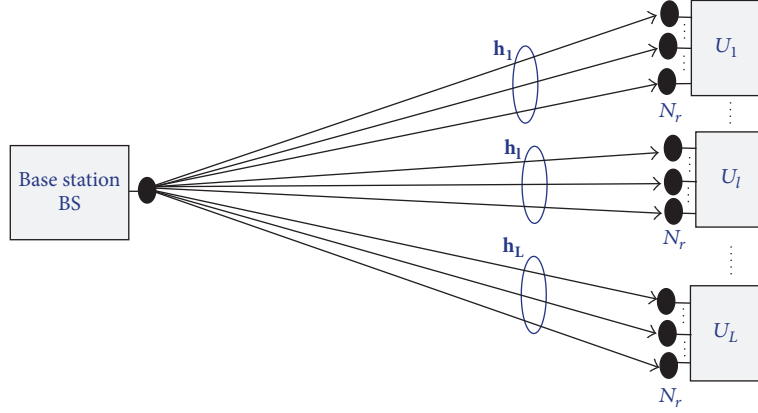


FIGURE 5: System model of the downlink SIMO-NOMA.

Now, nonordered channel gains for MRC can be given as follows:

$$\|\tilde{\mathbf{h}}_l\|^2 = \sum_{i=1}^{N_r} |h_{l,i}|^2, \quad l = 1, 2, \dots, L, \quad (23)$$

where $h_{l,i}$ denotes the channel coefficient between the BS and i th antenna of the l th user and are independent and identically distributed (i.i.d.) Nakagami- m random variables. By the help of the series expansion of incomplete Gamma function [72, eq. (8.352.6)], the cumulative distribution function (CDF) and probability density function (PDF) of Gamma random variable X , square of Nakagami- m random variable can be defined as follows:

$$F_X(x) = \frac{\gamma(m, mx/\Omega)}{\Gamma(m)} = 1 - e^{-mx/\Omega} \sum_{k=0}^{m-1} \left(\frac{mx}{\Omega}\right)^k \frac{1}{k!}, \quad (24)$$

$$f_X(x) = \left(\frac{m}{\Omega}\right)^m \frac{x^{m-1}}{\Gamma(m)} e^{-mx/\Omega},$$

where $\gamma(\cdot, \cdot)$ and $\Gamma(\cdot)$ are the lower incomplete Gamma function given by [72, eq. (8.350.1)] and the Gamma function given by [72, eq. (8.310.1)], respectively. m is parameter of Nakagami- m distribution, and $\Omega = E[|X|^2]$. With the help of the highest order statistics [73], we can write CDF of nonordered $\|\tilde{\mathbf{h}}_l\|^2$ as follows:

$$F_{\|\tilde{\mathbf{h}}_l\|^2}(x) = \frac{\gamma(mN_r, mx/\Omega)}{\Gamma(mN_r)}$$

$$= 1 - e^{-mx/\Omega} \sum_{s=0}^{mN_r-1} \left(\frac{mx}{\Omega}\right)^s \frac{1}{s!} \quad (25)$$

$$= \sum_{r=0}^1 \sum_{s=0}^{r(mN_r-1)} (-1)^r \vartheta_s(r, mN_r) x^s e^{-rmx/\Omega},$$

where $\Omega = E[\|\tilde{\mathbf{h}}_l\|^2]$ and $\vartheta_a(b, g_c)$ denotes multinomial coefficients which can be defined as [72, eq. (0.314)]

$$\vartheta_a(b, g_c) = \frac{1}{a d_0} \sum_{\rho=1}^a (\rho(b+1) - a) d_\rho \vartheta_{a-\rho}(b, g_c), \quad (26)$$

$$a \geq 1.$$

In (26), $d_\rho = (g_c/\Omega)^\rho/\rho!$, $\vartheta_0(b, g_c) = 1$, and $\vartheta_a(b, g_c) = 0$ if $\rho > g_c - 1$. Next, CDF of the ordered $\|\mathbf{h}_l\|^2$ can be expressed as [74]

$$F_{\|\mathbf{h}_l\|^2}(x) = \frac{L!}{(L-l)!(l-1)!} \sum_{t=0}^{L-l} \frac{(-1)^t}{l+t} \binom{L-l}{t}$$

$$\times [F_{\|\tilde{\mathbf{h}}_l\|^2}(x)]^{l+t} = \frac{L!}{(L-l)!(l-1)!}$$

$$\cdot \sum_{t=0}^{L-l} \sum_{r=0}^{l+t} \sum_{s=0}^{r(mN_r-1)} \frac{(-1)^{t+r}}{l+t}$$

$$\cdot \binom{L-l}{t} \binom{l+t}{r} \vartheta_s(r, mN_r) x^s e^{-rmx/\Omega}.$$

3.1.1. *Outage Probability of SIMO-NOMA.* The OP of the l th user can be obtained as follows:

$$P_{\text{out},l} = \Pr(\text{SINR}_{j \rightarrow l} < \gamma_{\text{th}_j})$$

$$= \Pr\left(\frac{a_j \gamma \|\mathbf{h}_l\|^2}{\gamma \|\mathbf{h}_l\|^2 \sum_{i=l+1}^L a_i + 1} < \gamma_{\text{th}_j}\right)$$

$$= \Pr\left(\|\mathbf{h}_l\|^2 < \frac{\gamma_{\text{th}_j}}{\gamma(a_j - \gamma_{\text{th}_j} \sum_{i=l+1}^L a_i)}\right)$$

$$\begin{aligned}
&= \Pr(\|\mathbf{h}_1\|^2 < \eta_l^*) = F_{\|\mathbf{h}_1\|^2}(\eta_l^*) = \frac{L!}{(L-l)!(l-1)!} \\
&\cdot \sum_{t=0}^{L-l} \sum_{r=0}^{l+t} \sum_{s=0}^{r(mN_r-1)} \frac{(-1)^{t+r}}{l+t} \\
&\cdot \binom{L-l}{t} \binom{l+t}{r} \vartheta_s(r, mN_r) \eta_l^{*s} e^{-rm\eta_l^*/\Omega},
\end{aligned} \tag{28}$$

where $\eta_l^* = \max[\eta_1, \eta_2, \dots, \eta_l]$ with $\eta_j = \gamma_{th_j}/\gamma(a_j - \gamma_{th_j} \sum_{i=l+1}^L a_i)$. γ_{th_j} denotes the threshold SINR of the j th user. Under the condition $a_j > \gamma_{th_j} \sum_{i=j+1}^L a_i$, the l th user can decode the j th user's signal successfully irrespective of the channel SNR.

3.1.2. Ergodic Sum Rate Analysis of SIMO-NOMA. Ergodic sum rate can be expressed as

$$\begin{aligned}
R_{\text{sum}} &= \sum_{l=1}^L E \left[\frac{1}{2} \log_2 (1 + \text{SINR}_l) \right] \\
&= \underbrace{\sum_{l=1}^{L-1} E \left[\frac{1}{2} \log_2 (1 + \text{SINR}_l) \right]}_{R_{\bar{L}}} \\
&\quad + \underbrace{E \left[\frac{1}{2} \log_2 (1 + \text{SINR}_L) \right]}_{R_L}.
\end{aligned} \tag{29}$$

Then, $R_{\bar{L}}$ can be expressed as

$$\begin{aligned}
R_{\bar{L}} &= \sum_{l=1}^{L-1} E \left[\frac{1}{2} \log_2 \left(1 + \frac{a_l \gamma \|\mathbf{h}_1\|^2}{\gamma \|\mathbf{h}_1\|^2 \sum_{i=l+1}^L a_i + 1} \right) \right] \\
&= \sum_{l=1}^{L-1} E \left[\frac{1}{2} \log_2 \left(1 + \frac{a_l}{\sum_{i=l+1}^L a_i + 1/\gamma} \right) \right].
\end{aligned} \tag{30}$$

Due to computational difficulty of calculating the exact expression of the ergodic sum rate, and, for the sake of simplicity, we will apply high SNR analysis in order to find the upper and lower bounds related to ergodic sum rate. Thus, when $\gamma \rightarrow \infty$ in (30), then $R_{\bar{L}}^\infty$ can be given by

$$R_{\bar{L}}^\infty = \frac{1}{2} \sum_{l=1}^{L-1} \log_2 \left(1 + \frac{a_l}{\sum_{i=l+1}^L a_i} \right). \tag{31}$$

Now, by using the identity $\int_0^\infty \ln(1+ay)f(y)dy = a \int_0^\infty ((1-F(y))/(1+ay))dy$, $\log_b a = \ln a / \ln b$, R_L can be written as

$$\begin{aligned}
R_L &= E \left[\frac{1}{2} \log_2 (1 + a_L \gamma \|\mathbf{h}_L\|^2) \right] \\
&= \frac{1}{2 \ln 2} E \left[\ln (1 + a_L \gamma \|\mathbf{h}_L\|^2) \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2 \ln 2} \int_0^\infty \ln(1 + a_L \gamma x) f_{\|\mathbf{h}_L\|^2}(x) dx \\
&= \frac{a_L \gamma}{2 \ln 2} \int_0^\infty \frac{1 - F_{\|\mathbf{h}_L\|^2}(x)}{1 + a_L \gamma x} dx,
\end{aligned} \tag{32}$$

Simply, by using (27) $F_{\|\mathbf{h}_L\|^2}$ can be expressed as

$$\begin{aligned}
F_{\|\mathbf{h}_L\|^2}(x) &= 1 \\
&\quad + \sum_{k=1}^L \sum_{n=0}^{k(mN_r-1)} \binom{L}{k} (-1)^k \vartheta_n(k, mN_r) x^n e^{-kmx/\Omega}.
\end{aligned} \tag{33}$$

By substituting (33) into (32),

$$\begin{aligned}
R_L &= \frac{a_L \gamma}{2 \ln 2} \sum_{k=1}^L \sum_{n=0}^{k(mN_r-1)} \binom{L}{k} (-1)^{k+1} \vartheta_n(k, mN_r) \\
&\quad \cdot \underbrace{\int_0^\infty \frac{x^n e^{-kmx/\Omega}}{1 + a_L \gamma x} dx}_I.
\end{aligned} \tag{34}$$

By defining $u = a_L \gamma x$, I can be written as follows:

$$I = \frac{1}{(a_L \gamma)^{n+1}} \int_0^\infty \frac{u^n e^{-kmu/a_L \gamma \Omega}}{1 + u} du. \tag{35}$$

Using [74, (eq. 11)], as $\gamma \rightarrow \infty$, then I can be approximated as

$$I \approx \xi = \begin{cases} \frac{\ln(a_L \gamma \Omega / mk)}{a_L \gamma}, & n = 0 \\ \frac{\Gamma(n) (\Omega / mk)^n}{a_L \gamma}, & n > 0. \end{cases} \tag{36}$$

By substituting (36) into (34), then R_L^∞ can be given by

$$R_L^\infty = \frac{a_L \gamma}{2 \ln 2} \sum_{k=1}^L \sum_{n=0}^{k(mN_r-1)} \binom{L}{k} (-1)^{k+1} \vartheta_n(k, mN_r) \xi. \tag{37}$$

Finally, by substituting (37) and (31) into (29), then asymptotic ergodic sum rate R_{sum}^∞ can be expressed as

$$\begin{aligned}
R_{\text{sum}}^\infty &= \frac{1}{2} \sum_{l=1}^{L-1} \log_2 \left(1 + \frac{a_l}{\sum_{i=l+1}^L a_i} \right) \\
&\quad + \frac{a_L \gamma}{2 \ln 2} \sum_{k=1}^L \sum_{n=0}^{k(mN_r-1)} \binom{L}{k} (-1)^{k+1} \vartheta_n(k, mN_r) \xi.
\end{aligned} \tag{38}$$

3.1.3. Numerical Results of SIMO-NOMA. We consider two users and their average power factors that provide $\sum_{i=1}^L a_i = 1$ are selected as $a_1 = 0.6$ and $a_2 = 0.4$, respectively. Also, in order to make a comparison between the performances

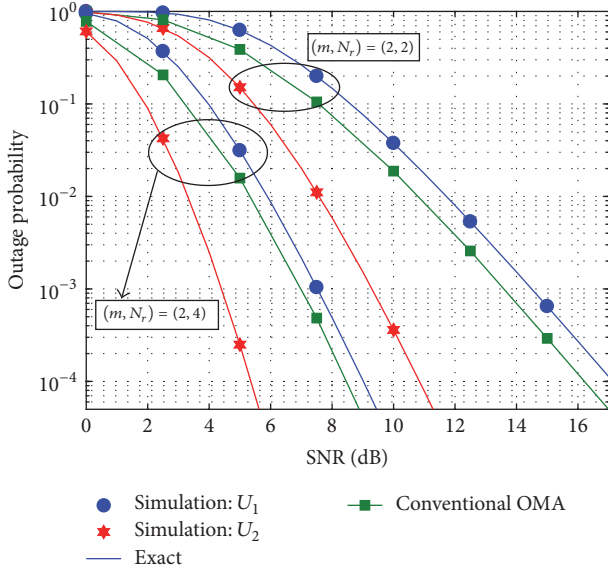


FIGURE 6: Outage probability of MIMO-NOMA system versus SNR for $L = 2$, $a_1 = 0.6$, $a_2 = 0.4$, $\gamma_{th_1} = 1$, $\gamma_{th_2} = 2$, and $\gamma_{th} = 5$.

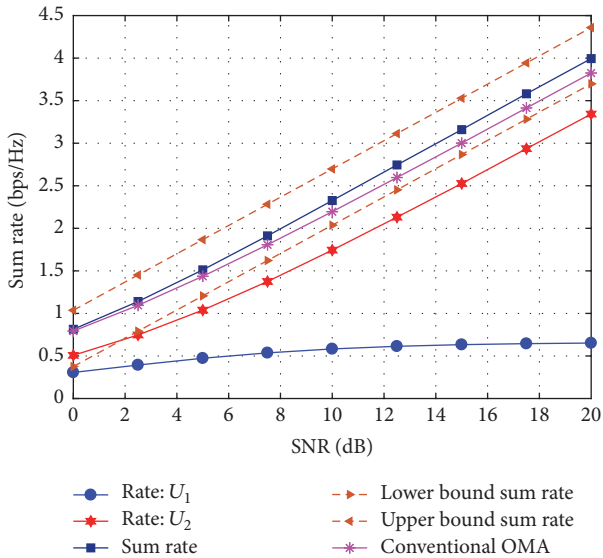


FIGURE 7: Ergodic sum rate of MIMO-NOMA system versus SNR for $L = 2$, $a_1 = 0.6$, $a_2 = 0.4$, $\gamma_{th_1} = 1$, $\gamma_{th_2} = 2$, $\gamma_{th} = 5$, and $(m, N_r) = (2, 2)$.

of conventional OMA and the proposed NOMA in terms of OP and ergodic sum rate over Nakagami- m fading channels, SNR threshold value of conventional OMA γ_{th} , which verifies $(1/2) \sum_{i=1}^L \log_2(1 + \gamma_{th_i}) = (1/2) \log_2(1 + \gamma_{th})$, is used.

Figure 6 shows the outage probability versus the system SNR over different Nakagami m parameters. In Figure 6, the simulations verify exact analytical results and a better outage performance at higher number of antennas is obtained.

Figure 7 depicts the ergodic sum rates of mobile users versus the system SNR. It is observed that ergodic rate for the first user is approximately constant over high SNR. This is due to high power allocation for the first user, such that it

considers the signal of the second user as noise, while ergodic rate for the second user proportionally increases with SNR because of no interference with the first one. Figures 6 and 7 show that NOMA outperforms conventional OMA in terms of outage probability and ergodic sum rate, respectively.

4. Cooperative NOMA

Cooperative communication, where the transmission between the source and destination is maintained by the help of one or multiple relays, has received significant attention of researchers since it extends the coverage area and increases system capacity while reducing the performance deteriorating effects of multipath fading [75, 76]. In cooperative communication systems, relays transmit the received information signals to the related destinations by applying forwarding protocols, such as amplify-and-forward (AF) and decode-and-forward (DF). In addition, in the last decade, the relays can be fundamentally categorized as half-duplex (HD) and full-duplex (FD) according to relaying operation. Differing from HD, FD relay maintains the data reception and transmission process simultaneously in the same frequency band and time slot [77]. Thus, FD relay can increase the spectral efficiency compared to its counterpart HD [78]. Therefore, the combination of cooperative communication and NOMA has been considered as a remarkable solution to further enhance the system efficiency of NOMA. Accordingly, in [79], a cooperative transmission scheme, where the users with stronger channel conditions are considered as relays due to their ability in the decoding information of other users in order to assist the users with poor channel conditions, has been proposed to be implemented in NOMA. In [80], by assuming the same scenario in [79], Kim et al. proposed a device-to-device aided cooperative NOMA system, where the direct link is available between the BS and one user, and an upper bound related to sum capacity scaling is derived. In addition, a new power allocation scheme is proposed to maximize the sum capacity. On the other hand, in [81], the authors analyze the performance of NOMA based on user cooperation, in which relaying is realized by one of the users, operating in FD mode to provide high throughput, by applying power allocation.

However, aforementioned user cooperation schemes are more appropriate for short-range communications, such as ultrawideband and Bluetooth. Therefore, in order to further extend the coverage area and to exploit the advantages of cooperation techniques, the concept of cooperative communication, where dedicated relays are used, has also been investigated in NOMA. In this context, in [82], a coordinated transmission protocol where a user communicates with BS directly while the other needs the help of a relay to receive the transmitted information from the BS has been employed in NOMA scheme in order to improve the spectral efficiency, and OP analysis is conducted for frequency-flat block fading channels by using DF relaying, as shown in Figure 8(a). In [83], the same scenario in [82] is considered, and OP and asymptotic expressions are obtained in approximated closed forms for AF relaying networks. Differing from [82] and [83], in [84], the authors proposed a cooperative relaying

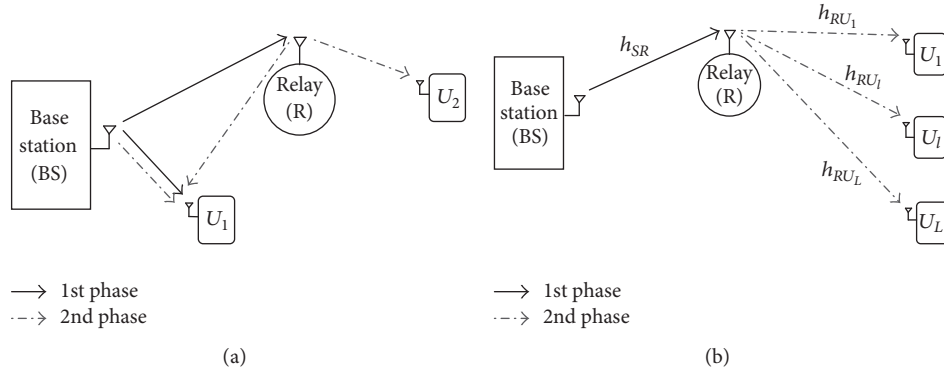


FIGURE 8: System model of cooperative NOMA downlink. (a) Coordinated direct and relay transmission. (b) A cooperative scheme without direct link.

system, where two symbols transmitted from the BS to the user by the help of a relay were combined at the BS by applying NOMA concept. The exact and asymptotic expressions related to achievable average rate are derived in i.i.d. Rayleigh fading channels and the results demonstrate that cooperative relaying based on NOMA outperforms the conventional one. Also, the authors of [85] analyzed the same transmission scheme in [84] over Rician fading channels. In order to further improve the achievable rate of the system investigated in [84], in [86], authors proposed a novel receiver scheme, where the transmitted symbols from the BS are combined at the destination according to MRC technique and investigated the system performance in terms of ergodic sum rate and OP. Their results demonstrate that the proposed scheme achieves better performance than the one in [84]. In addition, Wan et al. [87] investigated the same system in [86] by using two DF relays and assuming no direct link for cooperation and analyzed the system performance in terms of achievable sum rate. In [88], the authors investigate the performance of NOMA over i.i.d. Rayleigh fading channels by employing a downlink cooperative network in which the BS transmits the superimposed information to the mobile users through a relay and also the direct link is considered. The OP expression of the related user is obtained in closed form, and ergodic sum rate and asymptotic analyses are also maintained as performance metrics. The results show that the NOMA exhibits the same performance in terms of diversity order when compared to OMA by improving spectral efficiency and providing a better user fairness. Furthermore, in [89], performance of NOMA is investigated in relaying networks without the direct link over Nakagami- m fading environments for the network given in Figure 8(b) where all nodes and mobile users are assumed to have a single antenna. While closed-form OP expressions and simple bounds are obtained, ergodic sum rate and asymptotic analyses are also conducted. Under the consideration of imperfect CSI, the authors of [90] analyze the performance of NOMA system investigated in [89] in terms of OP. They provide exact OP and lower bound expressions in closed form and their results show that an error floor comes up due to the imperfect CSI at all SNR region. Similar to the scenario in [89], in [91], performance of NOMA with

fixed gain AF relaying is analyzed over Nakagami- m fading channels in case when the direct transmission also exists. For performance criterion, new closed-form expressions related to the exact and asymptotic OPs are obtained. Moreover, a buffer-aided cooperative technique, where the relay transmits and receives the information packets when source-relay and relay-destination links are in outage, respectively, has been taken into account by researchers in order to further enhance the reliability of the relaying systems and increase the system throughput [92]. Accordingly, in [93], the authors proposed a cooperative NOMA system with buffer-aided relaying technique consisting of one source and two users in which the stronger user is used as a buffer-aided relay. Differing from [93], Zhang et al. [94] proposed a buffer-aided NOMA relay network in which a dedicated relay was used to forward the information to two users, and exact OP of the system was obtained in single integral form and lower/upper bounds were derived in closed forms. In [95], for the same system in [94], an adaptive transmission scheme in which the working mode is adaptively chosen in each time slot is proposed to maximize the sum throughput of the considered NOMA system.

As can be seen from the aforementioned studies, the power allocation issue is vital for the performances of user destinations. In this context, there are several studies that focus on power allocation strategies for cooperative NOMA in the literature [96–99]. Accordingly, in [96], the authors proposed a novel two-stage power allocation scheme for cooperative NOMA with direct link consisting of one source, one relay, and one user destination in order to improve sum rate and OP of the system. In [97], Gau et al. proposed a novel dynamic algorithm that selects the optimal relaying mode and determines the optimal power allocation for cooperative NOMA, where the BS communicates with two users via a couple of dedicated relays. For the proposed approach, new closed-form expressions related to optimal power allocation were derived. In [98], the authors investigated a joint sub-carrier pairing and power allocation problem in cooperative NOMA which consists of one BS and two users (one of the users acts as a relay). Theoretical expressions related to joint

optimization approach are derived and superiority of the considered algorithms is demonstrated by simulations. In [99], in order to optimize the resource allocation for maximizing the average sum-rate, authors studied the performance of a single-cell NOMA system consisting of multiple source-destination pairs and one OFDM AF relay.

As well known from the literature, diversity techniques and using multiantenna strategies improve system performance significantly. Therefore, in [100], the same authors of [88] consider using multiple antennas at the BS and mobile users and analyze the OP behavior of the network over i.i.d. Rayleigh in case when the direct link does not exist. They apply TAS and MRC techniques at the BS and mobile users, respectively, while the relay has single antenna and show that using multiple antennas improves the system OP performance. Additionally, it is shown that NOMA provides a better OP performance than OMA when the distance between the BS and relay is sufficiently short. In [101], OP performance of the same system investigated in [100] was analyzed for Nakagami- m channels in case that fixed gain AF relay was used. In [102], performance of the same system in [100] was investigated over Nakagami- m fading environments in the presence of imperfect CSI. The system OP was obtained in closed form and tight lower/upper bounds were provided for further insights. In [103], the authors proposed an Alamouti space-time block coding scheme based on two-phase cooperative DF relaying for NOMA and obtained closed-form expressions for both OP and ergodic sum-rate. In [104], the authors analyzed the system performance of nonregenerative massive MIMO NOMA relay network in case that SIC and maximum mean square error SIC techniques were adopted at the receivers. In the system, multiple users and relays are equipped with single antenna while the BS has multiple antennas. For performance metrics, system capacity and sum rate expressions were derived in closed forms and authors demonstrated that the considered system outperforms massive MIMO OMA.

In addition to the aforementioned studies, using multirelays and/or relay selection techniques in cooperative NOMA concept are hot issues since using multiple relays improves the system performance significantly as already known from studies in the literature. Therefore, in [105], the authors proposed a novel NOMA relaying system based on hybrid relaying scheme, where some of relays adopted DF protocol while the others used AF for signal transmission, consisting of two sources and one user destination. For performance comparison with the conventional systems, channel capacity and average system throughput were investigated, and the proposed system was shown to achieve larger sum channel capacity and average system throughput than the conventional systems. Gendia et al. [106] investigated a cooperative NOMA with multiple relays in which all users except the user to whom the information signal would be transmitted were considered as relays. Comparisons with the other equivalent NOMA systems were done in terms of user-average bit error rate, ergodic sum rate, and fairness level by simulations. In [107], OP performance of a NOMA system, where the BS transmits the information signals to two users by using two relays, was analyzed when cooperative and TDMA schemes

were applied for transmission. The authors demonstrated that cooperative scheme outperforms TDMA one in terms of OP. Shin et al. [108] proposed a novel multiple-relay-aided uplink NOMA scheme for multicell cellular networks where the BS was equipped with multiantenna and limited by user numbers in each cell. Moreover, the feasibility conditions of the considered system were investigated. Besides multirelaying strategies, relay selection techniques were also investigated. Accordingly, in [109], the authors investigated the impact of two relay selection techniques on the performance of cooperative NOMA scheme without direct link. According to the results, with the relay selection strategies significant performance gain in terms of OP has been achieved in NOMA compared to counterpart OMA. In [110], performance of a cooperative NOMA with the best relay selection technique was analyzed in terms of average rate. The considered relay network consists of one BS, one user, and multiple relays and the direct link is also available. Authors demonstrated that the significant performance gain can be achieved by increasing the number of relays when compared to OMA one. Deng et al. [111] investigated the joint user and relay selection problem in cooperative NOMA relay networks, where multiple source users communicate with two destination users via multiple AF relays. In order to improve the system performance, the authors proposed an optimal relay selection scheme, where the best user-relay pair was selected. In [112], performance of cooperative NOMA with AF relays was analyzed by using partial relay selection technique. In the network, communication between the BS and two users was realized by selected relay, and also direct link between the BS and users was taken into account. While authors provided closed-form OP and sum rate expressions, asymptotic analysis at high SNR region was also conducted. It is shown that the performance can be improved by increasing the number of relays, but the same performance gain is obtained at high SNR region for more than two relays. In addition to above studies, Yang et al. [113] proposed a novel two-stage relay selection scheme for NOMA networks which consists of one source, multiple DF/AF relays, and two users. The considered selection strategy relies on satisfying the QoS of one user in the first stage while maximizing the rate of the other user in the second stage.

Besides that NOMA improves the system spectral efficiency, energy harvesting (EH) technology has also gained much attention because of its ability in increasing energy efficiency. Therefore, simultaneous wireless information and power transfer (SWIPT), which uses radio-frequency signals to enable self-sustainable communication, was proposed by Varshney [114] and regarded as an efficient solution over all emerging EH techniques due to the limitation of environmental energy sources. In this context, many studies combining cooperative NOMA with EH technologies were conducted in the literature [115–123]. In order to exploit the energy and spectral efficiency features of SWIPT and NOMA, Liu et al. [115] studied the application of SWIPT to cooperative NOMA, where users nearby to the BS act as EH relays. In addition, different user selection schemes were proposed in order to determine which nearby user would

cooperate with far user, and OP and throughput expressions related to the selection schemes were obtained in closed forms. In [116], a transceiver design problem in cooperative NOMA with SWIPT was studied. In the considered system, the stronger user acting as a relay and BS were equipped with multiple antennas while the other user had only single antenna. Optimal transmitter beamforming and ZF-based transmitter beamforming structures were proposed to maximize the rate of relay node. In [117, 118], the authors analyzed OP performance of NOMA-SWIPT relay networks over i.i.d. Rayleigh and Nakagami- m fading environments, respectively. Differing from the previous works, authors considered that the BS and multiple users were equipped with multiple antennas and communication between the BS and users was established only via an EH relay. They considered that TAS and MRC techniques were employed at the BS and users, respectively, and proved closed-form OP expressions for performance criterion. Similar to [115], in [119], a best-near best-far user selection scheme was proposed for a cellular cooperative NOMA-SWIPT system and OP analysis was conducted to demonstrate the superiority of the proposed scheme. In [120], the authors investigated TAS schemes in MISO-NOMA system based on SWIPT technique, where the BS with multiple antennas communicates with two users with single antenna and the stronger user is also used as an EH relay, in terms of OP and conducted diversity analysis. The impact of power allocation on cooperative NOMA-SWIPT networks was investigated by Yang et al. [121]. For performance comparisons with existing works, OP and high SNR analyses were conducted, and the proposed system was shown to improve the OP performance significantly. In [122], authors analyzed OP performance of a downlink NOMA with EH technique consisting of one BS and two users. While the BS and one of the users which was used as a relay were equipped with multiple antennas, the other user far from the BS had only single antenna. Closed-form OP expressions were derived for AF, DF, and quantize-map-forward relaying protocols over i.i.d. Rayleigh fading channels. Xu et al. [123] investigated joint beamforming and power splitting control problem in NOMA-SWIPT system studied in [120]. In order to maximize the rate of the relay user, power splitting ratio and beamforming vectors were optimized. Moreover, SISO-NOMA system was also studied.

While most of the prior works on the cooperative NOMA systems have focused on the use of HD relaying technique, there are also some studies that consider using FD relaying technique in order to further increase spectral efficiency of NOMA systems. In [124], performance of cooperative SISO-NOMA relaying system consisting of one BS and two users was investigated. The user near BS was considered as an FD relay which employed compress-and-forward protocol for poor user. Authors provided theoretical expressions of achievable rate region based on the noisy network coding. Zhong and Zhang [125] proposed using FD relay instead of HD for the investigated system in [82], where one user can communicate with the BS directly while the other needs a relay cooperation. In order to demonstrate the superiority of using FD relay, authors provided exact OP and ergodic

sum capacity expressions. In [126], OP performance of cooperative NOMA system in which the strong user helps the other by acting as an FD-DF relay was analyzed in terms of OP. Moreover, an adaptive multiple access scheme that selects access mode between proposed NOMA, conventional NOMA, and OMA was investigated in order to further enhance the system OP. Differing from [126], authors of [127] investigated optimizing the maximum achievable rate region of cooperative NOMA system in which the BS also operated in FD mode. Therefore, the authors proposed three approaches for maximization problem, such as fixed transmit power, nonfixed transmit power, and transmit power corrupted by error vector magnitude. In [128], a hybrid half/full-duplex relaying scheme was proposed to implement in cooperative NOMA and power allocation problem was investigated in terms of achievable rate. In addition, NOMA with HD and NOMA with FD systems were separately investigated by providing closed-form optimal expressions related to powers. Hybrid NOMA scheme was shown to outperform the other NOMA schemes. The same hybrid NOMA system in [128] was also investigated by Yue et al. [129] in terms of OP, ergodic rate, and energy efficiency. In addition, the authors also investigated the system when the direct link was not available between the BS and poor user. In [130], OP and ergodic sum rate performance of a cooperative NOMA system with FD relaying was investigated in case that the direct link was not available. Theoretical expressions were derived in closed forms. Moreover, in order to maximize the minimum achievable rate, optimization problem for power allocation was also studied.

In the next section, we provide an overview of the cooperative NOMA system which is investigated in [89] to provide an example of cooperative NOMA.

4.1. Performance Analysis of Cooperative NOMA. Consider a dual hop relay network based on downlink NOMA as given in Figure 8(b) which consists of one BS (S), one AF HD relay (R), and L mobile users. In the network, all nodes are equipped with a single antenna, and direct links between the BS and mobile users can not be established due to the poor channel conditions and/or the mobile users are out of the range of BS. We assume that all channel links are subjected to flat Nakagami- m fading. Therefore, channel coefficients of S - R and R - U_l are denoted by h_{SR} and h_{RU_l} with the corresponding squared means $E[|h_{SR}|^2] = \Omega_{SR}$ and $E[|h_{RU_l}|^2] = \Omega_{RU_l}$, respectively, where $l = 1, \dots, L$. In order to process NOMA concept, without loss of generality, we consider ordering the channel gains of L users as $|h_{RU_1}|^2 \leq |h_{RU_2}|^2 \leq \dots \leq |h_{RU_L}|^2$. In the first phase, the superimposed signal s given in (1) is transmitted from the BS to the relay and then the received signal at R can be modeled as

$$y_R = h_{SR} \sum_{i=1}^L \sqrt{a_i} P_s x_i + n_R, \quad (39)$$

where n_R is the complex additive Gaussian noise at R and distributed as $CN(0, \sigma_R^2)$.

In the second phase, after the relay applies AF protocol, the received signal at U_l can be written as

$$y_{RU_l} = \sqrt{P_R} G h_{SR} h_{RU_l} \sum_{i=1}^L \sqrt{a_i P_s} x_i + \sqrt{P_R} G h_{RU_l} n_R + n_{U_l}, \quad (40)$$

where n_{U_l} is the complex additive Gaussian noise at U_l and distributed as $CN(0, \sigma_{U_l}^2)$, and P_R is the transmit power at R . G denotes the amplifying factor and can be chosen as

$$G = \sqrt{\frac{P_R}{P_s |h_{SR}|^2 + \sigma_R^2}}. \quad (41)$$

In order to provide notational simplicity, we assume that $P_s = P_R = P$, $\sigma_R^2 = \sigma_{U_l}^2 = \sigma^2$. In addition, $\gamma = P/\sigma^2$ denotes the average SNR.

After the SIC process implemented at the receiver of U_l , the SINR for the l th user can be obtained as [89]

$$\gamma_{RU_l} = \frac{a_l \gamma^2 |h_{SR}|^2 |h_{RU_l}|^2}{\gamma^2 |h_{SR}|^2 |h_{RU_l}|^2 \Psi_l + \gamma (|h_{SR}|^2 + |h_{RU_l}|^2) + 1}, \quad (42)$$

where $\Psi_l = \sum_{i=l+1}^L a_i$. Then, the received SINR by the L th user can be simply expressed as [89]

$$\gamma_{RU_L} = \frac{a_L \gamma^2 |h_{SR}|^2 |h_{RU_L}|^2}{\gamma (|h_{SR}|^2 + |h_{RU_L}|^2) + 1}. \quad (43)$$

Since channel parameters are Nakagami- m distributed, $|\tilde{h}_X|^2$ squared envelope of any unordered link X , where $X \in \{SR, RU_l\}$, follows Gamma distribution with CDF

$$F_{|\tilde{h}_X|^2}(x) = \frac{\gamma (m_X, x (m_X/\Omega_X))}{\Gamma(m_X)} = 1 - e^{-x(m_X/\Omega_X)} \sum_{n=0}^{m_X-1} \left(\frac{m_X}{\Omega_X} x \right)^n \frac{1}{n!}. \quad (44)$$

In (44), right hand side of the equation is obtained by using the series expansion form of incomplete Gamma function [72, eq. (8.352.6)] and m_X denotes the Nakagami- m parameter belonging to the link X .

Furthermore, the PDF and CDF of the ordered squared envelope $|h_X|^2$ can be written by using (44) as [89]

$$f_{|h_X|^2}(x) = Q \sum_{k=0}^{L-1} (-1)^k C_k^{L-1} f_{|\tilde{h}_X|^2}(x) [F_{|\tilde{h}_X|^2}(x)]^{L+k-1}, \quad (45)$$

$$F_{|h_X|^2}(x) = Q \sum_{k=0}^{L-1} \frac{(-1)^k}{l+k} C_k^{L-1} [F_{|\tilde{h}_X|^2}(x)]^{L+k}, \quad (46)$$

where $Q = L!/(L-l)!(l-1)!$ and $C_k^K = \binom{K}{k}$ represents the binomial combination.

4.1.1. Outage Probability of Cooperative NOMA. By using the approach given in [89], the OP of the l th user can be written as

$$P_{\text{out},l} = 1 - \Pr \left(|h_{RU_l}|^2 > \eta_l^*, |h_{SR}|^2 > \frac{\eta_l^* (1 + \gamma |h_{RU_l}|^2)}{\gamma (|h_{RU_l}|^2 - \eta_l^*)} \right). \quad (47)$$

The OP expression given in (47) can be mathematically rewritten as

$$P_{\text{out},l} = \underbrace{\int_0^{\eta_l^*} f_{|h_{RU_l}|^2}(x) dx}_{J_1} + \underbrace{\int_{\eta_l^*}^{\infty} f_{|h_{RU_l}|^2}(x) F_{|h_{SR}|^2} \left(\frac{\eta_l^* (1 + \gamma x)}{\gamma (x - \eta_l^*)} \right) dx}_{J_2}. \quad (48)$$

Then, by using (44) and (45), J_2 can be calculated as

$$J_2 = 1 - J_1 - Q \sum_{k=0}^{L-1} \sum_{n=0}^{m_{SR}-1} (-1)^k C_k^{L-1} \frac{1}{n!} \cdot \int_{\eta_l^*}^{\infty} f_{|\tilde{h}_{RU_l}|^2}(x) \underbrace{(F_{|\tilde{h}_{RU_l}|^2}(x))^{L+k-1}}_{\varphi} \times e^{-(\eta_l^* (1 + \gamma x) m_{SR} / \gamma \Omega_{SR} (x - \eta_l^*))} \left(\frac{\eta_l^* (1 + \gamma x) m_{SR}}{\gamma \Omega_{SR} (x - \eta_l^*)} \right)^n dx. \quad (49)$$

In (49), by using binomial expansion [72, eq. (1.111)], φ can be obtained in closed form as

$$\varphi = \sum_{t=0}^{l+k-1} \sum_{p=0}^{t(m_{RU}-1)} C_t^{l+k-1} (-1)^t \cdot e^{-x(m_{RU}t/\Omega_{RU})} x^p \vartheta_p(t, m_{RU}), \quad (50)$$

where $\vartheta_a(b, g_c)$ denotes multinomial coefficient given in (26).

Furthermore, if we substitute derivative of (44) and (50) into (49) and then by using some algebraic manipulations, J_2 can be obtained in closed form. Then, by substituting J_2 into (48), we can obtain the OP of l th user in closed form as

$$P_{\text{out},l} = 1 - Q \sum_{k,n,t,p,i,q} (-1)^{t+k} C_k^{L-m} C_t^{l+k-1} C_i^n C_q^{p+m_{RU}-1} \cdot \frac{\vartheta_p(t, m_{RU})}{n! \Gamma(m_{RU})} \times \left(\frac{m_{RU}}{\Omega_{RU}} \right)^{(2m_{RU}-q+i-1)/2} \cdot \frac{\rho^{(i+q+1)/2}}{(t+1)^{(q-i+1)/2}} \left(\frac{m_{SR}}{\Omega_{SR}} \right)^{n-i} \eta_l^{*n-i+p+m_{RU}-1-q} \times e^{-(\eta_l^* m_{RU}(t+1)/\Omega_{RU})} e^{-(\eta_l^* m_{SR}/\Omega_{SR})} \times 2 \times K_{q-i+1} \left(2 \sqrt{\frac{\rho m_{RU} (t+1)}{\Omega_{RU}}} \right), \quad (51)$$

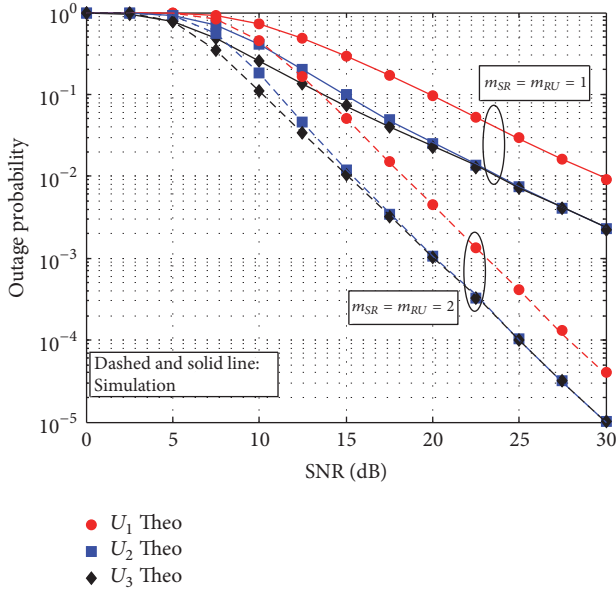


FIGURE 9: Outage probability of NOMA versus SNR in case $d_{SR} = 0.5$ and different Nakagami- m parameters.

where the binomial expansion [72, eq. (1.111)] and the integral representation in [72, eq. (3.471.9)] are used for the derivation. In (51), $\rho = \eta_l^* m_{SR} (1 + \gamma \eta_l^*) / \gamma \Omega_{SR}$ and $\sum_{k,n,t,p,i,q} \equiv \sum_{k=0}^{L-1} \sum_{n=0}^{m_{SR}-1} \sum_{t=0}^{l+k-1} \sum_{p=0}^{t(m_{RU}-1)} \sum_{i=0}^n \sum_{q=0}^{p+m_{RU}-1} (\cdot)$ notations are used to provide a short hand representation. $K_\nu(\cdot)$ denotes the ν th order modified Bessel function of second kind [72, eq. (8.407.1)]. The OP expression in (51) is in a simpler form when compared to equivalent representations in the literature.

4.1.2. Numerical Results of Cooperative NOMA. In this section, we provide numerical examples of the provided theoretical results obtained for the OP of NOMA and validate them by Monte Carlo simulations. We assume that the distances between the BS and the mobile users are normalized to one, so that $\Omega_{SR} = d_{SR}^{-\kappa}$ and $\Omega_{RU} = (1 - d_{SR})^{-\kappa}$, where $\kappa = 3$ is the path loss exponent. In all figures, $L = 3$ users and $a_1 = 1/2$, $a_2 = 1/3$, $a_3 = 1/6$, $\gamma_{th_1} = 0.9$, $\gamma_{th_2} = 1.5$, $\gamma_{th_3} = 2$ parameters have been used.

In Figure 9, we present the OP performance of NOMA versus SNR. As can be seen from the figure, theoretical results are well matched with simulations. In addition, OP performances of the second and third users are better than that of the first user and also the same at high SNR region. Moreover, as the channel parameters increase, the OPs of all users increase.

Figure 10 plots the OP performance of NOMA versus the normalized distance between the BS and the relay. As seen from the figure, while the optimal relay location of the user with the strongest channel condition is near the BS, the other users' optimal relay locations are far from the BS since the user with worse channel has higher power allocation coefficient.

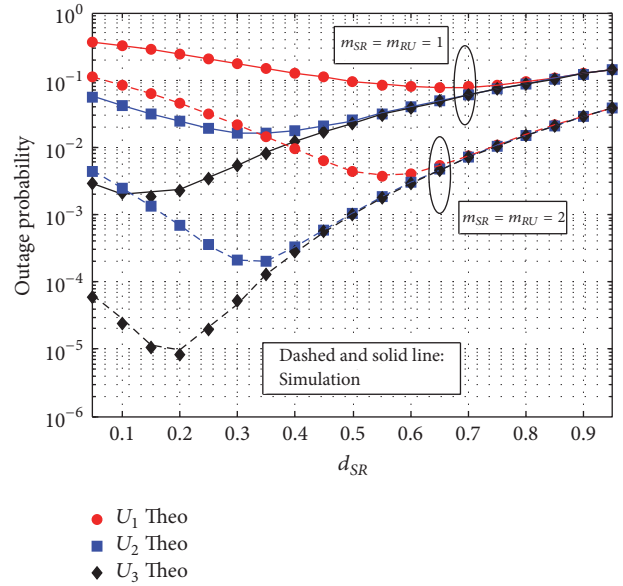


FIGURE 10: Outage probability of NOMA versus d_{SR} in case $\gamma = 20$ dB and different Nakagami- m parameters.

5. Practical Implementation Aspects

In the literature, power allocation and user clustering are generally considered as the main problems in NOMA systems, and several strategies are proposed to provide efficient solutions to these issues. As also considered in [131–133], these problems are formulated as an optimization problem and the corresponding solution procedures are also proposed. Besides these, studies, such as [54, 134, 135], propose approaches that are suitable to real-time applications. Imperfect CSI is assumed in the corresponding system models. However, real-time implementation challenges are not considered in most of the studies and the associated implementation design, which may provide effective solutions to these challenges, is not mentioned. In this section, these challenges are highlighted and important design components are explained. In the following subsection, some studies that include real-time implementation of NOMA are mentioned and challenges of such real-time implementations will be detailed.

5.1. Related Works. The number of studies that target real-time implementation of NOMA is very limited. To the best of the authors' knowledge, beyond three main studies, such content is not included in any other study at the time of preparation of this paper. In [136], single user-(SU-) MIMO is integrated to downlink and uplink NOMA, and extensive computer simulations provide detailed rate evaluation between OMA and NOMA methods. Moreover, a comprehensive testbed is created to experiment downlink NOMA with SU-MIMO setup under real-time impairments. Turbo encoding is also utilized in the implementation and a SIC decoding structure, which also includes turbo decoding and MIMO detection, is proposed. Due to usage of a wider

bandwidth, NOMA provides data rate improvement of 61% in this experiment scenario. Reference [137] targets improper power allocation issue, which is seen as a performance limiting factor in conventional NOMA models. By exploiting the physical-layer network coding (PNC) in NOMA, the authors propose network-coded multiple access (NCMA). Adaptation of PNC provides an additional transmission dimension, and the received signals via two different dimensions increase the throughput significantly when compared to the conventional NOMA systems. It is validated by experimental results that the proposed NCMA variations provide noticeable performance improvements under the power-balanced or near power-balanced scenarios. As the final study, in [138], software defined radio (SDR) implementation of downlink NOMA is realized to evaluate the performance differences between NOMA and OMA techniques. Moreover, protocol stack of LTE is modified to propose a suitable protocol stack for NOMA. Besides these multilayer modifications, detailed experiments are also carried out. Measurement results demonstrate the performance advantages of NOMA over OMA.

Since superposition coding and NOMA are very similar in context, studies on superposition coding also contain the same valuable outcomes. In [139], advantages of superposition coding over time division multiplexing approach in terms of improving the quality of the poor links are validated via an SDR platform. Accordingly, the packet error rate is measured and need of a joint code optimization is shown. Moreover, an improved packet error rate performance that is obtained with superposition coding, when compared to the results of time division multiplexing utilization, is demonstrated. Similarly in [140], the authors propose a scheduler based on superposition coding and it is demonstrated that superposition coding based resource allocation can provide a data rate improvement up to 25% when compared to the orthogonal access techniques.

These studies provide significant insights about real-time implementation aspects of NOMA. However, several practical challenges are not yet considered in available works.

5.2. Implementation Challenges. Practical implementation challenges of NOMA are considered in some surveys. In [141], the authors focus on multicell NOMA and the related design issues in the environment in the presence of a strong intercell interference (ICI). Since future wireless networks are expected to be densely deployed, NOMA technique is considered to be a candidate technique. ICI should be considered due to the potential effects of interference between adjacent BSs. Theoretical details of single-cell and multicell NOMA solutions are detailed and the capacity analysis is provided. Moreover, some major implementation issues are highlighted. Hardware complexity and error propagation issues of SIC implementation are detailed. Then, the importance of CSI is highlighted and the damaging effects of imperfect CSI on the performance of NOMA are explained. Multiuser power allocation and clustering are also emphasized. To limit ICI between adjacent cells, the authors propose that users should be clustered properly and power allocation mechanism should be operated efficiently. Integration of

fractional frequency reuse with NOMA is also considered as a major challenge and such integration should be allocated properly to obtain significant gains. Lastly, security is highlighted as another challenge, and the implementation of physical layer security techniques is seen as a difficult task. As demonstrated with computer simulations targeting to demonstrate the performance limitation of interference, proper ICI cancellation is very significant to obtain a robust performance in multicell NOMA systems.

In [142], challenges of downlink and uplink NOMA implementations and their implementation differences are explained. As the first challenge, implementation complexity is highlighted, where it is pointed out that downlink NOMA brings more complexity because of the utilization of iterative detection procedures multiple times at multiple receive nodes, when compared to the central receiver node, as applicable in uplink NOMA systems. Secondly, intra-cell/intracluster interference is stated as a crucial issue for both systems due to interference effects between users. As the third challenge, SIC receivers which are implemented differently in downlink and uplink cases are considered. Lastly, ICI is elaborated. It is shown that ICI is more effective in uplink case and could limit performance significantly. However, it is not that effective in downlink case and the observed performance degradation is comparable to that of observed in OMA systems. Moreover, some critical points are listed. Firstly, propagation errors in SIC receivers are mentioned as an important performance limiting factor and interference cancellation schemes are considered necessary to improve these effects. Secondly, multicell NOMA is highlighted, where obtaining the same single-cell NOMA gains over OMA in multicell scenarios becomes challenging. User grouping/scheduling, power allocation, and ICI mitigation are also considered as crucial items to obtain an improved performance. Besides these implementation issues, integration of NOMA-based wireless backhauling to small cells and cooperative schemes are highlighted as necessary precautions to increase NOMA's applicability in real-time.

In [143], implementation issues of NOMA are discussed and listed. Decoding complexity, error propagation, and errors that faced power balanced scenarios are also mentioned. As less considered issues, quantization errors that lead to degradation of weak signals, power allocation complexity due to difficulty of optimization of proper power levels to all users, residual timing offset that leads to synchronization loss, and error increment are highlighted. Furthermore, signaling and processing overhead due to learning procedure of CSI are also listed as a critical inefficiency source.

Some of the main problems that are mentioned in these studies and other issues that are not yet discussed in the literature will be listed and detailed below.

(1) Hardware Complexity. When compared to OMA, NOMA causes increased complexity on the hardware side due to SIC implementation. To obtain the users' symbols that transmit or receive with lower power symbols, high power symbols are required to be estimated first with the SIC detector. If the number of users especially is high or fast signal transmission is required, the SIC procedure that is used multiple times,

in addition to the detection delay, could cause important limitations for battery-limited devices. Since longer battery life is desired in consumer electronics, implementation of NOMA, particularly in dense networks, could be inefficient. This issue may limit usage of NOMA. Effective user clustering and power allocation are crucial to alleviate this problem.

(2) *Error Propagation in SIC Implementation.* According to the main principle of NOMA, on the receiver side, the user with better channel conditions is estimated first via SIC detection. Therefore, the success of the reception of main signal depends on successful estimation of the high power signals. Since channel and hardware impairments are effective in the reception process, SIC detection can be negatively affected. It is not straightforward for NOMA systems to ideally estimate channel, due to the presence of carrier frequency offset (CFO), timing offset (TO), and other hardware related impairments. Thus, erroneous detection and error propagation are probable in the SIC detection process. To overcome this and to improve the transmission quality, more robust solutions are necessary. Rather than changing the main detector components, improving the estimation quality of mentioned impairments is a more effective approach to obtain a practical performance gain.

(3) *Optimal Pilot Allocation.* Since multiple signals are transmitted in an overlapped fashion, interference emerges and error performance starts to degrade in NOMA, when compared to OMA systems. It is a clear fact that perfect or near-perfect CSI is a must to obtain a good performance. Pilot positions and the number of allocated pilots are important design considerations in NOMA implementation. These are critical even in OMA systems due to uncertain channel characteristics in wireless communication environments. However, due to the inherent interference, optimal pilot allocation is more critical for NOMA systems and careful design is required. Therefore, channel characteristics should be tracked efficiently and accurately to allocate sufficient number of pilots at proper positions, which could result in good error performance in NOMA systems.

(4) *Instantaneous CSI Requirement.* Besides pilot allocation issues in NOMA implementations, another basic CSI estimation issue exists in this process. Allocation of a previously allocated frequency band to a secondary user brings a serious problem; CSI for the transmission of this user should be estimated with orthogonal transmissions. This inevitably blocks the transmission of main user and results in an unfavorable situation. It is not clear whether this issue can be tolerated or not in real-time. Moreover, in dense networks, instantaneous band allocation may be required and, in these cases, this issue may become more critical. Effective and practical solution to this problem is very important for the future of NOMA systems. As a road map suggestion, pilot contamination problem in massive MIMO systems may be considered and corresponding solutions like [144] may be applied to NOMA systems. However, differences between the logs of these techniques should also be taken into account.

(5) *Carrier Frequency Offset and Timing Offset Estimation.* Due to the nature of wireless devices, CFO and TO emerge frequently during communication. Low-quality clocks especially that are included in such devices cause significant CFO and TO, thus, leading to a significantly degraded transmission quality. Usage of multicarrier waveforms like OFDM renders robust CFO and TO estimation and provides the necessary correction. In the point-to-point OMA transmissions, joint estimation of CFO and TO is quite straightforward due to distinguishability of received signals. Even in these cases, these impairments could cause serious performance degradation. However, this is not valid for NOMA because of the reception of signals in an overlapped fashion. This issue has not yet been considered in the literature. Effective solutions and practical approaches are required to guarantee a good transmission quality in NOMA. Highly accurate synchronization support to devices can overcome such disturbances; however, lower cost expectations prevent such a solution. Therefore, particularly, in uplink transmissions, distinguishability of overlapped signals should be achieved.

5.3. *Lessons Learned.* In order to capture the full set of advantages of NOMA in real-time that are validated in the theoretical studies, possible major challenges should be investigated and a comprehensive implementation strategy that overcomes these challenges should be determined. There are few studies in the literature that list these challenges, but there are some challenges that have not yet been considered. From this perspective, in this section, previously mentioned challenges are evaluated and important ones are given with other undetected major challenges. These also provide topics that deserve attention from the researchers who target improving NOMA's applicability.

6. Conclusion

NOMA schemes are proposed to improve the efficient usage of limited network sources. OMA based approaches that use time, frequency, or code domain in an orthogonal manner cannot effectively utilize radio resources, limiting the number of users that can be served simultaneously. In order to overcome such drawbacks and to increase the multiple access efficiency, NOMA technique has been recently proposed. Accordingly, users are separated in the power domain. Such a power-domain based multiple access scheme provides effective throughput improvements, depending on the channel conditions.

In OMA, differences between channels and conditions of users cannot be effectively exploited. It is quite possible for a user to be assigned with a large frequency band while experiencing deteriorating channel conditions. Such user cases limit the effectiveness of OMA based approaches. However, according to the NOMA principle, other users who may be experiencing better channel conditions can use these bands and increase their throughput. Moreover, corresponding users who are the primary users of these bands continue to use these bands. In such deployments, power level of users is selected in a way to target a certain maximum error rate. Furthermore, the performance of NOMA can

be significantly improved using MIMO and cooperative communication techniques.

In this paper, we provide a unified model system model for NOMA, including MIMO and cooperative communication scenarios. Implementation aspects and related open issues are detailed. A comprehensive literature survey is also given to provide an overview of the state-of-the-art.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] G. Wunder, P. Jung, M. Kasparick et al., "5G NOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, 2014.
- [2] J. G. Andrews, S. Buzzi, and W. Choi, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.
- [3] T. Rappaport, S. Sun, R. Mayzus et al., "Millimeter wave mobile communications for 5G cellular: it will work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.
- [4] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.
- [5] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.
- [6] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with successive interference cancellation for future radio access," in *Proceedings of APWCS, 2012*, Kyoto, Japan, 2012.
- [7] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, 2014.
- [8] S. Verdú, *Multuser Detection*, Cambridge University Press, New York, NY, USA, 1st edition, 1998.
- [9] Z. Yuan, G. Yu, and W. Li, "Multi-User Shared Access for 5G," *Telecommunications Network Technology*, vol. 5, no. 5, pp. 28–30, May 2015.
- [10] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 332–336, IEEE, London, UK, September 2013.
- [11] 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.
- [12] J. Zeng, D. Kong, X. Su, L. Rong, and X. Xu, "On the performance of pattern division multiple access in 5G systems," in *Proceedings of the 8th International Conference on Wireless Communications and Signal Processing, WCSP 2016*, pp. 1–5, Yangzhou, China, October 2016.
- [13] 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.
- [14] L. Dai, B. Wang, Y. Yuan, S. Han, C. 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.
- [15] "Framework and overall objectives of the future development of IMT for 2020 and beyond," Tech. Rep. ITU-R M.2083-0, 2015, <http://www.itu.int/ITU-R/>.
- [16] S. Hao, J. Zeng, X. Su, and L. Rong, "Application scenarios of novel multiple access (NMA) technologies for 5G," in *Advances in Intelligent Systems and Computing*, vol. 570, pp. 1029–1033, Springer, Cham, Switzerland, 2017.
- [17] "Overview of new radio interface," Tech. Rep. 3GPP RI-162332, Fujitsu, Busan, Korea, April 2016.
- [18] "Multiple access for 5G new radio interface," Tech. Rep. 3GPP RI-162305, CATT, Busan, Korea, April 2016.
- [19] "Candidate solution for new multiple access," Tech. Rep. 3GPP RI-162306, CATT, Busan, Korea, April 2016.
- [20] 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 the IEEE 77th Vehicular Technology Conference (VTC '13)*, pp. 1–5, Dresden, Germany, June 2013.
- [21] Z. Ding, Y. Liu, J. Choi et al., "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017.
- [22] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.
- [23] "Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)," Tech. Rep. 3GPP TR 36.859, 2015.
- [24] "Evaluation methodologies for downlink multiuser superposition transmissions," Tech. Rep. 3GPP RI-153332, NTT DOCOMO Inc., Fukuoka, Japan, May 2015.
- [25] "Deployment scenarios for downlink multiuser superposition transmissions," Tech. Rep. 3GPP RI-152062, NTT DOCOMO Inc., Belgrade, Serbia, April 2015.
- [26] "Candidate non-orthogonal multiplexing access scheme," Tech. Rep. 3GPP RI-153335, MediaTek Inc., Fukuoka, Japan, May 2015.
- [27] "System-level evaluation results for downlink multiuser superposition schemes," Tech. Rep. 3GPP RI-154536, NTT DOCOMO Inc., Beijing, China, August 2015.
- [28] "Link-level evaluation results for downlink multiuser superposition schemes," Tech. Rep. 3GPP RI-154537, NTT DOCOMO Inc., Beijing, China, August 2015.
- [29] "New work item proposal: Downlink multiuser superposition transmission for LTE," Tech. Rep. 3GPP RI-160680, MediaTek Inc., Gothenburg, Sweden, March 2016.
- [30] L. Zhang, W. Li, Y. Wu et al., "Layered-Division-Multiplexing: Theory and Practice," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 216–232, 2016.
- [31] "New study item proposal: study on non-orthogonal multiple access for NR," Tech. Rep. 3GPP RP-170829, ZTE-CATT-Intel-Samsung, Dubrovnik, Croatia, March 2016.
- [32] White Paper, "Rethink Mobile Communications for 2020+," FUTURE Mobile Commun. Forum 5G SIG, 2014, <http://www.future-forum.org/dl/141106/whitepaper.zip>.
- [33] R. Kizilirmak, "Non-Orthogonal Multiple Access (NOMA) for 5G Networks," in *Towards 5G Wireless Networks-A Physical Layer Perspective*, H. Bizaki, Ed., pp. 83–98, 2016.
- [34] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

- [35] 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.
- [36] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the Ergodic Capacity of MIMO NOMA Systems," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 405–408, 2015.
- [37] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, pp. 2123–2129, 2016.
- [38] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity Comparison between MIMO-NOMA and MIMO-OMA with Multiple Users in a Cluster," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2413–2424, 2017.
- [39] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the Sum Rate of MIMO-NOMA and MIMO-OMA Systems," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 534–537, 2017.
- [40] A. F. Molisch and M. Z. Win, "MIMO systems with antenna selection," *IEEE Microwave Magazine*, vol. 5, no. 1, pp. 46–56, 2004.
- [41] A. P. Shrestha, T. Han, Z. Bai, J. M. Kim, and K. S. Kwak, "Performance of transmit antenna selection in non-orthogonal multiple access for 5G systems," in *Proceedings of the 8th International Conference on Ubiquitous and Future Networks, ICUFN 2016*, pp. 1031–1034, Vienna, Austria, July 2016.
- [42] X. Liu and X. Wang, "Efficient antenna selection and user scheduling in 5G massive MIMO-NOMA system," in *Proceedings of the 83rd IEEE Vehicular Technology Conference, VTC Spring 2016*, Nanjing, China, May 2016.
- [43] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "Antenna selection for MIMO-NOMA networks," in *Proceedings of the 2017 IEEE International Conference on Communications, ICC 2017*, Paris, France, May 2017.
- [44] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2239–2251, 2006.
- [45] M. Kaliszán, E. Pollakis, and S. Stańczak, "Multigroup multicast with application-layer coding: Beamforming for maximum weighted sum rate," in *Proceedings of the 2012 IEEE Wireless Communications and Networking Conference, WCNC 2012*, pp. 2270–2275, France, April 2012.
- [46] B. Kimy, S. Lim, H. Kim et al., "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proceedings of the 2013 IEEE Military Communications Conference, MILCOM 2013*, pp. 1278–1283, San Diego, Calif, USA, November 2013.
- [47] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 791–800, 2015.
- [48] Y. Hayashi, Y. Kishiyama, and K. Higuchi, "Investigations on power allocation among beams in nonorthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proceedings of the 2013 IEEE 78th Vehicular Technology Conference, VTC Fall 2013*, Las Vegas, Nev, USA, September 2013.
- [49] M. S. Ali, E. Hossain, and D. I. Kim, "Non-Orthogonal Multiple Access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, 2017.
- [50] Q. Sun, S. Han, Z. Xu, S. Wang, I. Chih-Lin, and Z. Pan, "Sum rate optimization for MIMO non-orthogonal multiple access systems," in *Proceedings of the 2015 IEEE Wireless Communications and Networking Conference, WCNC 2015*, pp. 747–752, New Orleans, LA, USA, March 2015.
- [51] 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, 2016.
- [52] X. Sun, D. Duran-Herrmann, Z. Zhong, and Y. Yang, "Non-orthogonal multiple access with weighted sum-rate optimization for downlink broadcast channel," in *Proceedings of the 34th Annual IEEE Military Communications Conference, MILCOM 2015*, pp. 1176–1181, Tampa, Fla, USA, October 2015.
- [53] 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.
- [54] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, "Low Complexity Beamforming and User Selection Schemes for 5G MIMO-NOMA Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2708–2722, 2017.
- [55] Z. Ding, R. Schober, and H. V. Poor, "A General MIMO Framework for NOMA Downlink and Uplink Transmission Based on Signal Alignment," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4438–4454, 2016.
- [56] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Communications Letters*, vol. 21, no. 1, pp. 84–87, 2017.
- [57] Z. Ding, R. Schober, and H. V. Poor, "On the design of MIMO-NOMA downlink and uplink transmission," in *Proceedings of the 2016 IEEE International Conference on Communications, ICC 2016*, Kuala Lumpur, Malaysia, May 2016.
- [58] J. Cui, Z. Ding, and P. Fan, "Power minimization strategies in downlink MIMO-NOMA systems," in *Proceedings of the 2017 IEEE International Conference on Communications, ICC 2017*, Paris, France, May 2017.
- [59] V.-D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O.-S. Shin, "Precoder Design for Signal Superposition in MIMO-NOMA Multicell Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2681–2695, 2017.
- [60] Z. Ding and H. V. Poor, "Design of Massive-MIMO-NOMA with Limited Feedback," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 629–633, 2016.
- [61] C. Xu, Y. Hu, C. Liang, J. Ma, and L. Ping, "Massive MIMO, Non-Orthogonal Multiple Access and Interleave Division Multiple Access," *IEEE Access*, vol. 5, pp. 14728–14748, 2017.
- [62] J. Ma, C. Liang, C. Xu, and L. Ping, "On Orthogonal and Superimposed Pilot Schemes in Massive MIMO NOMA Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2696–2707, 2017.
- [63] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing iterative detection for MIMO-NOMA systems with massive access," in *Proceedings of the 59th IEEE Global Communications Conference, GLOBECOM 2016*, Washington, DC, USA, December 2016.
- [64] L. Liu, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving iterative LMMSE detection for MIMO-NOMA systems," in *Proceedings of the 2016 IEEE International Conference on Communications, ICC 2016*, Kuala Lumpur, Malaysia, May 2016.
- [65] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and Energy-Efficient Beamspace MIMO-NOMA for Millimeter-Wave Communications Using Lens Antenna Array," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2370–2382, 2017.

- [66] Q. Sun, S. Han, C.-L. I, and Z. Pan, "Energy efficiency optimization for fading MIMO non-orthogonal multiple access systems," in *Proceedings of the IEEE International Conference on Communications, ICC 2015*, pp. 2668–2673, London, UK, June 2015.
- [67] P. Wu, Z. Jie, X. Su, H. Gao, and T. Lv, "On energy efficiency optimization in downlink MIMO-NOMA," in *Proceedings of the 2017 IEEE International Conference on Communications Workshops, ICC Workshops 2017*, pp. 399–404, France, May 2017.
- [68] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal Precoding for a QoS Optimization Problem in Two-User MISO-NOMA Downlink," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1263–1266, 2016.
- [69] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6174–6189, 2016.
- [70] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA Design for Small Packet Transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393–1405, 2016.
- [71] Z. Chen and X. Dai, "MED Precoding for Multiuser MIMO-NOMA Downlink Transmission," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5505–5509, 2017.
- [72] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic Press, San Diego, Calif, USA, 7th edition, 2007.
- [73] H. A. David and H. N. Nagaraja, *Order Statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.
- [74] F. Xu, F. C. M. Lau, and D.-W. Yue, "Diversity order for amplify-and-forward dual-hop systems with fixed-gain relay under Nakagami fading channels," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 92–98, 2010.
- [75] A. Sendonaris, E. Erkip, and B. Aazhang, "Increasing uplink capacity via user cooperation diversity," in *Proceedings of the 1998 IEEE International Symposium on Information Theory, ISIT 1998*, p. 156, August 1998.
- [76] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.
- [77] J. Choi II, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Proceedings of the 16th Annual Conference on Mobile Computing and Networking (MobiCom '10)*, pp. 1–12, ACM, September 2010.
- [78] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: challenges, solutions, and future research directions," *Proceedings of the IEEE*, vol. 104, no. 7, pp. 1369–1409, 2016.
- [79] Z. Ding, M. Peng, and H. V. Poor, "Cooperative Non-Orthogonal Multiple Access in 5G Systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, 2015.
- [80] J.-B. Kim, I.-H. Lee, and J. Lee, "Capacity Scaling for D2D Aided Cooperative Relaying Systems Using NOMA," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 42–45, 2018.
- [81] X. Liu, X. Wang, and Y. Liu, "Power allocation and performance analysis of the collaborative NOMA assisted relaying systems in 5G," *China Communications*, vol. 14, no. 1, pp. 50–60, 2017.
- [82] J.-B. Kim and I.-H. Lee, "Non-orthogonal multiple access in coordinated direct and relay transmission," *IEEE Communications Letters*, vol. 19, no. 11, pp. 2037–2040, 2015.
- [83] X. Liang, Y. Wu, D. W. Ng, Y. Zuo, S. Jin, and H. Zhu, "Outage Performance for Cooperative NOMA Transmission with an AF Relay," *IEEE Communications Letters*, vol. 21, no. 11, pp. 2428–2431, 2017.
- [84] 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.
- [85] R. Jiao, L. Dai, J. Zhang, R. MacKenzie, and M. Hao, "On the Performance of NOMA-Based Cooperative Relaying Systems over Rician Fading Channels," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11409–11413, 2017.
- [86] M. Xu, F. Ji, M. Wen, and W. Duan, "Novel Receiver Design for the Cooperative Relaying System with Non-Orthogonal Multiple Access," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1679–1682, 2016.
- [87] D. Wan, M. Wen, H. Yu, Y. Liu, F. Ji, and F. Chen, "Non-orthogonal multiple access for dual-hop decode-and-forward relaying," in *Proceedings of the 59th IEEE Global Communications Conference, GLOBECOM 2016, USA*, December 2016.
- [88] 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.
- [89] 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.
- [90] J. Men, J. Ge, and C. Zhang, "Performance analysis for downlink relaying aided non-orthogonal multiple access networks with imperfect CSI over Nakagami-m fading channels," *IEEE Access*, vol. 5, pp. 998–1004, 2017.
- [91] X. Yue, Y. Liu, S. Kang, and A. Nallanathan, "Performance analysis of NOMA with fixed gain relaying over Nakagami-m fading channels," *IEEE Access*, vol. 5, pp. 5445–5454, 2017.
- [92] B. Xia, Y. Fan, J. Thompson, and H. V. Poor, "Buffering in a three-node relay network," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4492–4496, 2008.
- [93] Z. Liang, X. Chen, and J. Huang, "Non-orthogonal multiple access with buffer-aided cooperative relaying," in *Proceedings of the 2nd IEEE International Conference on Computer and Communications, ICC 2016*, pp. 1535–1539, China, October 2016.
- [94] Q. Zhang, Z. Liang, Q. Li, and J. Qin, "Buffer-Aided Non-Orthogonal Multiple Access Relaying Systems in Rayleigh Fading Channels," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 95–106, 2017.
- [95] S. Luo and K. C. Teh, "Adaptive transmission for cooperative NOMA system with buffer-aided relaying," *IEEE Communications Letters*, vol. 21, no. 4, pp. 937–940, 2017.
- [96] W. Duan, M. Wen, Z. Xiong, and M. H. Lee, "Two-Stage Power Allocation for Dual-Hop Relaying Systems With Non-Orthogonal Multiple Access," *IEEE Access*, vol. 5, pp. 2254–2261, 2017.
- [97] R.-H. Gau, H.-T. Chiu, C.-H. Liao, and C.-L. Wu, "Optimal Power Control for NOMA Wireless Networks with Relays," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 22–25, 2018.
- [98] X. Li, C. Li, and Y. Jin, "Joint Subcarrier Pairing and Power Allocation for Cooperative Non-orthogonal Multiple Access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10577–10582, 2017.

- [99] S. Zhang, B. Di, L. Song, and Y. Li, "Sub-Channel and Power Allocation for Non-Orthogonal Multiple Access Relay Networks with Amplify-and-Forward Protocol," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2249–2261, 2017.
- [100] J. Men and J. Ge, "Non-Orthogonal Multiple Access for Multiple-Antenna Relaying Networks," *IEEE Communications Letters*, vol. 19, no. 10, pp. 1686–1689, 2015.
- [101] M. Aldababsa and O. Kucur, "Outage performance of NOMA with TAS/MRC in dual hop AF relaying networks," in *Proceedings of the 2017 Advances in Wireless and Optical Communications (RTUWO)*, pp. 137–141, Riga, Latvia, November 2017.
- [102] Y. Zhang, J. Ge, and E. Serpedin, "Performance Analysis of Non-Orthogonal Multiple Access for Downlink Networks with Antenna Selection Over Nakagami-m Fading Channels," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10590–10594, 2017.
- [103] M. F. Kader and S. Y. Shin, "Cooperative Relaying Using Space-Time Block Coded Non-orthogonal Multiple Access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 5894–5903, 2017.
- [104] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan, and T. Sato, "Performance Analysis of Non-Regenerative Massive-MIMO-NOMA Relay Systems for 5G," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4777–4790, 2017.
- [105] Y. Liu, G. Pan, H. Zhang, and M. Song, "Hybrid Decode-Forward Amplify-Forward Relaying with Non-Orthogonal Multiple Access," *IEEE Access*, vol. 4, pp. 4912–4921, 2016.
- [106] A. H. Gendia, M. Elsabrouty, and A. A. Emran, "Cooperative multi-relay non-orthogonal multiple access for downlink transmission in 5G communication systems," in *Proceedings of the 2017 Wireless Days, WD 2017*, pp. 89–94, Portugal, March 2017.
- [107] H. Sun, Q. Wang, R. Q. Hu, and Y. Qian, "Outage probability study in a NOMA relay system," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference, WCNC 2017*, USA, March 2017.
- [108] W. Shin, H. Yang, M. Vaezi, J. Lee, and H. V. Poor, "Relay-Aided NOMA in Uplink Cellular Networks," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1842–1846, 2017.
- [109] Z. Ding, H. Dai, and H. V. Poor, "Relay Selection for Cooperative NOMA," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 416–419, 2016.
- [110] J.-B. Kim, M. S. Song, and I.-H. Lee, "Achievable rate of best relay selection for non-orthogonal multiple access-based cooperative relaying systems," in *Proceedings of the 2016 International Conference on Information and Communication Technology Convergence, ICTC 2016*, pp. 960–962, Republic of Korea, October 2016.
- [111] D. Deng, L. Fan, X. Lei, W. Tan, and D. Xie, "Joint User and Relay Selection for Cooperative NOMA Networks," *IEEE Access*, vol. 5, pp. 20220–20227, 2017.
- [112] S. Lee, D. B. da Costa, Q.-T. Vien, T. Q. Duong, and R. T. de Sousa, "Non-orthogonal multiple access schemes with partial relay selection," *IET Communications*, vol. 11, no. 6, pp. 846–854, 2017.
- [113] Z. Yang, Z. Ding, Y. Wu, and P. Fan, "Novel Relay Selection Strategies for Cooperative NOMA," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10114–10123, 2017.
- [114] L. R. Varshney, "Transporting information and energy simultaneously," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '08)*, pp. 1612–1616, IEEE, Toronto, Canada, July 2008.
- [115] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative Non-orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 938–953, 2016.
- [116] R. Sun, Y. Wang, X. Wang, and Y. Zhang, "Transceiver design for cooperative non-orthogonal multiple access systems with wireless energy transfer," *IET Communications*, vol. 10, no. 15, pp. 1947–1955, 2016.
- [117] Y. Zhang and J. Ge, "Performance analysis for non-orthogonal multiple access in energy harvesting relaying networks," *IET Communications*, vol. 11, no. 11, pp. 1768–1774, 2017.
- [118] W. Han, J. Ge, and J. Men, "Performance analysis for NOMA energy harvesting relaying networks with transmit antenna selection and maximal-ratio combining over nakagami-m fading," *IET Communications*, vol. 10, no. 18, pp. 2687–2693, 2016.
- [119] N. T. Do, D. B. Da Costa, T. Q. Duong, and B. An, "A BNBF User Selection Scheme for NOMA-Based Cooperative Relaying Systems with SWIPT," *IEEE Communications Letters*, vol. 21, no. 3, pp. 664–667, 2017.
- [120] N. T. Do, D. Benevides Da Costa, T. Q. Duong, and B. An, "Transmit antenna selection schemes for MISO-NOMA cooperative downlink transmissions with hybrid SWIPT protocol," in *Proceedings of the 2017 IEEE International Conference on Communications, ICC 2017*, France, May 2017.
- [121] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "The Impact of Power Allocation on Cooperative Non-orthogonal Multiple Access Networks with SWIPT," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4332–4343, 2017.
- [122] M. Ashraf, A. Shahid, J. W. Jang, and K.-G. Lee, "Energy Harvesting Non-Orthogonal Multiple Access System with Multi-Antenna Relay and Base Station," *IEEE Access*, vol. 5, pp. 17660–17670, 2017.
- [123] Y. Xu, C. Shen, Z. Ding et al., "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4874–4886, 2017.
- [124] J. So and Y. Sung, "Improving Non-Orthogonal Multiple Access by Forming Relaying Broadcast Channels," *IEEE Communications Letters*, vol. 20, no. 9, pp. 1816–1819, 2016.
- [125] C. Zhong and Z. Zhang, "Non-Orthogonal Multiple Access with Cooperative Full-Duplex Relaying," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2478–2481, 2016.
- [126] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device-aided cooperative nonorthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4467–4471, 2017.
- [127] H. Huang, J. Xiong, J. Yang, G. Gui, and H. Sari, "Rate Region Analysis in a Full-Duplex-Aided Cooperative Nonorthogonal Multiple-Access System," *IEEE Access*, vol. 5, pp. 17869–17880, 2017.
- [128] G. Liu, X. Chen, Z. Ding, Z. Ma, and F. R. Yu, "Hybrid Half-Duplex/Full-Duplex Cooperative Non-Orthogonal Multiple Access With Transmit Power Adaptation," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 506–519, 2018.
- [129] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, "Exploiting Full/Half-Duplex User Relaying in NOMA Systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 560–575, 2018.
- [130] L. Zhang, J. Liu, M. Xiao, G. Wu, Y.-C. Liang, and S. Li, "Performance Analysis and Optimization in Downlink NOMA Systems with Cooperative Full-Duplex Relaying," *IEEE Journal*

- on *Selected Areas in Communications*, vol. 35, no. 10, pp. 2398–2412, 2017.
- [131] 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.
- [132] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, “Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access,” in *Proceedings of the 81st IEEE Vehicular Technology Conference, VTC Spring 2015*, UK, May 2015.
- [133] F. Liu, P. Mahonen, and M. Petrova, “Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access,” in *Proceedings of the 26th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC 2015*, pp. 1127–1131, China, September 2015.
- [134] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, “Energy-efficient resource scheduling for NOMA systems with imperfect channel state information,” in *Proceedings of the 2017 IEEE International Conference on Communications, ICC 2017*, France, May 2017.
- [135] 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.
- [136] F.-L. Luo and C. Zhang, *Non-Orthogonal Multiple Access (NOMA): Concept and Design*, Wiley-IEEE Press, 2016.
- [137] H. Pan, L. Lu, and S. C. Liew, “Practical Power-Balanced Non-Orthogonal Multiple Access,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2312–2327, 2017.
- [138] X. Wei, H. Liu, Z. Geng et al., “Software Defined Radio Implementation of a Non-Orthogonal Multiple Access System Towards 5G,” *IEEE Access*, vol. 4, pp. 9604–9613, 2016.
- [139] S. Vanka, S. Srinivasa, and M. Haenggi, “A practical approach to strengthen vulnerable downlinks using superposition coding,” in *Proceedings of the 2012 IEEE International Conference on Communications, ICC 2012*, pp. 3763–3768, Canada, June 2012.
- [140] P. Vizi, S. Vanka, S. Srinivasa, M. Haenggi, and Z. Gong, “Scheduling using Superposition Coding: Design and software radio implementation,” in *Proceedings of the 2011 IEEE Radio and Wireless Symposium, RWS 2011*, pp. 154–157, USA, January 2011.
- [141] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, “Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges,” *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, 2017.
- [142] H. Tabassum, M. S. Ali, E. Hossain, M. J. Hossain, and D. I. Kim, “Non-orthogonal multiple access (NOMA) in cellular uplink and downlink: Challenges and enabling techniques,” <http://arxiv.org/abs/1608.05783>.
- [143] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, “Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.
- [144] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, “Pilot contamination and precoding in multi-cell TDD systems,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2640–2651, 2011.

Research Article

Weighted Proportional Fair Scheduling for Downlink Nonorthogonal Multiple Access

Marie-Rita Hojeij ^{1,2}, Charbel Abdel Nour ¹,
Joumana Farah ³ and Catherine Douillard ¹

¹IMT-Atlantique, CNRS UMR 6285 Lab-STICC, UBL, Brest, France

²Faculty of Engineering, Holy Spirit University of Kaslik, Jounieh, Lebanon

³Faculty of Engineering, Lebanese University, Roumieh, Lebanon

Correspondence should be addressed to Charbel Abdel Nour; charbel.abdelnour@imt-atlantique.fr

Received 5 January 2018; Revised 28 March 2018; Accepted 12 April 2018; Published 16 May 2018

Academic Editor: Muhammad Z. Shakir

Copyright © 2018 Marie-Rita Hojeij et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A weighted proportional fair (PF) scheduling method is proposed in the context of nonorthogonal multiple access (NOMA) with successive interference cancellation (SIC) at the receiver side. The new scheme introduces weights that adapt the classical PF metric to the NOMA scenario, improving performance indicators and enabling new services. The distinguishing value of the proposal resides in its ability to improve long-term fairness and total system throughput while achieving a high level of fairness in every scheduling slot. Finally, it is shown that the additional complexity caused by the weight calculation has only a limited impact on the overall scheduler complexity, while simulation results confirm the claimed improvements, making the proposal an appealing alternative for resource allocation in a cellular downlink system.

1. Introduction

Radio access technologies apply multiple access schemes to provide the means for multiple users to access and share resources at the same time. In the 3.9 and fourth generation of mobile communication systems, such as Long-Term Evolution (LTE) [1] and LTE-Advanced [2, 3], orthogonal multiple access (OMA) based on orthogonal frequency division multiplexing (OFDM) and single carrier frequency division multiple access (SC-FDMA) were adopted, respectively, for downlink and uplink transmissions. Orthogonal multiple access techniques have gained their success from their ability to achieve good system-level throughput performance in packet-domain services, while requiring a reasonable complexity, especially due to the absence of multiuser detection.

However, with the proliferation of Internet applications, between the end of 2016 and 2022, total mobile traffic is expected to increase by 8 times [4]. At the same time, communications networks are required to further enhance system efficiency, latency, and user fairness. To this

end, nonorthogonal multiple access (NOMA) has recently emerged as a promising candidate for future radio access [5]. By exploiting an additional multiplexing domain, the power domain, NOMA allows the cohabitation of two or more users per subcarrier. User multiplexing is conducted at the transmitter side, on top of the OFDM layer, and multiuser signal separation takes place at the receiver side, using successive interference cancellation (SIC) [6–12].

The main appeal of NOMA is that it improves user fairness while maximizing the total user throughput. The majority of existing works dealing with scheduling in NOMA have investigated and proposed new techniques for improving the system-level performance in terms of system capacity and cell-edge user throughput [13, 14].

In [15], throughput performance is assessed for an uplink nonorthogonal multiple access system where optimized scheduling techniques are proposed and evaluated. A cost function is assigned to each possible pair of users, in order to maximize either the sum rate or the weighted sum rate. The user pairing problem is solved by the Hungarian method and

significant improvements in sum rates and cell-edge rates are shown compared to OMA.

In [16], several new strategies for the allocation of radio resources (in terms of bandwidth and power) in a downlink NOMA system have been investigated and evaluated. The main objective of [16] is to minimize the number of allocated subbands, while guaranteeing a requested service data rate for each user. In this sense, several design issues have been explored: choice of user pairing, subband assignment, optimal and suboptimal power allocation, and dynamic switching to OMA. Simulation results show that the proposed resource allocation techniques provide better performance when NOMA is used, compared to OMA.

In addition to proposing new scheduling techniques for a NOMA-based system, some papers have investigated the commonly used PF scheduler for the good tradeoff it provides between system capacity and user fairness. Several enhancements have been proposed to the PF scheduler in order to further improve the system-level performance of a NOMA system.

In [17], an improved downlink NOMA scheduling scheme based on the PF scheduler is proposed and evaluated. In this sense, modifications to the PF scheduling metric have been introduced in order to improve the fairness at every frame assignment. Results have shown improved performance compared to a NOMA-based system that considers the conventional PF scheduler.

In [11], a weighted PF-based multiuser scheduling scheme is proposed in the context of a nonorthogonal access downlink system for the aim of further enhancing the gain of the cell-edge user. A frequency block access policy is proposed for cell-interior and cell-edge user groups using fractional frequency reuse (FFR), with significant improvements in the user fairness and system frequency efficiency.

Proposing enhancements to the PF scheduling metric has also been investigated in an OMA-based system. Similar to the work done in [11], several papers have proposed weighted versions of the PF scheduler, with the aim of improving user fairness in the OMA context.

In [18], fair weights have been implemented for opportunistic scheduling of heterogeneous traffic types for OMA networks. For designing fair weights, the proposed scheduler takes into account the average channel status as well as resource requirements in terms of traffic types. Simulation analysis demonstrates the efficiency of the proposed scheme in terms of resource utilization and its flexibility with regard to network characteristics changes due to user mobility.

In [19], the problem of fairness deficiency encountered by the PF scheduler when the mobiles experience unequal path loss is investigated. To mitigate this issue, a modified version of the PF scheduler introducing distance compensation factors has been proposed. This solution was shown to achieve both high capacity and high fairness.

In [20], a weighted PF algorithm is proposed in order to maximize best-effort service utility. The reason behind introducing weight factors into the PF metric is to exploit the inherent near-far diversity given by the path loss. The proposed algorithm enhances both best-effort service utility

and throughput performance, with a complexity similar to the complexity of the conventional PF scheduler.

Designing fair weights within the PF scheduling metric has shown remarkable improvements in the system's performance of OMA and NOMA-based systems, especially at the level of user fairness.

The above-mentioned studies have shown the advantages of introducing fair weights within the PF scheduler, especially when combined with a NOMA system. However, increasing the achieved user rate at every slot has not been tackled in these studies. Indeed, such a feature can have a positive impact on the perceived quality of service (QoS), especially for multimedia services on one side, and can reduce the amount of required buffering and memory at the user terminal on the other side. Accordingly, we aim in this paper to combine the advantages of NOMA in terms of spectral efficiency with an implementation of fair weights at the scheduling level in order to improve both the achieved user rate and the user fairness at each time slot. We propose indeed a weighted PF metric where several designs of the introduced weights are evaluated. The proposed scheme aims at providing fairness among users for each channel realization. By doing so, not only is short-term fairness achieved but also user capacity and long-term fairness are enhanced accordingly. On the other hand, the proposed schemes mitigate the problem of zero-rate incidence, inherent to PF scheduling, by attempting to provide nonzero rate to each user in any time scale of interest. This will further enhance the quality of experience (QoE) of all users.

This paper is organized as follows. In Section 2, we introduce the system model and give a general description of the NOMA-based PF scheduler. Section 3 details the proposed weighted schemes in the NOMA context. In Section 4, we apply the fair weights to a resource allocation system based on OMA. Section 5 describes a specific treatment to be applied to the first time slot in order to further enhance fairness. In Section 6, we propose some changes to the weighted metrics in order to give the possibility of delivering different levels of quality of service. Simulation results are given and analyzed in Section 7, while Section 8 concludes the paper.

2. System Description

2.1. Basic NOMA System. In this section, we describe the basic concept of NOMA including user multiplexing at the transmitter of the base station (BS) and signal separation at the receiver of the user terminal.

In this paper, a downlink system with a single input single output (SISO) antenna configuration is considered. The system consists of K users per cell, with a total bandwidth B divided into S subbands.

Among the K users, a subset of users $U_s = \{k_1, k_2, \dots, k_n, \dots, k_{n(s)}\}$ is selected to be scheduled over each frequency subband s ($1 \leq s \leq S$). The n th user ($1 \leq n \leq n(s)$) scheduled at subband s is denoted by k_n , and $n(s)$ indicates the number of users nonorthogonally scheduled at subband s . At the BS transmitter side, the information sequence of each scheduled user at subband s is independently coded and modulated, resulting in symbol x_{s,k_n} for the n th scheduled user. Therefore,

the signal transmitted by the BS on subband s , x_s , represents the sum of the coded and modulated symbols of the $n(s)$ scheduled users:

$$x_s = \sum_{n=1}^{n(s)} x_{s,k_n}, \quad \text{with } E \left[|x_{s,k_n}|^2 \right] = P_{s,k_n}, \quad (1)$$

where P_{s,k_n} is the power allocated to user k_n at subband s . The received signal vector of user k_n at subband s , y_{s,k_n} , is represented by

$$y_{s,k_n} = h_{s,k_n} x_{s,k_n} + w_{s,k_n}, \quad (2)$$

where h_{s,k_n} is the channel coefficient between user k_n and the BS at subband s . w_{s,k_n} represents the received Gaussian noise plus intercell interference experienced by user k_n at subband s . Let P_{\max} be the maximum allowable power transmitted by the BS. Hence, the sum power constraint is formulated as follows:

$$\sum_{s=1}^S \sum_{n=1}^{n(s)} P_{s,k_n} = P_{\max}. \quad (3)$$

The SIC process [21] is conducted at the receiver side, and the optimal order for user decoding is in the increasing order of the channel gains observed by users, normalized by the noise and intercell interference $h_{s,k_n}^2/n_{s,k_n}$, where n_{s,k_n} is the average power of w_{s,k_n} . Therefore, any user can correctly decode the signals of other users whose decoding order comes before that user. In other words, user k_n at subband s can remove the interuser interference from the j th user, k_j , at subband s , provided $h_{s,k_j}^2/n_{s,k_j}$ is lower than $h_{s,k_n}^2/n_{s,k_n}$, and it treats the received signals from other users with higher $h_{s,k_j}^2/n_{s,k_j}$ as noise [7, 22].

Assuming successful decoding and no error propagation and supposing that intercell interference is randomized such that it can be considered as white noise [9, 15], the throughput of user k_n at subband s , R_{s,k_n} , is given by

$$R_{s,k_n} = \frac{B}{S} \log_2 \left(1 + \frac{h_{s,k_n}^2 P_{s,k_n}}{\sum_{j \in N_s, h_{s,k_n}^2/n_{s,k_n} < h_{s,k_j}^2/n_{s,k_j}} h_{s,k_n}^2 P_{s,k_j} + n_{s,k_n}} \right). \quad (4)$$

It should be noted that most of the papers dealing with resource allocation in downlink NOMA [10, 22–24] consider a maximum number of users per subband of two in order to limit the SIC complexity in the mobile receiver, except for [9] and [25], where this number, respectively, reaches 3 and 4. However, in the last two cases, static power allocation is assumed, which simplifies the power allocation step but degrades throughput performance. It has also been stated that the performance gain obtained with 3 or 4 users per subband is minor in comparison to the case with 2 users.

2.2. Conventional PF Scheduling Scheme. The PF scheduling algorithm has been proposed to ensure balance between cell throughput and user fairness. Kelly et al. [26] have defined the proportional fair allocation of rates and used a utility function to represent the degree of satisfaction of allocated users. In [27], the practical implementation of the PF scheduler is detailed: at the beginning of every scheduling slot, each user provides the base station with its channel state (or equivalently its feasible rate). The scheduling algorithm keeps track of the average throughput $T_k(t)$ of each user in a past window of length t_c . In the scheduling slot t , user k^* is selected to be served based on [27]

$$k^* = \arg \max_k \frac{R_k(t)}{T_k(t)}, \quad (5)$$

where $R_k(t)$ is the feasible rate of user k for scheduling slot t .

In [28], an approximated version of the PF scheduler for multiple users transmission is presented. This version has been adopted in the majority of the works dealing with NOMA [22, 23, 25] in order to select users to be nonorthogonally scheduled on available resources.

For a subband s under consideration, the PF metric is estimated for each possible combination of users U , and the combination that maximizes the PF metric is denoted by U_s :

$$U_s = \arg \max_U \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)}. \quad (6)$$

$R_{s,k}(t)$ denotes the instantaneous achievable throughput of user k at subband s and scheduling time slot t .

Note that the total number of combinations tested for each considered subband is

$$N_U = \binom{K}{1} + \binom{K}{2} + \cdots + \binom{K}{N(s)}. \quad (7)$$

Equation (7) represents the general formulation of the case, where a maximum of $N(s)$ users are to be multiplexed in the power domain. In this case, all the combinations of possible users up to $N(s)$ should be verified via the PF metric and the particular user combination leading to the best PF metric value should be chosen for allocation.

$R_{s,k}(t)$ is calculated based on (4), whereas $T_k(t)$ is recursively updated as follows [28]:

$$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} \sum_{s=1}^S R_{s,k}(t). \quad (8)$$

Parameter t_c defines the throughput averaging time window. In other words, this is the time horizon in which we want to achieve fairness. t_c is chosen to guarantee a good tradeoff between system performance (in terms of fairness) and system capacity. We assume in the following a t_c window of 100 time slots. With a time slot duration equal to 1 ms, a 100 ms average user throughput $T_k(t)$ is therefore considered.

3. Proposed Weighted NOMA-Based Proportional Fairness (WNOFPF) Scheduler

The PF scheduler aims at both achieving high data rates and ensuring fairness among users, but it only considers long-term fairness. In other words, a duration of t_c time slots is needed to achieve fairness among users. However, short-term fairness and fast convergence towards required performance are an important issue to be addressed in upcoming mobile standards [4].

Since all possible combinations of candidate users are tested for each subband, a user might be selected more than once and attributed multiple subbands during the same time slot. On the other hand, it can also happen that no subband and therefore no transmission rate are allocated for multiple scheduling slots to a user with a high historical rate. This behavior can be very problematic in some applications, especially those requiring a quasi-constant QoE such as multimedia transmissions. In such cases, buffering may be needed. However, such a scenario may not be compatible with applications requiring low latency transmission.

Therefore, we propose several weighted PF metrics that aim at

- (i) enhancing the user capacity, thus increasing the total achieved user throughput;
- (ii) reducing the convergence time towards required fairness performance;
- (iii) enhancing fairness among users (both long-term and short-term fairness);
- (iv) limiting the fluctuations of user data rates;
- (v) incorporating the delivery of different levels of QoS.

The proposed scheduler consists of introducing fair weights into the conventional PF scheduling metric. The main goal of the weighted metrics is to ensure fairness among users in every scheduling slot.

To do so, we start by modifying the PF metric expression so as to take into account the status of the current assignment in time slot t . Therefore, the scheduling priority given for each user is based not only on its historical rate but also on its current total achieved rate (throughput achieved during the current scheduling slot t), as proposed in [17].

Scheduling is performed subband by subband and on a time slot basis. For each subband s , the conventional PF metric $\text{PF}_s^{\text{NOMA}}$ and a weight factor $W(U)$ are both calculated for each candidate user set U . Then, the scheduler selects the set of scheduled users U_s which maximizes the weighted metric $\text{PF}_s^{\text{NOMA}}(U) \times W(U)$. The corresponding scheduling method is referred to as weighted NOMA PF scheduler, denoted by $\text{WPF}_s^{\text{NOMA}}$. The resource allocation metric can be formulated as follows:

$$\begin{aligned} \text{WPF}_s^{\text{NOMA}}(U) &= \text{PF}_s^{\text{NOMA}}(U) \times W(U) \\ U_s &= \underset{U}{\text{argmax}} \text{WPF}_s^{\text{NOMA}}(U). \end{aligned} \quad (9)$$

The proposed weight calculation for each candidate user set U relies on the sum of the weights of the multiplexed users.

$$W(U) = \sum_{k \in U} W_k(t) \quad (10)$$

with

$$W_k(t) = R_{\text{avg}}^e(t) - R_k(t), \quad k \in U. \quad (11)$$

$R_{\text{avg}}^e(t)$ is the expected achievable bound for the average user data rate in the current scheduling slot t . It is calculated as follows:

$$R_{\text{avg}}^e(t) = b \cdot R_{\text{avg}}(t-1). \quad (12)$$

Since we tend to enhance the achieved user rate in every slot, each user must target a higher rate compared to the rate previously achieved. Therefore, parameter b is chosen to be greater than 1.

The average user data rate, $R_{\text{avg}}(t)$, used in (12), is updated at the end of each scheduling slot based on the following:

$$R_{\text{avg}}(t) = \frac{1}{K} \sum_{k=1}^K \sum_{s=1}^S R_{s,k}(t), \quad (13)$$

where $R_{s,k}(t)$ is the data rate achieved by user k on subband s .

On the other hand, $R_k(t)$, the actual achieved data rate by user k during scheduling slot t , is calculated as

$$R_k(t) = \sum_{s \in S_k} R_{s,k}(t), \quad k \in U \quad (14)$$

with S_k being the set of subbands allocated to user k during time slot t . At the beginning of every scheduling slot, S_k is emptied; each time user k is being allocated a new subband, and S_k and $R_k(t)$ are both updated.

The main idea behind introducing weights is to minimize the rate gap among scheduled users in every scheduling slot, thus maximizing fairness among them. A user set U is provided with a high priority among candidate user sets if it contains nonorthogonally multiplexed users experiencing a good channel quality on subband s , having low or moderate historical rates, or/and having large rate distances between multiplexed users' actual achieved rates and their expected achievable average user throughput. The highest level of fairness is achieved when all users reach the expected user average rate $R_{\text{avg}}^e(t)$. By applying the proposed scheduling procedure, we aim to enhance long-term and short-term fairness at the same time.

It was shown in [28] that the scheduling metric PF^{NOMA} , defined in (6), strikes a good tradeoff between throughput and fairness, since it maximizes the sum of users service utility which can be formally written as

$$\text{PF}^{\text{NOMA}} = \max_{\text{scheduler}} \sum_{k=1}^K \log T_k. \quad (15)$$

Therefore, any enhanced scheduling metric like WNOFPF can strike a better throughput-fairness balance (by achieving

a higher service utility), compared to the conventional PF scheduler, provided that

$$\sum_{k=1}^K \log T_k \geq \sum_{k=1}^K \log T'_k, \quad (16)$$

where the historical rates T_k and T'_k correspond to the schedulers using the WNOPF metric and the conventional PF metric, respectively.

Inspired by the work in [20], in the sequel, we show how this goal can be achieved by an appropriate design of the weights which verifies the constraints provided in Proposition 1.

Proposition 1. *To make (16) valid, for a NOMA-based system, the following inequality should be verified:*

$$\prod_{k=1}^K W(U_k) \prod_{k=1}^K E[R_{s,k}] \geq \prod_{k=1}^K E[R'_{s,k}]. \quad (17)$$

$E[R_{s,k}]$ and $E[R'_{s,k}]$ are the statistical average of the instantaneous transmittable rate of user k on a subband s , when WNOPF and the conventional PF scheduler are applied, respectively. U_k denotes a scheduled user set containing user k , U is a possible candidate user set, and $W(U_k)$ is the weight of the set U_k .

Proof. Equation (16) can be written as

$$\prod_{k=1}^K T_k \geq \prod_{k=1}^K T'_k. \quad (18)$$

If we consider that $T_k = I_{k,\text{tot}}/(t_c \Delta T)$, where $I_{k,\text{tot}}$ is the total amount of information that can be received by user k , for a total observation time $t_c \Delta T$, and ΔT is the scheduling time slot length, we obtain

$$\prod_{k=1}^K \frac{I_{k,\text{tot}}}{t_c \Delta T} \geq \prod_{k=1}^K \frac{I'_{k,\text{tot}}}{t_c \Delta T}. \quad (19)$$

If we denote by N_k the number of allocated time slots for user k within t_c and by n_k the statistical average of the number of allocated subbands to user k per time slot, (19) can be rewritten as

$$\prod_{k=1}^K \frac{N_k n_k E[R_{s,k}] \Delta T}{t_c \Delta T} \geq \prod_{k=1}^K \frac{N'_k n'_k E[R'_{s,k}] \Delta T}{t_c \Delta T}. \quad (20)$$

Using a simple rearrangement, we get

$$\frac{\prod_{k=1}^K (N_k/t_c) S(n_k/S)}{\prod_{k=1}^K (N'_k/t_c) S(n'_k/S)} \geq \frac{\prod_{k=1}^K E[R'_{s,k}]}{\prod_{k=1}^K E[R_{s,k}]} \quad (21)$$

If $\Pr_k (= N_k/t_c)$ denotes the probability of user k being scheduled per time slot and $\text{pr}_k (= n_k/S)$ denotes the probability of user k being scheduled per subband, (21) can be reformulated as

$$\frac{\prod_{k=1}^K \Pr_k \text{pr}_k}{\prod_{k=1}^K \Pr'_k \text{pr}'_k} \geq \frac{\prod_{k=1}^K E[R'_{s,k}]}{\prod_{k=1}^K E[R_{s,k}]} \quad (22)$$

pr_k can be regarded as the probability of a set U_k , that is, $\Pr(U_k)$, being chosen among all possible candidate sets U to be scheduled per subband.

Let us consider two sets of users U_1 and U_2 . If the probability of the user set U_1 is greater than that of U_2 , U_1 will be chosen to be scheduled. Equivalently, the corresponding scheduling metric for user set U_1 will be in this case larger than that of U_2 . Therefore, in fact, a user set is chosen if the corresponding PF metric is the largest.

Thus, (22) can be equivalent to the following equation:

$$\frac{\prod_{k=1}^K \Pr_k \text{PF}^{\text{NOMA}}(U_k) W(U_k)}{\prod_{k=1}^K \Pr'_k \text{PF}^{\text{NOMA}}(U_k)} \geq \frac{\prod_{k=1}^K E[R'_{s,k}]}{\prod_{k=1}^K E[R_{s,k}]} \quad (23)$$

Note that, in a NOMA-based system, the probability of a user being scheduled per time slot remains the same when using the proposed weighted metric or the conventional PF metric, since users are distributed with uniform and random probability over the entire network in each time slot. Thus, we adopt the following approximation:

$$\Pr_k \approx \Pr'_k. \quad (24)$$

Additional observations and verifications related to this approximation are given in Section 7. Therefore, (23) and (24) can also be formulated as (17). \square

Other configurations of rate-distance weights can also be introduced. A promising one is obtained by substituting (25) for (9) and (10):

$$U_s = \arg \max_U \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)} W_k(t), \quad k \in U. \quad (25)$$

Here, the conventional NOMA-based PF metric and the weights are jointly calculated for each user k in candidate user set U . By doing so, we assign to each user its weight while ignoring the cross effect $(R_{s,k|U}(t)/T_{k|U}(t))W_{k'|U}(t)$ produced by (9), where k and k' are nonorthogonally multiplexed users in the same U . This joint-based incorporation of weights is denoted by J-WNOPF in the following evaluations.

4. Proposed Weighted OMA-Based PF Scheduler (WOPF)

In the majority of existing works dealing with fair scheduling, OMA-based systems are considered. For this reason, we propose applying the weighted proportional fair scheduling metric introduced in this paper to an OMA-based system as well. This allows the contribution of NOMA within our framework to be evaluated. In the OMA case, nonorthogonal cohabitation is not allowed. Instead, a subband s is allocated to only one user, based on the following metric:

$$k^* = \arg \max_k \frac{R_{s,k}(t)}{T_k(t)} W_k(t), \quad (26)$$

where $W_k(t)$ is the weight assigned to user k , calculated similarly to the weights in WNOPF. The conventional OMA-based PF scheduling metric is denoted by PF^{OMA} , whereas

the resulting scheduling algorithm combining OMA with the proposed weighted PF is denoted by WOPF.

OMA can be regarded as a special case of NOMA, where only one user is allowed to be scheduled per subband. Therefore, in order to achieve a higher user service utility with WOPF than with the conventional PF scheduler in OMA, Proposition 1, detailed and proven in Section 3, should also be verified for an OMA-based system. For this purpose, (17) is modified as follows:

$$\prod_{k=1}^K W_k \prod_{k=1}^K E[R_{s,k}] \geq \prod_{k=1}^K E[R'_{s,k}], \quad (27)$$

where W_k is the weight assigned to user k .

Note that, as in the NOMA case, we assume that the probability of a user being scheduled per time slot remains the same when using the proposed weighted metric or the conventional PF metric.

5. Proposed Scheduling Metric for the First Scheduling Slot

In the first scheduling slot, the historical rates and the expected user average data rate are all set to zero. Hence, the selection of users by the scheduler is only based on the instantaneous achievable throughputs. Therefore, fairness is not achieved in the first scheduling slot, and the following slots are penalized accordingly. To counteract this effect, we propose treating the first scheduling slot differently, for all the proposed weighted metrics.

For each subband s , the proposed scheduling process selects U_s among the candidate user sets based on the following criterion:

$$U_s = \arg \max_U \sum_{k \in U} \frac{R_{s,k}(t=1)}{R_k(t=1)}. \quad (28)$$

Note that when WOPF is considered, the maximum number of users per set U is limited to 1.

$R_k(t=1)$, the actual achieved throughput, is updated each time a subband is allocated to user k during the first scheduling slot. By doing so, we give priority to the user experiencing a good channel quality with regard to its actual total achieved data rate, thus enhancing fairness in the first slot.

6. Incorporation of Premium Services

In this section, we propose some changes to the proposed weighted metrics in order to give the possibility of delivering different levels of quality of service. In other words, the proposed metrics should have the ability to provide different priorities to different users or to guarantee a certain level of performance to a data flow. To do so, (11) is modified as follows:

$$W_k(t) = R_{\text{service}} - R_k(t), \quad k \in U, \quad (29)$$

where R_{service} is the data rate requested by a certain group of users, corresponding to a certain level of performance.

As an example, we detail an example of 3 services, although the proposed modifications can be applied to an arbitrary number of services. R_{service} is then defined as follows:

$$R_{\text{service}} = \begin{cases} R_{\text{basic}}, & \text{if } k \text{ requests a basic service} \\ R_{\text{silver}}, & \text{if } k \text{ requests a silver service} \\ R_{\text{gold}}, & \text{if } k \text{ requests a gold service,} \end{cases} \quad (30)$$

$k \in U.$

This modification aims to guarantee a minimum requested service data rate for each user and also tends to enhance the overall achieved fairness between users belonging to the same group, that is, asking for the same service.

7. Numerical Results

7.1. System Model Parameters and Performance Evaluation.

This subsection presents the system level simulation parameters used to evaluate the proposed scheduling techniques. The parameters considered in this work are based on existing LTE/LTE-Advanced specifications [29]. We consider a baseline SISO antenna configuration. The maximum transmission power of the base station is 46 dBm. The system bandwidth is 10 MHz and is divided into 128 subbands when not further specified. The noise power spectral density is $4 \cdot 10^{-18}$ mW/Hz. Users are deployed randomly in the cell and the cell radius is set to 500 m. Distance-dependent path loss is considered with a decay factor of 3.76. Extended typical urban (ETU) channel model is assumed, with time-selectivity corresponding to a mobile velocity of 50 km/h, at the carrier frequency of 2 GHz. In both OMA and NOMA scenarios, equal repartition of power is considered among subbands, as considered in [9, 23, 24]. In the case of NOMA, fractional transmit power allocation (FTPA) [30] is used to allocate power among scheduled users within a subband. Without loss of generality, NOMA results are shown for the case where the maximum number of scheduled users per subband is set to 2 ($n(s) = 2$).

As for parameter b in (12), after several tests, the best performance was observed for b equal to 1.5. In fact, the system has a rate saturation bound with respect to parameter b , since when we further increase b , similar performance is maintained.

7.2. Performance Evaluation. In this part, we mainly consider four system-level performance indicators: achieved system capacity, long-term fairness, short-term fairness, and cell-edge user throughput.

Several techniques are evaluated and compared. The following acronyms are used to refer to the main studied methods:

- (i) PF^{NOMA}: conventional PF scheduling metric in a NOMA-based system
- (ii) WNOPF: proposed weighted PF scheduling metric in a NOMA-based system
- (iii) J-WNOPF: proposed weighted PF scheduling metric with a joint incorporation of weights in a NOMA-based system

- (iv) $\text{PF}_{\text{modified}}^{\text{NOMA}}$: a modified version of the PF scheduling metric proposed in [17], where the actual assignment of each frame is added to the historical rate
- (v) PF^{OMA} : conventional PF scheduling metric in an OMA-based system
- (vi) WOPF: proposed weighted PF scheduling metric in an OMA-based system

In order to assess the fairness performance achieved by the different techniques, a fairness metric needs to be defined first. Gini fairness index [31] measures the degree of fairness that a resource allocation scheme can achieve. It is defined as

$$G = \frac{1}{2K^2\bar{r}} \sum_{x=1}^K \sum_{y=1}^K |r_x - r_y| \quad (31)$$

with

$$\bar{r} = \frac{\sum_{k=1}^K r_k}{K}. \quad (32)$$

r_k is the throughput achieved by user k . When long-term fairness is evaluated, r_k is considered as the total throughput achieved by user k averaged over a time-window length t_c :

$$r_k = \frac{1}{t_c} \sum_{t=1}^{t_c} R_k(t). \quad (33)$$

Otherwise, when fairness among users is to be evaluated within each scheduling slot, short-term fairness is considered and r_k is taken equal to $R_k(t)$, the actual throughput achieved by user k during scheduling slot t .

Gini fairness index takes values between 0 and 1, where $G = 0$ corresponds to the maximum level of fairness among users, while a value of G close to 1 indicates that the resource allocation scenario is highly unfair.

First, we check the validity of Proposition 1 detailed in Sections 3 and 4 and the validity of the assumption done in (24). Figure 1 shows the observed ratio between Pr_k and Pr'_k , denoted by Ratio 1, for different values of the number of users per cell. Figure 1 also shows the ratio between the left-hand and the right-hand expressions of (17), denoted by Ratio 2. Results show that Ratio 1 is very close to 1, which means that the probability of a user being scheduled per time slot remains the same under the proposed weighted metric or under the conventional PF metric. In addition, Ratio 2 is shown to be greater than 1 regardless of the number of users per cell, which verifies Proposition 1, defined in (17). The results of a similar verification for an OMA system are observed in Figure 2.

Figure 3 shows the system capacity achieved with each of the simulated methods for different numbers of users per cell. Curves in solid lines represent the NOMA case, whereas curves with dotted lines refer to OMA.

We can observe that the throughput achieved with all the simulated methods increases as the number of users per cell is increased, even though the total number of used subbands is constant. This is due to the fact that the higher the number of users per cell is, the better the multiuser diversity is exploited by the scheduling scheme, as also observed in [18].

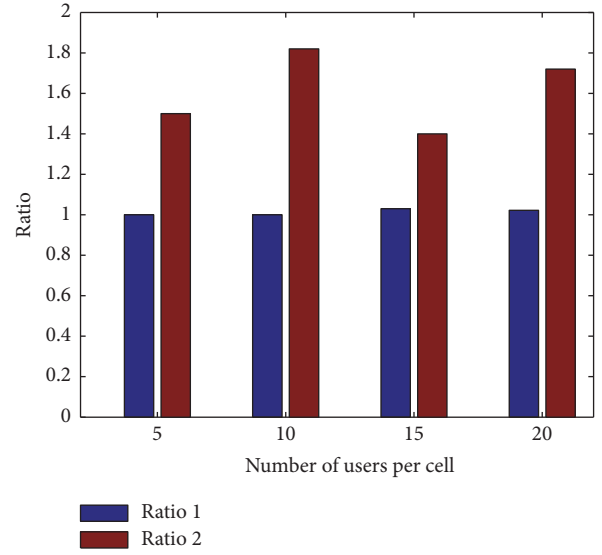


FIGURE 1: Observed ratios related to (17) and (26) versus number of users per cell, NOMA-based system.

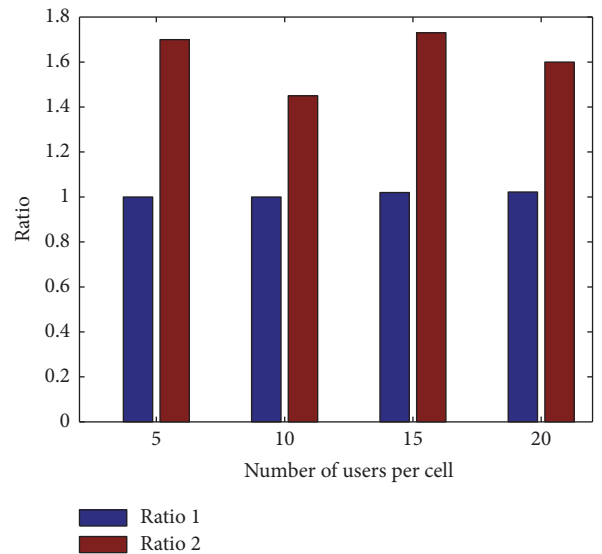


FIGURE 2: Observed ratio related to (26) and (29) versus number of users per cell, OMA-based system.

The gain achieved by WNOPF, when compared to the other proposed weighted metric J-WNOPF, is mainly due to the fact that the joint incorporation of weights does not take into consideration the cross effect produced by nonorthogonally multiplexed users.

The gain in performance obtained by the introduction of weights in the scheduling metric, compared to the conventional PF^{NOMA} metric, stems from the fact that, for every channel realization, the weighted metrics try to ensure similar rates to all users, even those experiencing bad channel conditions. With PF^{NOMA} , such users would not be chosen frequently, whereas appropriate weights give them a higher chance to be scheduled more often.

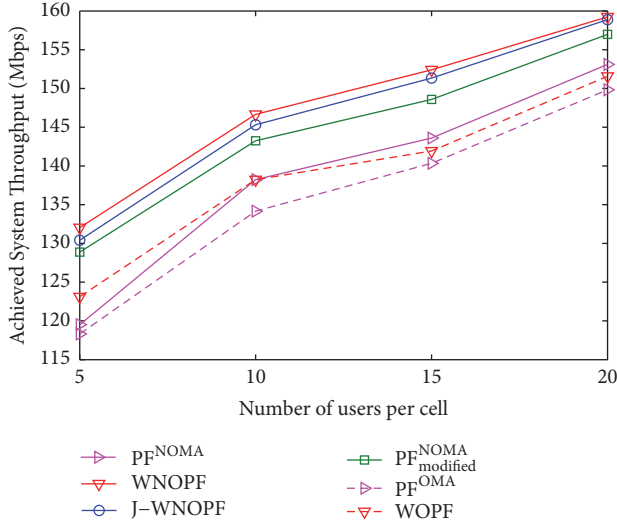


FIGURE 3: System throughput achieved with the proposed scheduling schemes versus number of users per cell.

Figure 3 also shows an improved performance of the proposed metrics when compared to the modified PF scheduling metric $PF_{\text{modified}}^{\text{NOMA}}$ described in [17]. Although they both consider the current assignment in their metric calculation, they still differ by the fact that the proposed weighted metrics target a higher rate compared to the rate previously achieved, therefore tending to increase the achieved user rate in every slot.

When the proposed scheduling metrics are applied in an OMA context, WOPF provides higher throughputs than PF^{OMA} , due to the same reason why WNOPF outperforms PF^{NOMA} . Figure 3 also shows a significant performance gain achieved by NOMA over OMA. All weighted scheduling metrics applying NOMA outperform the simulated metrics based on OMA, including WOPF. This gain is due to the efficient nonorthogonal multiplexing of users. It should also be noted that the gain achieved by WNOPF over PF^{NOMA} is greater than the one achieved by WOPF over PF^{OMA} : combining fair weights with NOMA definitely yields the best performance.

Long-term fairness is an important performance indicator for the allocation process. Figure 4 shows this metric as a function of the number of users per cell. Long-term fairness is improved when fair weights are introduced, independently of the access technique (OMA or NOMA). The reason is that, when aiming to enhance fairness in every scheduling slot, long-term fairness is enhanced accordingly. Again, in terms of fairness, the proposed weighted metrics outperform the modified PF metric [17], $PF_{\text{modified}}^{\text{NOMA}}$. This is due to the fact that WNOPF and J-WNOPF not only consider the current rate assignment but also tend to minimize the rate gap among scheduled users in every channel realization, thus maximizing fairness among them.

Figure 5 shows the achieved system throughput as a function of the number of subbands S , for 15 users per cell. We can see that the proposed weighted metrics outperform

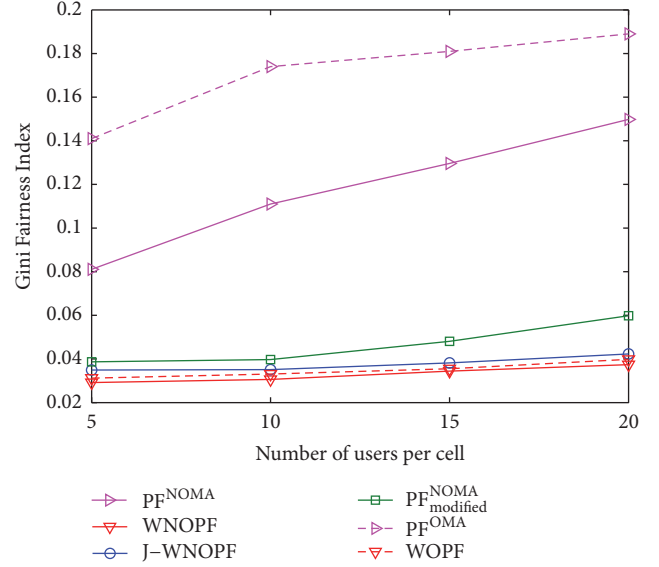


FIGURE 4: Gini fairness index of the proposed scheduling schemes versus number of users per cell.

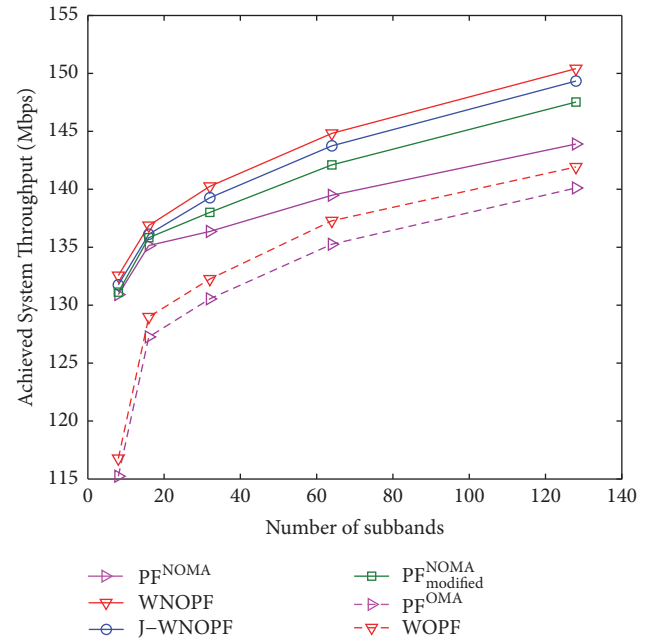
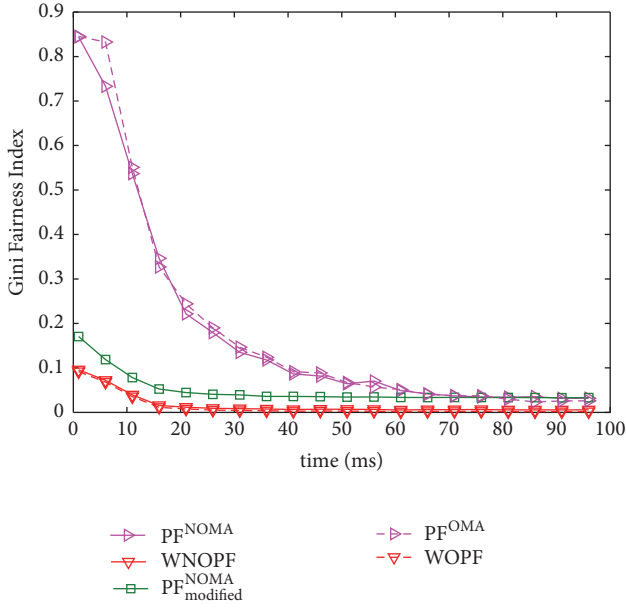


FIGURE 5: Achieved system throughput versus S , for $K = 15$.

the conventional PF scheduling scheme, for both access techniques OMA and NOMA, even when the number of subbands is limited. It is also shown that the proposed weighted metrics outperform the modified PF metric [17], $PF_{\text{modified}}^{\text{NOMA}}$.

Since WNOPF proves to give better performance than J-WNOPF, in terms of system capacity and fairness, J-WNOPF will not be considered in the subsequent results.

Since one of the main focuses of this study is to achieve short-term fairness, the proposed techniques should be compared based on the time required to achieve the final


 FIGURE 6: Gini fairness index versus scheduling time index t .

fairness level. Figure 6 shows the Gini fairness index versus the scheduling time index t . The proposed weighted metric WNOPF achieves high fairness from the beginning of the allocation process and converges to the highest level of fairness (lowest value of index $G = 0.0013$) in a limited number of allocation steps or time slots. On the contrary, PF^{NOMA} shows unfairness among users for a much longer time. Weighted metrics not only show faster convergence to a high fairness level but also give a lower Gini indicator at the end of the window length, when compared to conventional PF^{NOMA} . Figure 6 shows also that the proposed weighted metric WNOPF still outperforms the modified PF method proposed in [14].

In order to assess the QoE achieved by the proposed scheduling schemes, we evaluate the time required for each user to be served for the first time, referred to as the rate latency, as well as the variations of its achieved rate over time. For this purpose, Figure 7 shows the achieved rate versus time for the user experiencing the largest rate latency, for the different scheduling schemes.

When the conventional PF^{NOMA} is used, no rate is provided for this user, for the first five scheduling slots. In addition, large rate fluctuations are observed through time. In contrast, when weighted metrics and a special treatment of the first time slot are considered, a nonzero rate is assigned for the least privileged users from the first scheduling slot and remains stable for all the following slots. This behavior results from the fact that, at the beginning of the scheduling process (first scheduling slot), historical rates are set to zero, and PF^{NOMA} uses only instantaneous achievable throughputs to choose the best candidate user set. Therefore, users experiencing bad channel conditions have a low chance to be chosen. The corresponding achieved data rates are then equal to zero. On the other side, using the

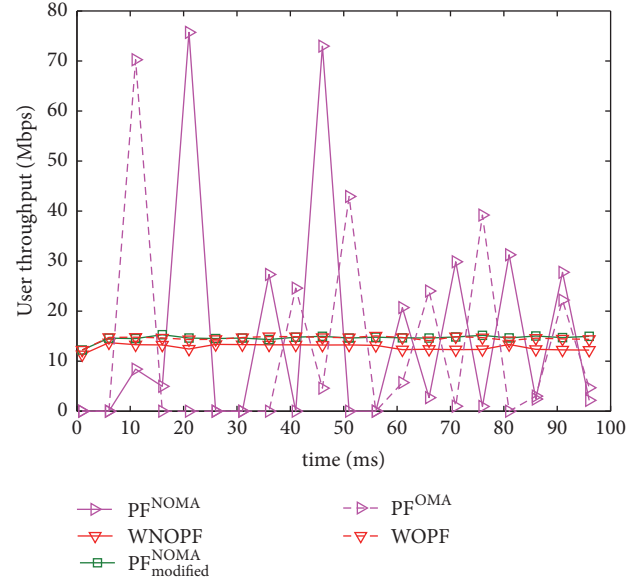


FIGURE 7: User throughput versus time for NOMA-based scheduling schemes.

proposed scheduling, the treatment of the first scheduling slot is conducted differently and users are chosen depending on their actual rates (measured during the actual scheduling period). In this case, zero rates are eliminated. Hence, latency is greatly reduced.

For the next scheduling slots, historical rates are taken into account. For PF^{NOMA} , users experiencing a large $T_k(t)$ have less chance to be chosen and may not be chosen at all. In this case, the use of buffering becomes mandatory and the size of the buffer should be chosen adequately to prevent overflow when peak rates occur, as a result of a high achieved throughput (high $R_{s,k}(t)$). Based on calculation, the average size of the buffer should be around 110 Mbit, for the simulation case at hand. However, in the case of the weighted proposed metrics, buffering is not needed, since only small variations between user data rates are observed, and a better QoE is achieved. Similar performance improvement is obtained for the orthogonal case in the same aforementioned conditions.

Finally, we have analyzed the effect of the proposed scheduling scheme on the cell-edge user throughput in Figure 8. Again, the proposed weighted metrics outperform the conventional PF scheduling scheme for both access techniques, OMA and NOMA. In addition, WNOPF shows the best performance. Therefore, we can state that the incorporation of fair weights with a NOMA-based system proves to be the best combination.

In order to evaluate the performance of the proposed weighted metrics when premium services are considered, Tables 1 and 2 show the Gini fairness index values for two different scenarios, where three levels of services are requested: basic, silver, and gold. The number of users per group is set to 5.

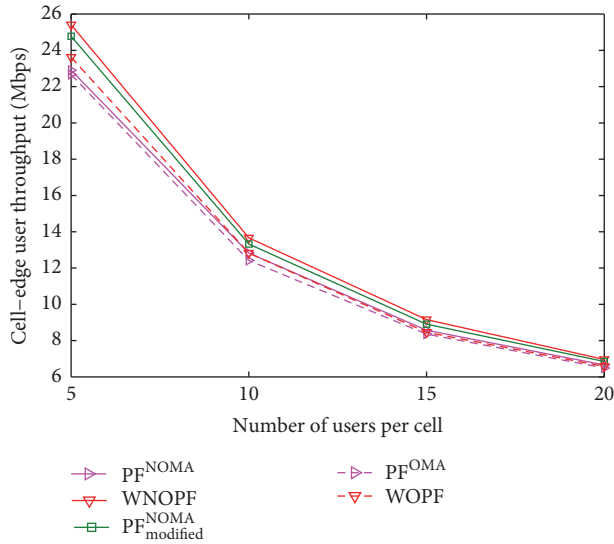


FIGURE 8: Cell-edge user throughput versus number of users per cell.

TABLE 1: Gini fairness index and data rate achieved per group for Scenario 1 (100% success).

Service	Gini fairness index	Achieved data rate per group (Mbps)
Basic	0.0491	25.7
Silver	0.0724	51
Gold	0.0042	76.3

TABLE 2: Gini fairness index and data rate achieved per group for Scenario 2 (no success).

Service	Gini fairness index	Achieved data rate per group (Mbps)
Basic	0.0522	30.2
Silver	0.0613	49.6
Gold	0.0049	75.2

Scenario 1. The corresponding data rates of the three levels are set to 5 Mbps, 10 Mbps, and 15 Mbps, respectively.

Scenario 2. The corresponding data rates of the three levels are set to 10 Mbps, 20 Mbps, and 30 Mbps, respectively.

The objective behind adding such scenario is to show the behavior of the system regarding user fairness when the target data rate is set to be relatively very high, thus when target data rate is not met.

In Scenario 1, all users succeed in reaching their requested service data rates, and results of Table 1 show a high level of fairness achieved among users requesting the same service. However, when Scenario 2 is applied, no success could be obtained but fairness is still maintained among users. This analysis can show that the proposed algorithm tries to boost and improve the data rates of all users in such a way to still maintain an improved performance at the level of user fairness even when the target data rate is not met.

7.3. Computational Complexity. With the aim of assessing the implementation feasibility of the different proposed schedulers, we measured the computational load of the main allocation techniques to be integrated at the BS.

From a complexity point of view, the proposed scheduling metric WNO PF differs from the conventional PF metric in the weight calculation. For a number of users per subband limited to 2 in NOMA, the number of candidates per subband is $\binom{1}{K} + \binom{2}{K}$. When listing the operations of the proposed allocation technique, we obtain that the proposed metric WNO PF increases the PF computational load by $(26/3)KS + S$ ($\approx O(KS)$) multiplications and $-K^3S + (3/2)K^2S^2 - (4/6)K^2S - (3/6)KS$ ($\approx O((3/2)K^2S^2 - K^3S)$) additions.

In order to compute the PF metric for a candidate user set containing only 1 user, $4 + S$ multiplications and $1 + (3/2)S$ additions are needed. For each candidate user set containing 2 multiplexed users, $13 + 2S$ multiplications and $6 + 3S$ additions are required.

By taking account of the calculations of the terms $h^{-2\alpha}$, h^2 , and $h^2/(N_0B/S)$ performed at the beginning of the allocation process, the classical NOMA PF requires a total of $3KS + C_K^1S(4 + S) + C_K^2S(13 + 2S)$ multiplications, which are equal to $K^2S^2 + (1/2)KS + (13/2)K^2S$ ($\approx O(K^2S^2)$) and $C_K^1S(1 + 3S/2) + C_K^2S(6 + 3S)$ additions, which are equal to $(3/2)K^2S^2 + (1/2)KS + (13/2)K^2S$ ($\approx O(K^2S^2)$). Therefore, we can see that the increase in the number of multiplications is minor in comparison with that of the conventional PF, while the number of additions is almost doubled.

8. Conclusion

In this paper, we have proposed new weighted scheduling schemes for both NOMA and OMA multiplexing techniques. They target maximizing fairness among users, while improving the achieved capacity. Several fair weights designs have been investigated. Simulation results show that the proposed schemes allow a significant increase in the total user throughput and the long-term fairness, when compared to OMA and classic NOMA-based PF scheduler. Combining NOMA with fair weights shows the best performance. Furthermore, the proposed weighted techniques achieve a high level of fairness within each scheduling slot, which improves the QoE of each user. In addition, the proposed weighted metrics give the possibility of delivering different levels of QoS, which can be very useful for certain applications. The study conducted here with two scheduled users per subband can be easily adapted to a larger number of users. Initially developed for single-antenna systems, the proposed scheduling technique can be extended to support multi-antenna systems. Such an extension could be performed using our previous work in [32]. We are currently undergoing further research to reduce the complexity of the PF scheduler by introducing an iterative allocation scheme.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Part of this work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660), which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in FANTASTIC-5G. This work has also been funded with support from the Lebanese University and the French-Lebanese CEDRE program.

References

- [1] 3GPP TS36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description.
- [2] 3GPP TR36.913 (V8.0.0), 3GPP, and TSG RAN, Requirements for further advancements for E-UTRA (LTE-Advanced), 2008.
- [3] 3GPP TR36.814 (V9.0.0), Further Advancements for E-UTRA Physical Layer Aspects, 2010.
- [4] “Ericsson Mobility Report, on the pulse of the networked society,” <https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-june-2017.pdf>, June 2017.
- [5] Y. Chen, A. Bayesteh, Y. Wu et al., “Toward the standardization of non-orthogonal multiple access for next generation wireless networks,” *IEEE Communications Magazine*, vol. 56, no. 3, pp. 19–27, 2018.
- [6] G. Caire and S. Shamai, “On the achievable throughput of a multi-antenna Gaussian broadcast channel,” *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.
- [7] T. Takeda and K. Higuchi, “Enhanced user fairness using non-orthogonal access with SIC in cellular uplink,” in *Proceedings of the IEEE 74th Vehicular Technology Conference, VTC Fall 2011*, September 2011.
- [8] K. Higuchi and Y. Kishiyama, “Non-orthogonal access with random beamforming and intra-beam SIC for cellular mimo downlink,” in *Proceedings of the IEEE 78th Vehicular Technology Conference (VTC '13)*, pp. 1–5, Las Vegas, Nev, USA, September 2013.
- [9] 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 the 2012 9th International Symposium on Wireless Communication Systems, ISWCS 2012*, pp. 476–480, August 2012.
- [10] 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 the IEEE 77th Vehicular Technology Conference (VTC '13)*, pp. 1–5, Dresden, Germany, June 2013.
- [11] J. Umehara, Y. Kishiyama, and K. Higuchi, “Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink,” in *Proceedings of the IEEE 14th International Conference on Communication Systems (ICCS '12)*, pp. 324–328, Singapore, November 2012.
- [12] Sharp corporation, “Evolving RAN towards Rel-12 and beyond, RWS-120039,” in *Proceedings of the 3GPP Workshop on Release 12 Onward*, Ljubljana, Slovenia, 2012.
- [13] H. Xing, Y. Liu, A. Nallanathan, Z. Ding, and H. V. Poor, “Optimal throughput fairness trade-offs for downlink non-orthogonal multiple access over fading channels,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2017.
- [14] Q. Pham and W. Hwang, “ α -Fair resource allocation in non-orthogonal multiple access systems,” *IET Communications*, vol. 12, no. 2, pp. 179–183, 2018.
- [15] J. Schaefferle, “Throughput of a wireless cell using superposition based multiple-access with optimized scheduling,” in *Proceedings of the 2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications, PIMRC 2010*, pp. 212–217, September 2010.
- [16] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, “New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access,” *Wireless Personal Communications*, vol. 87, no. 3, pp. 837–867, 2016.
- [17] E. Okamoto, “An improved proportional fair scheduling in downlink non-orthogonal multiple access system,” in *Proceedings of the 82nd IEEE Vehicular Technology Conference, VTC Fall 2015*, September 2015.
- [18] M. Mehrjoo, M. K. Awad, M. Dianati, and X. S. Shen, “Design of fair weights for heterogeneous traffic scheduling in multichannel wireless networks,” *IEEE Transactions on Communications*, vol. 58, no. 10, pp. 2892–2902, 2010.
- [19] C. Gueguen and S. Baey, “Compensated proportional fair scheduling in multiuser OFDM wireless networks,” in *Proceedings of the IEEE International Conference on Wireless & Mobile Computing, Networking & Communications*, 2008.
- [20] C. Yang, W. Wang, Y. Qian, and X. Zhang, “A weighted proportional fair scheduling to maximize best-effort service utility in multicell network,” in *Proceedings of the 2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, September 2008.
- [21] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge, UK, 2005.
- [22] S. Tomida and K. Higuchi, “Non-orthogonal access with SIC in cellular downlink for user fairness enhancement,” in *Proceedings of the 2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2011)*, pp. 1–6, Chiang Mai, Thailand, December 2011.
- [23] 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 the 2013 21st International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2013*, pp. 770–774, November 2013.
- [24] B. Kimy, S. Lim, H. Kim et al., “Non-orthogonal multiple access in a downlink multiuser beamforming system,” in *Proceedings of the 2013 IEEE Military Communications Conference, MILCOM 2013*, pp. 1278–1283, November 2013.
- [25] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, “System-level performance evaluation of downlink non-orthogonal multiple access (NOMA),” in *Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC*, pp. 611–615, September 2013.
- [26] F. P. Kelly, A. K. Maulloo, and D. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 206–217, 1997.
- [27] P. Viswanath, D. N. Tse, and R. Laroia, “Opportunistic beamforming using dumb antennas,” *Institute of Electrical and*

Electronics Engineers Transactions on Information Theory, vol. 48, no. 6, pp. 1277–1294, 2002.

- [28] M. Kountouris and D. Gesbert, “Memory-based opportunistic multi-user beamforming,” in *Proceedings of the International Symposium on Information Theory, ISIT 2005*, pp. 1426–1430, Adelaide, Australia, September 2005.
- [29] 3GPP and TR25-814 (V7.1.0), Physical Layer Aspects for Evolved UTRA, 2006.
- [30] 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 the IEEE Globecom Workshops*, pp. 66–70, Atlanta, Ga, USA, December 2013.
- [31] M. Dianati, X. Shen, and S. Naik, “A new fairness index for radio resource allocation in wireless networks,” in *Proceedings of the 2005 IEEE Wireless Communications and Networking Conference, WCNC 2005: Broadband Wirelss for the Masses—Ready for Take-off*, pp. 712–717, March 2005.
- [32] M.-J. Youssef, J. Farah, C. A. Nour, and C. Douillard, “Waterfilling-based resource allocation techniques in downlink Non-Orthogonal Multiple Access (NOMA) with single-user MIMO,” in *Proceedings of the 2017 IEEE Symposium on Computers and Communications, ISCC 2017*, pp. 499–506, July 2017.

Research Article

NOMA for Multinumerology OFDM Systems

Ayman T. Abusabah ¹ and Huseyin Arslan ^{1,2}

¹*School of Engineering and Natural Sciences, Istanbul Medipol University, 34810 Istanbul, Turkey*

²*Department of Electrical Engineering, University of South Florida, Tampa, FL 33620, USA*

Correspondence should be addressed to Ayman T. Abusabah; asabah@st.medipol.edu.tr

Received 23 November 2017; Revised 21 February 2018; Accepted 28 March 2018; Published 9 May 2018

Academic Editor: Oğuz Kucur

Copyright © 2018 Ayman T. Abusabah and Huseyin Arslan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nonorthogonal multiple access (NOMA) is a promising technique which outperforms the traditional multiple access schemes in many aspects. It uses superposition coding (SC) to share the available resources among the users and adopts successive interference cancellation (SIC) for multiuser detection (MUD). Detection is performed in power domain where fairness can be supported through appropriate power allocation. Since power domain NOMA utilizes SC at the transmitter and SIC at the receiver, users cannot achieve equal rates and experience higher interference. In this paper, a novel NOMA scheme is proposed for multinumerology orthogonal frequency division multiplexing system, that is, different subcarrier spacings. The scheme uses the nature of mixed numerology systems to reduce the constraints associated with the MUD operation. This scheme not only enhances the fairness among the users but improves the bit error rate performance as well. Although the proposed scheme is less spectrally efficient than conventional NOMA schemes, it is still more spectrally efficient than orthogonal multiple access schemes.

1. Introduction

Cumulative and incessant demands on new services and applications, in addition to the great expansion in the number of connected devices, have led to a huge data traffic explosion and appearance of different application classes and classifications [1]. Thus, a strong need to boost the expected high data traffic has been recently emerged. Also, since it became obvious that next generation has to support high data rates, applications industry and academia agreed on the necessity of new and flexible radio access technologies (RATs) [2].

Multiple access (MA) techniques play a major role in the overall communication process. Since the first generation up until the fourth generation, MA techniques were distributing the users over the available resources orthogonally. For instance, time division multiple access, frequency division multiple access, code division multiple access, and orthogonal frequency division multiple access (OFDMA) are considered as orthogonal multiple access (OMA) techniques. Eventually, researchers have moved away from utilizing the resources in an orthogonal way to a nonorthogonal one, which they call nonorthogonal multiple access (NOMA) [3].

Power domain NOMA adopts multiplexing multiple users through sharing the same resources at the transmitter (TX). Furthermore, it uses successive interference cancellation (SIC) as a multiuser detection technique to separate the users through power differences at the receiver (RX) side. NOMA is considered as a promising multiple access technique for the fifth-generation wireless networks due to massive connectivity, low latency, and high spectral efficiency (SE) [4].

It is shown that, by adopting NOMA with orthogonal frequency division multiplexing- (OFDM-) based method as RAT, multiple users can be allocated on a subcarrier at the same time. However, the cochannel interference (CCI) per subcarrier increases as more users are multiplexed on the same subcarrier, which degrades the system performance [5]. In [6], the authors have studied users' pairing; then they concluded that NOMA outperforms OMA especially with users whose channel conditions are more distinctive, which is practically difficult to occur.

Power control/allocation has been studied in many works [7]. To guarantee the fairness among NOMA users, more power is required for users with poor channel conditions

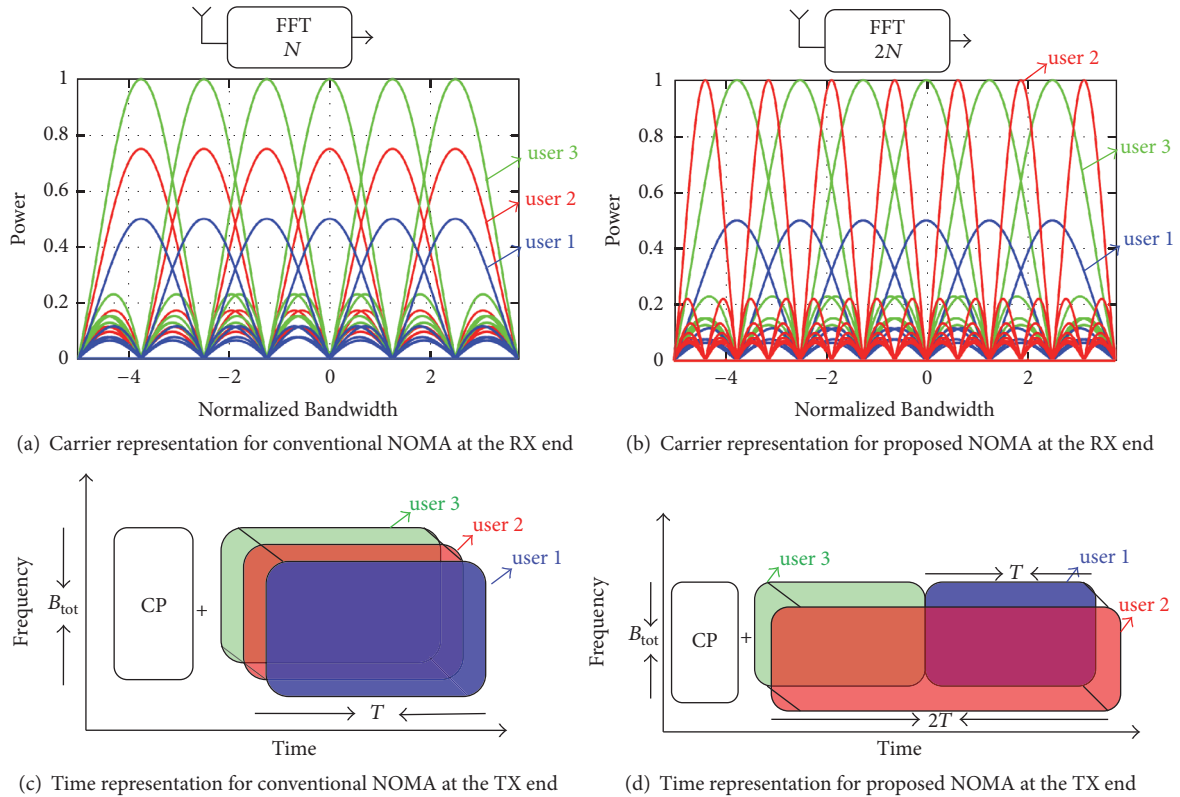


FIGURE 1: Signals superposition in time and frequency domains.

and less power for users with better channel conditions [8]. However, if the users have similar channel conditions, OMA can guarantee better fairness and conventional power domain NOMA cannot strictly guarantee the users' quality of service (QoS) targets [9], which could be critical for some scenarios with strict fairness constraints.

In this work, we propose an OFDM based NOMA scheme. The scheme utilizes the numerology concept to reduce the constraints associated with the conventional NOMA schemes. Simply, the users utilize different subcarrier spacings, wide subcarriers, and narrow subcarriers. A three-user scenario is considered in this paper. In the conventional scheme, the three users are supposed to share the same subcarriers as shown in Figure 1(a), while in the proposed scheme one user is assigned narrower and less frequently spaced subcarriers as illustrated in Figure 1(b).

The subcarriers configuration of the proposed scheme is characterized by the fact that the wide subcarrier users are fully overlapped within the same wide subcarriers and make a zero crossing at the peaks of the other wide subcarriers. Furthermore, the narrow subcarriers do not impose any interference at the peaks of wide subcarriers, that is, by avoiding the transmission on the half of narrow subcarriers. As a result, the wide subcarriers are not affected by any external interference. On the other hand, an interference is imposed by the tails of the wide subcarriers on the peaks of the narrow subcarriers.

Even though the narrow subcarriers share the bandwidth with the wide subcarriers, the detection, of wide subcarrier users, is independent of narrow subcarrier users; therefore, SIC can be used to detect the wide subcarrier users based on their power differences. On the other hand, the narrow subcarriers are detected once the interference imposed by wide subcarriers is eliminated.

By assigning one of the users narrower subcarriers, we reduce the amount of CCI imposed on the wide subcarrier users. Furthermore, the interference imposed on the narrow subcarrier user can be easily canceled by eliminating the wide subcarrier signals, which enhances the bit error rate (BER) performance for each user.

The narrow subcarrier user has an extra degree of freedom (DOF) as its power level is independent of the detection process; that is, it is not restricted to the wide subcarrier users and does not affect their detection process. In other words, the SIC process does not depend on the power level of the narrow subcarriers which grants more flexibility for power assigning. Based on that, the proposed scheme has the advantage of providing a fairer rate allocation to the users compared with the conventional scheme especially when the channel conditions of the users are similar.

The subcarriers configuration of Figure 1(b) can be simply accomplished by composing the symbols of wide subcarrier users synchronously in the time domain, that is, each with a length of one OFDM symbol slot. Then, the extended

symbol, that is, narrow subcarrier user, is added. However, a novel structure is adopted for the transmission. As shown in Figure 1(d), the OFDM symbol slots of wide subcarrier users are orthogonally constructed and then the extended symbol of narrow subcarrier user is added. In this case, the wide subcarrier users are composed at the RX end and the proposed subcarrier configuration is obtained. In both transmission structures, the resulting signal consists of two OFDM symbol slots and the utilized resources are the same.

To achieve this purpose, a fast Fourier transform (FFT) operation, with the length of two OFDM symbol slots, is performed at the RX end. By doing this, the wide subcarrier users are multiplexed at the RX end. Extending the length of FFT window at the RX end allows us to equalize the channel using one cyclic prefix (CP) [10], which is a good solution to increase the SE even with an absolute OFDMA system.

The rest of the paper is organized as follows: Section 2 presents the signal configuration of the proposed scheme. Problem description, features, and potentials for our technique are provided in Section 3. The system model for conventional multicarrier NOMA scheme is given in Section 4. The design analysis is provided in Section 5. Simulation results are shown in Section 6. Finally, Section 7 concludes our paper.

2. Signal Configuration

The frequency and time representations, for conventional NOMA scheme, are shown in Figures 1(a) and 1(c), respectively. Three different power signals are multiplexed utilizing the same resources, where each signal constitutes one OFDM symbol slot; then a CP is appended. Therefore, the SE is expected to be doubled 3 times compared with OFDMA system without considering the CP redundancy.

On the other hand, in proposed NOMA, one of the users (user-2) is assigned narrower and less frequent subcarriers as illustrated in Figure 1(b), which simply multiplies the symbol duration by two as depicted in Figure 1(d). Since the new structure uses only T seconds out of a possible $2T$, the SE, defined as bps/Hz, is halved for the wide subcarrier users. Meanwhile, as the total energy budget is the same, the power used by wide subcarrier users is twice as before. As the SE increases logarithmically with power, this increase does not compensate for the .5 loss in the SE; thus, the overall SE improvement over OFDMA is greater than 1 but strictly less than 1.5.

As mentioned earlier, the subcarriers' configuration in Figure 1(b) can be obtained if both symbols of user 1 and user 3 share the first half of user 2 symbol. However, in Figure 1(d), user 1 and user 3 are constructed orthogonally at the TX. In this case, the subcarriers' configuration of the proposed scheme is accomplished at the RX end where the wide subcarrier users (user 1 and user 3) are multiplexed by adopting an extended FFT operation.

In Figure 1(d), user 3 symbol is multiplexed with the first half of user 2 symbol and user 1 does the same with the second half. This can be seen from Figure 1(b) as well, where user 2 has a contribution from user 1 and user 3 at the peaks. Note that, user 1 and user 3 are orthogonal with respect to each

other at the TX side. However, processing them together at the RX side, that is, by using an extended FFT window, makes them overlapping and therefore nonorthogonal. As the basic NOMA concept based on multiplexing the users at the TX utilizing the same resources, the proposed scheme can be considered as a half NOMA.

The key advantage of using larger FFT window size at RX is the capability of equalizing the channel using one CP for the whole OFDM symbols and thus increasing the SE. This property can be even used for a pure OFDMA system. For example, in the absence of user 2 in Figure 1(d), the system becomes an OFDMA system. Then, if FFT operation, with the length of two OFDM symbols, is performed at the RX, the transmitted symbols are composed and one CP can be used for equalization rather than two. In our case, the composed signals are detected in power domain and user 2 is able to share the other users' resources without introducing any extra interference.

Note that user 2 signal (narrow subcarriers) makes a zero crossing with the other signals (wide subcarriers) at the peaks. This can be clearly concluded from frequency domain representation. The same result is not obvious from time domain representation. In Section 5, it is proven mathematically that, by adopting an extended FFT window size at the RX end, the wide subcarrier users are multiplexed although they are assigned different OFDM symbol slots at the TX and the narrow subcarrier users do not affect their detection process.

3. Problem Description

To describe the features of our proposed scheme, two scenarios are considered. The first scenario appears in Figure 2(a) where three downlink users have distinctive channel conditions $|h|$, while the second scenario appears in Figure 2(b), where two users have similar channel conditions; that is, two users are at the cell edge and one user is close to the base station (BS).

In the case of the users whose channel conditions are similar, like the second scenario, either we assign similar power allocation (PA) to achieve the fairness and, therefore, SIC cannot work properly due to its inherent nature which depends on power differences for separation, or we assign different PA which leads explicitly to unfairness distribution.

3.1. Conventional NOMA (CN). If users have distinctive $|h|$ as represented in Figure 2(a), under perfect channel state information assumption at the BS, achieving the fairness can be ensured through a proper PA. Furthermore, degradation in the performance due to the number of assigned users is expected. For instance, user 3 signal has to be detected with the presence of user 2 and user 1 signals by considering them as a noise, which degrades the BER performance.

In the second scenario, user 2 and user 3 experience the same channel effect. Therefore, if the fairness is a system requirement, both users have to be assigned similar PA. However, the natural work of SIC depends on the power differences to facilitate the separation process. Thus, with similar PA, the internal interference cannot be avoided and the performance can be extremely degraded.

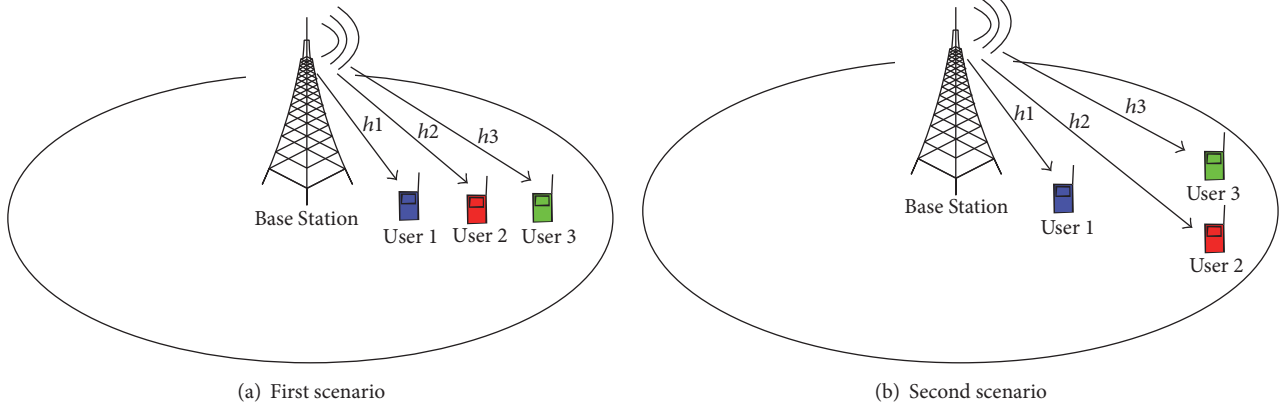


FIGURE 2: Two different user distribution scenarios: (a) $|h_{1,w}| > |h_{2,w}| > |h_{3,w}|$ and (b) $|h_{1,w}| > (|h_{2,w}| = |h_{3,w}|)$.

3.2. *Proposed NOMA (PN)*. According to the first scenario, by assigning user 2 narrow subcarriers, we reduce the interference imposed on user 1 and user 3 signals. Thus, SIC process becomes easier since it has to differentiate two signals rather than three based on their power differences. Actually, the narrow subcarrier user is selected so that the other users, that is, wide subcarrier users, obtain more distinctive channel conditions. Besides, the PA of user 2 is determined independently which grants more flexibility for the system design.

This becomes very beneficial for the second scenario where achieving the fairness is an issue. As mentioned before, user 2 and user 3 cannot be paired as they have similar channel conditions. Nevertheless, if user 2 is assigned narrow subcarriers, user 3 and user 1 can be paired and a proper PA is determined. Naturally, the PA of user 2 is not restricted by the other users.

4. Conventional NOMA System Model

Conventional multicarrier downlink NOMA system is formulated by considering I downlink users around a BS as shown in Figure 2. Users are distributed randomly and served by one BS and the total bandwidth B_{tot} consists of N_{sc} number of orthogonal subcarriers in frequency domain. Transceivers are supposed to be equipped with one antenna, and I users share N_{sc} OFDM subcarriers through superposition coding.

The BS assigns different power levels depending on users' conditions in order to achieve the fairness in the system and provide the capability of detection at the RX side using SIC. So, high power is assigned to the poor users, that is, far users, and low power to ones whose channel conditions are good. In other words, NOMA exploits the heterogeneity of users' distribution and then allows the separation in the power domain [11].

The BS is transmitting the signal $x_{i,w}$ to the i th user ($i = \{1, 2, \dots, I\}$) on the w th subcarrier ($w = \{1, 2, \dots, N_{\text{sc}}\}$) with transmission power $P_{i,w}$, and, then, the received signal by user i on subcarrier w is given by [12] as follows:

$$y_{i,w} = h_{i,w} \sum_{u=1}^I \sqrt{P_{u,w}} x_{u,w} + z_{i,w}, \quad (1)$$

where $z_{i,w}$ represents the additive white Gaussian noise (AWGN) for the i th user on subcarrier w with a zero mean and $\sigma_{i,w}^2$ variance, that is, $z_{i,w} \sim \mathcal{N}(0, \sigma_{i,w}^2)$, and $h_{i,w}$ denotes the channel gain between the BS and the received user at the w th subcarrier including both effects of large and small scale fading. Path loss and shadowing are considered as small fading effects. On the other hand, block Rayleigh is adopted for large-scale fading.

Without loss of generality, the channels are sorted as $|h_{1,w}|^2 > |h_{2,w}|^2 > \dots > |h_{i,w}|^2 > \dots > |h_{I,w}|^2 > 0$. For a given subcarrier, a user who enjoys a better downlink channel quality can decode and remove the CCI from a user who has a worse downlink channel quality by employing SIC [5]; thus, user i enjoys a better channel quality than user $(i + 1)$. At the i th user, if SIC is carried out perfectly, then achieving the fairness follows Shannon's equation [13], where the achievable rate of the i th user for B_{tot} Hz system bandwidth at the RX side is given by

$$R_i = B_{\text{sc}} \sum_{w=1}^{N_{\text{sc}}} \log_2(1 + \text{SINR}_{i,w}), \quad (2)$$

$$\text{SINR}_{i,w} = \left(\frac{P_{i,w} |h_{i,w}|^2}{|h_{i,w}|^2 \sum_{u=1}^{i-1} P_{u,w} + \sigma_{i,w}^2} \right),$$

where $\text{SINR}_{i,w}$ is the instantaneous signal-to-interference-plus-noise ratio by user i on the w th subcarrier and $B_{\text{sc}} = B_{\text{tot}}/N_{\text{sc}}$ is the subcarrier bandwidth. Note that if the strongest user, that is, user 1, decodes and cancels all other users' signals successively, then the achievable data rate is given by $R_1 = B_{\text{sc}} \sum_{w=1}^{N_{\text{sc}}} \log_2(1 + P_{1,w} |h_{1,w}|^2 / \sigma_{1,w}^2)$.

5. Signal Representation and Design Analysis

This section considers the design analysis of our proposed scheme. The mathematical model of the proposed scheme is established. The time domain signal in Figure 1(d) is formulated. Then, it is shown that, by adopting FFT operation with a length of two OFDM symbols at the RX side, the wide subcarrier users are composed although they are orthogonally transmitted.

Generation of wide subcarrier signals can be done using an inverse fast Fourier transform (IFFT) process with a length of N samples. On the other hand, narrow subcarrier signals can be also generated using IFFT process with a length of $M = QN$ samples, where Q is the ratio between the two different subcarrier spacings or the two different symbol lengths.

According to Figure 1(b), user 1 and user 3 signals are generated using IFFT with N points, while user 2 signal is generated utilizing IFFT with $M = 2N$ points. It is worth mentioning that, to avoid direct interference with wide subcarriers, we do not use all of the narrow subcarriers for transmission. In our example, narrow-odd subcarriers are used for user 2 data transmission while narrow-even ones are filled with zeros using subcarrier mapping (SM).

An OFDM transmission symbol, of wide subcarrier user, is given by the N point complex modulation sequence

$$\begin{aligned} x_{w_a}(n) &= \sqrt{P\alpha_a} \text{IFFT}(X_{w_a}) \\ &= \frac{1}{N} \sqrt{P\alpha_a} \sum_{k=0}^{N-1} X_{w_a}(k) \cdot e^{j2\pi nk/N}, \end{aligned} \quad (3)$$

for $n = 0, 1, \dots, N-1$, $a = 0, 1, \dots, A$,

where $X_{w_a}(k)$ is the complex modulated symbol of a th user on k th subcarrier; that is, A is the number of wide subcarrier users, α_a is the assigned PA factor to the a th user, and $\sum_{a=1}^A P\alpha_a$ is the amount of power that is specified for wide subcarrier users.

In a similar way, the OFDM transmission symbol, for narrow subcarrier user, is given by the M point complex modulation sequence

$$\begin{aligned} x_{nr_b}(m) &= \sqrt{P\beta_b} \text{IFFT}(\widehat{X}_b) \\ &= \frac{1}{M} \sqrt{P\beta_b} \sum_{l=0}^{M-1} \widehat{X}_b(l) \cdot e^{j2\pi ml/M}, \end{aligned} \quad (4)$$

for $m = 0, 1, \dots, M-1$, $b = 0, 1, \dots, B$.

$$\begin{aligned} \widehat{X}_b(l) &= \begin{cases} X_{nr_b} \left(\frac{l-1}{Q} \right), & l = Qk+1, (l = 1, 3, \dots, M-1) \\ 0, & \text{o.w.,} \end{cases} \end{aligned} \quad (5)$$

where $\widehat{X}_b(l)$ is the complex modulated symbol of b th user on l th subcarrier after SM; that is, B is the number of narrow subcarrier users, β_b is the assigned PA factor to the b th user, $\sum_{b=1}^B P\beta_b$ is the amount of power that is specified for narrow subcarrier users, $I = A + B$ is the total number of users, and the maximum assigned power from the BS to all users is $P = \sum_{a=1}^A P\alpha_a + \sum_{b=1}^B P\beta_b$; that is, $B = 0$ represents the conventional NOMA scheme case.

The wide subcarrier signals (x_w) are assigned different time slots to form one block xx_w with a length of M samples. Thereafter, the narrow subcarrier signals (x_{nr}) are added

together forming another block xx_{nr} which is already with a length of M samples. Finally, the resultant blocks are added synchronously to produce one block s with a length of M samples for the transmission; this process can be expressed as follows:

$$s = xx_w + xx_{nr},$$

$$xx_w = [x_{w_1}, \dots, x_{w_A}]_{M=AN}, \quad xx_{nr} = \sum_{b=1}^B x_{nr_b}. \quad (6)$$

According to the proposed scheme in Figure 1, $A = 2$, $B = 1$, $M = 2N$, and $Q = 2$. So (6) can be written as follows:

$$\begin{aligned} xx_w &= [x_{w_1}, x_{w_2}]_{2N}, \\ xx_{nr} &= [x_{nr_1}]_{M=2N}. \end{aligned} \quad (7)$$

By assuming $m = n + qN$, (4) can be expressed as follows:

$$\begin{aligned} x_{nr}(n + qN) &= \frac{1}{M} \sum_{k=0}^{N-1} \sqrt{P\beta_1} X_{nr}(Qk+1) \\ &\quad \cdot e^{j2\pi(n+qN)(Qk+1)/M}, \end{aligned} \quad (8)$$

for $n = 0, 1, \dots, N-1$, $q = 0, 1, \dots, Q-1$.

Since $Q = 2$, the first half and second half of x_{nr} signal can be represented by setting q to 0 and 1, respectively:

$$\begin{aligned} x_{nr}(n) &= \frac{1}{M} \sum_{k=0}^{N-1} \sqrt{P\beta_1} X_{nr}(Qk+1) \\ &\quad \cdot e^{j2\pi n(Qk+1)/M}, \quad \text{for } q = 0, \\ x_{nr}(n + N) &= \frac{1}{M} \sum_{k=0}^{N-1} \sqrt{P\beta_1} X_{nr}(Qk+1) \\ &\quad \cdot e^{j2\pi(n+N)(Qk+1)/M}, \quad \text{for } q = 1, \\ x_{nr}(n + N) &= -x_{nr}(n). \end{aligned} \quad (9)$$

Thus, the second half of the signal x_{nr} is just a reversal copy of the first half because of odd subcarriers usage. From (9), the transmitted signal s in (6) can be expressed as follows:

$$s(r) = \begin{cases} x_{w_1}(r) + x_{nr_1}(r), & 0 < r < N-1 \\ x_{w_2}(r) - x_{nr_1}(r), & N < r < 2N-1, \end{cases} \quad (10)$$

where r represents the composed signal sample index ($r = \{0, 1, \dots, M-1\}$).

After composing the signals, to avoid intersymbol interference and enable frequency domain equalization (FDE) at the RX, a copy from the resultant tail is appended as a CP where its duration has to be larger than the maximum excess delay of the channel.

At the RX end and after removing the CP, FFT operation, with a length of $M = 2N$ points, is performed as follows:

$$S(v) = \sum_{r=0}^{M-1} s(r) \cdot e^{-j2\pi rv/M} \quad \text{for } v = 0, 1, \dots, M-1, \quad (11)$$

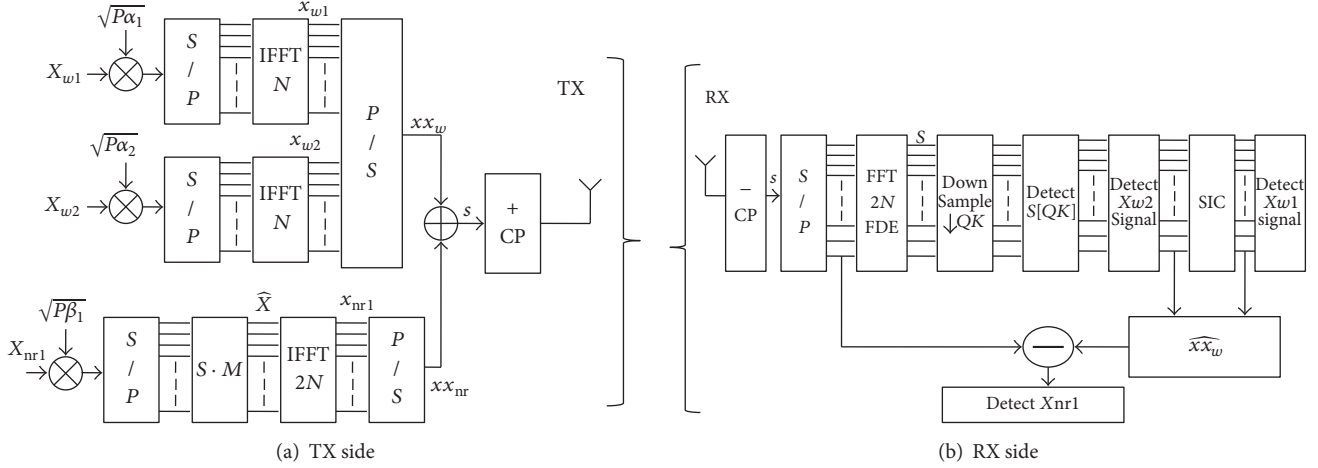


FIGURE 3: Transceiver design for proposed NOMA scheme adopting three downlink users.

where S are the received complex symbols after FFT operation. Afterwards, FDE takes the responsibility to get rid of channel's sparsity where single tap equalization is available. To compute the output on the even and odd subcarriers, we assume that $v = Qk + q$, and, then, (11) can be represented as

$$\begin{aligned}
 S(Qk + q) &= \sum_{r=0}^{N-1} (x_{w_1}(r) + x_{nr_1}(r)) \cdot e^{-j2\pi r(Qk+q)/QN} \\
 &+ \sum_{r=N}^{2N-1} (x_{w_2}(r) - x_{nr_1}(r)) \\
 &\cdot e^{-j2\pi r(Qk+q)/QN}.
 \end{aligned} \quad (12)$$

By setting $q = 0$ and assuming $z = r - N$ for the second part of (12), then, the output on even subcarriers is proven to be

$$S(Qk) = \sqrt{P\alpha_1}X_{w_1}(k) + \sqrt{P\alpha_2}X_{w_2}(k). \quad (13)$$

Explicitly, (13) proves that although X_{w_1} and X_{w_2} signals are constructed orthogonally at the TX, they are multiplexed by utilizing larger FFT window at the RX side. In addition, (13) ensures the absence of narrow subcarriers contribution to wide subcarriers; thereafter, SIC separates the wide subcarrier signals based on power differences.

User 1 and user 3 signals are constructed again; then, the reconstructed signal \widehat{xx}_w is subtracted from the received signal s with a view to detect user 2 signal. Transceiver block diagram is given in Figure 3 for the proposed NOMA scheme.

6. Performance Evaluation

In this section, we evaluate the performance of proposed NOMA scheme through simulation. System parameters are presented in Table 1.

6.1. BER. The BER performance is supposed to be enhanced by performing proposed NOMA due to many reasons. Mainly, wide subcarrier users experience lower number

TABLE 1: Simulation parameters.

Parameter name	Value
Number of wide subcarriers (N)	64
Number of narrow subcarriers (M)	128
Modulation type	QPSK
Total system bandwidth	5 MHz
Total power	10 dBm
The first scenario $ h_{i,w} ^2/\sigma_{i,w}^2$	20 dB, 17 dB, 0 dB for $i = 1, 2, 3$
The second scenario $ h_{i,w} ^2/\sigma_{i,w}^2$	20 dB, 0 dB, 0 dB, for $i = 1, 2, 3$

of interferer users. Moreover, the interference imposed on narrow subcarrier user can be eliminated by detecting and canceling wide subcarrier users and utilizing the unused narrow subcarriers. Furthermore, the narrow subcarrier user can enjoy any power level. Thus, the BER performance is enhanced significantly. Note that the SINR values are considered at the RX side after FFT process.

The $\text{SINR}_{i,w}^c$ values utilizing conventional NOMA scheme for user 1, user 2, and user 3 on subcarrier w , with successful decoding and no error propagation assumption, are given by

$$\begin{aligned}
 \text{SINR}_{1,w}^c &= \left(\frac{\alpha_1 P |h_{1,w}|^2}{\sigma_{1,w}^2} \right), \\
 \text{SINR}_{2,w}^c &= \left(\frac{\alpha_2 P |h_{2,w}|^2}{|h_{2,w}|^2 \alpha_1 P + \sigma_{2,w}^2} \right), \\
 \text{SINR}_{3,w}^c &= \left(\frac{\alpha_3 P |h_{3,w}|^2}{|h_{3,w}|^2 (\alpha_1 + \alpha_2) P + \sigma_{3,w}^2} \right),
 \end{aligned} \quad (14)$$

while $\text{SINR}_{i,w}^p$ values utilizing proposed NOMA scheme are expressed as follows:

$$\text{SINR}_{1,w}^p = \left(\frac{\alpha_1 P |h_{1,w}|^2}{\sigma_{1,w}^2} \right),$$

$$\begin{aligned} \text{SINR}_{2,w}^P &= \left(\frac{\beta_1 P |h_{2,w}|^2}{\sigma_{2,w}^2} \right), \\ \text{SINR}_{3,w}^P &= \left(\frac{\alpha_2 P |h_{3,w}|^2}{|h_{3,w}|^2 \alpha_1 P + \sigma_{3,w}^2} \right). \end{aligned} \quad (15)$$

Using the same PA for both schemes, we can notice from (14) and (15) that the first user experiences the same SINR values, while a big enhancement, in SINR values, is noticeable for the second and third user utilizing proposed NOMA.

6.2. Fairness Factor (F). To evaluate the fairness level for conventional and proposed NOMA we define the factor F as in [14], where F measures the equality of users' rate R for a given system and it is given by

$$F = \frac{\left(\sum_{i=1}^I R_i \right)^2}{I \sum_{i=1}^I (R_i)^2}. \quad (16)$$

For instance, if all users get the same amount of R , then the value F will be close to 1.

The goal of PA mechanism is to maximize the sum capacity under a fairness constraint for NOMA systems. The optimization problem is formulated as

$$\begin{aligned} \max_{\alpha_a, \beta_b} \quad & B_{\text{sc}} \sum_{i=1}^I \sum_{w=1}^{N_{\text{sc}}} \log_2 (1 + \text{SINR}_{i,w}) \\ \text{s.t.} \quad & \sum_{i=1}^I \sum_{w=1}^{N_{\text{sc}}} P_{i,w} \leq P \\ & P_{i,w} \geq 0, \quad \forall i, \forall w \\ & F = \bar{F}, \end{aligned} \quad (17)$$

where \bar{F} is the target fairness index in the network. The PA coefficients (α_a, β_b) are obtained through exhaustive search using algorithm 1 in [15].

According to the first scenario, the optimal PA coefficients utilizing conventional NOMA and proposed NOMA schemes are equal to $[\alpha_1 \alpha_2 \alpha_3] = [0.01 \ 0.15 \ 0.84]$ and $[\alpha_1 \beta_2 \alpha_2] = [0.08 \ 0.36 \ 0.56]$, respectively. On the other hand, based on the second scenario, the optimal PA coefficients utilizing conventional NOMA and proposed NOMA are found to be $[\alpha_1 \alpha_2 \alpha_3] = [0.02 \ 0.39 \ 0.59]$ and $[\alpha_1 \beta_2 \alpha_2] = [0.13 \ 0.34 \ 0.53]$, respectively.

Based on the obtained optimal PA coefficients, the BER performance is evaluated. The normalized channel gains ($|h_{i,w}|^2 / \sigma_{i,w}^2$) are set as in Table 1 and the fairness index \bar{F} is assumed to be 0.7. The individual BER for the first and the second scenarios are shown in Figures 4 and 5, respectively. Using the optimal PAs obtained for the second scenario, the fairness level, of conventional and proposed NOMA schemes, is evaluated as depicted in Figure 6. The results show clear dominance of proposed NOMA over conventional NOMA in terms of BER and fairness level.

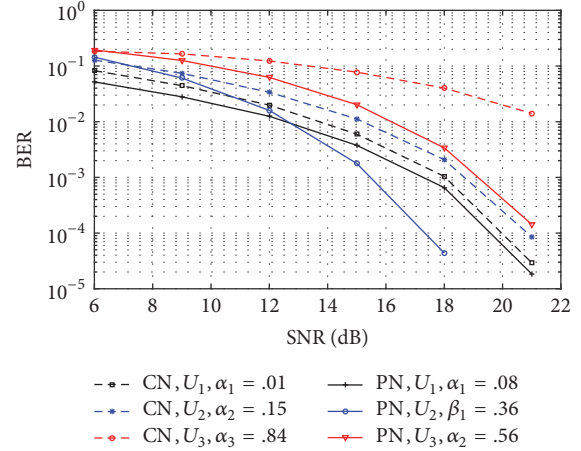


FIGURE 4: BER performance utilizing the optimum PA for conventional and proposed NOMA in the first scenario, $\bar{F} = 0.7$.

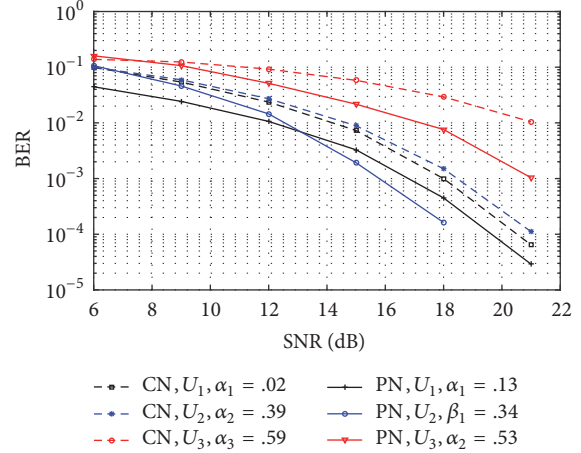


FIGURE 5: BER performance utilizing the optimum PA for conventional and proposed NOMA in the second scenario, $\bar{F} = 0.7$.

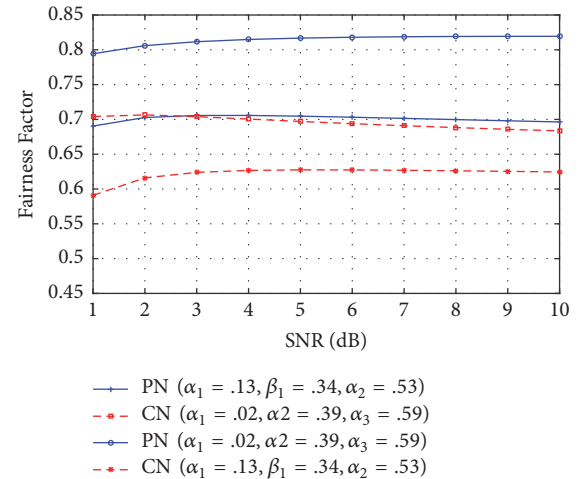


FIGURE 6: Fairness level of conventional and proposed NOMA utilizing the optimum PA for the second scenario.

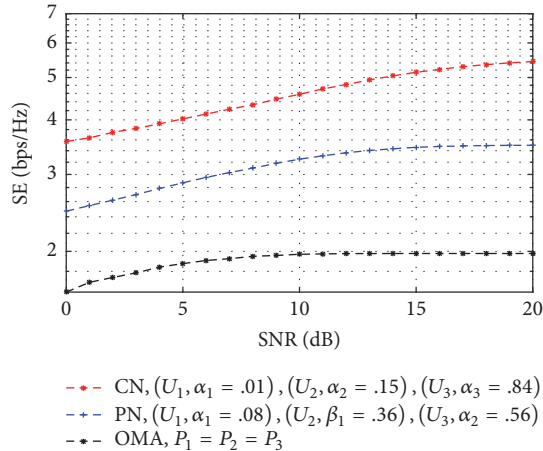


FIGURE 7: Average spectral efficiency adopting conventional NOMA, proposed NOMA, and OMA using the optimum PA for the first scenario.

6.3. Spectral Efficiency (SE). The spectral efficiency $\eta_i = R_i/B_{\text{tot}}$ represents the amount of the carried data over the available resources. In [16], it is proven that the conventional NOMA schemes can offer a better SE than OMA. In conventional NOMA, the SE is expected to improve dramatically since three users utilize the available resources, that is, time and frequency units. Unfortunately, this comes at a price of unguaranteed QoS and loss of fairness. On the other hand, although proposed NOMA offers less SE than conventional NOMA due to additional time unit usage, that is, two time slots, it still provides superior SE compared with OMA. In particular, the superior SE in proposed NOMA arises from the capability of the narrow subcarrier user to share the available resources in addition to CP reduction between OFDM symbols.

The first scenario in Figure 2(a) is considered in the evaluation of the average SE for OMA, conventional NOMA, and proposed NOMA. For OMA, the bandwidth B_{tot} and the total power P are split equally between the assigned users. The SE performance is illustrated in Figure 7.

It is obvious from Figure 7 that the SE of proposed NOMA is decreased compared to conventional NOMA; however, it is still more spectrally efficient than OMA.

7. Conclusion

In this work, some of the NOMA-OFDM system based problems have been addressed by employing numerology concept cleverly. More DOF are given to one of the composed users by assigning narrower subcarriers. Based on this DOF, the constraints associated with the conventional NOMA schemes have been reduced and the BER performance has been improved too. Furthermore, the proposed method has proven its superiority in affording a fairer rate allocation to the users compared to conventional method.

A new methodology of multiplexing the users, in power domain, has been implemented by adopting a larger FFT window at the RX end. As a result, the guard durations

between OFDM symbols became unnecessary. Thus, our proposed scheme is more spectrally efficient than OMA schemes.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank Khawar Hussain, Research Assistant, for assistance and comments that greatly improved the manuscript.

References

- [1] U. L. Rohde, A. K. Poddar, I. Eisele, and E. Rubiola, "Next generation 5G radio communication NW," in *Proceedings of the 2017 Joint Conference of the European Frequency and Time Forum and IEEE International Frequency Control Symposium ((EFTF/IFC))*, pp. 113–116, Besançon, France, July 2017.
- [2] J. G. Andrews, S. Buzzi, and W. Choi, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.
- [3] L. Dai, B. Wang, Y. Yuan, S. Han, C. 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.
- [4] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the Spectral Efficiency and Security Enhancements of NOMA Assisted Multicast-Unicast Streaming," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 3151–3163, 2017.
- [5] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1177–1191, 2017.
- [6] 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.
- [7] 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.
- [8] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.
- [9] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.
- [10] J.-Y. Chouinard, X. Wang, and Y. Wu, "MSE-OFDM: A new OFDM transmission technique with improved system performance," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05*, pp. III865–III868, USA, March 2005.
- [11] 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 the 2013 21st International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2013, pp. 770–774, Japan, November 2013.

- [12] 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.
- [13] C. E. Shannon, “Communication in the Presence of Noise,” *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.
- [14] R. K. Jain, D. M. W. Chiu, and W. R. Hawe, “A quantitative measure of fairness and discrimination for resource allocation in shared computer system,” vol. 38, pp. 20–21, Eastern Research Laboratory, Digital Equipment Corporation, Hudson, Mass, USA, 1984.
- [15] T. Manglayev, R. C. Kizilirmak, and Y. H. Kho, “Optimum power allocation for non-orthogonal multiple access (NOMA),” in *Proceedings of the 10th IEEE International Conference on Application of Information and Communication Technologies, AICT 2016*, Azerbaijan, October 2016.
- [16] P. Sedtheetorn and T. Chulajata, “Spectral efficiency evaluation for non-orthogonal multiple access in Rayleigh fading,” in *Proceedings of the 18th International Conference on Advanced Communications Technology, ICACT 2016*, pp. 747–750, kor, February 2016.

Research Article

Nonuniform Code Multiple Access

Cheng Yan ^{1,2}, Ningbo Zhang ^{1,2} and Guixia Kang ^{1,2}

¹Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Beijing, China

²Wuxi BUPT Sensory Technology and Industry Institute Co., Ltd., Wuxi, China

Correspondence should be addressed to Guixia Kang; gxkang@bupt.edu.cn

Received 21 October 2017; Accepted 13 March 2018; Published 18 April 2018

Academic Editor: Güneş K. Kurt

Copyright © 2018 Cheng Yan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For sparse code multiple access advanced (SCMAA), the quality of initial information on each resource node and the convergence reliability of the detected user in each decision process were unsatisfactory at the message passing algorithm (MPA) receiver. Driven by these problems, this paper proposes a nonuniform code multiple access (NCMA) scheme. In the codebook design of NCMA, different transmitted layers are generated from different complex multidimension constellations, respectively, and a novel basic complex multidimension constellation design is proposed to increase the minimum intrapartition distance. Then a novel criterion of permutation set is proposed to maximize the sum of distances between interfering dimensions of transmitted codewords multiplexed on any resource node, where the number of nonzero elements of transmitted codewords is more than 1. On the other side, an advanced MPA receiver is proposed to improve the reliability of detection on each transmitted layer of NCMA. Simulation results show that the block error rate performance of NCMA outperforms SCMAA and sparse code multiple access (SCMA) under the same spectral efficiency.

1. Introduction

Higher spectral efficiency is one of main requirements in future 5G system [1]. Compared with 4G system, future 5G system improves spectral efficiency by 5~15 times [1]. Driven by this requirement, nonorthogonal multiple access, such as sparse code multiple access (SCMA), is proposed. SCMA [2–5] was a multidimension codebook-based nonorthogonal multiple access [5, 6]. In SCMA, there were J transmitted layers multiplexed on K resource nodes. Each layer (a transmitted layer represents a transmitted user) had its dedicated codebook. A codebook contained a plurality of K -dimension codewords [3, 4]. A K -dimension codeword was a sparse column vector, where there were $N < K$ nonzero elements, and was generated from a complex N -dimension constellation point by a binary mapping matrix. In order to improve spectral efficiency, more than one layer was multiplexed on limited resource nodes. The constellation length and size were the same in all the transmitted layers of SCMA.

In the SCMA scheme, the initial information of message passing algorithm (MPA) receiver was susceptible to noise

and multipath fading, and the criterion of permutation set failed to increase power differences between transmitted codewords [4, 7]. Driven by these problems, a sparse code multiple access advanced (SCMAA) scheme was proposed [7]. Under the same minimum Euclidean distance, SCMAA increased the sum of distances between interfering dimensions of transmitted codewords multiplexed on each resource node, which could improve the quality of initial information of MPA receiver on its corresponding resource node compared with SCMA [7–9]. However, in the SCMAA scheme, the increase of the sum of distances between interfering dimensions of transmitted codewords multiplexed on each resource node was limited by the suboptimal minimum intrapartition distance (the minimum intrapartition distance is the minimum Euclidean distance between basic complex multidimension constellation points in each partition). Moreover, the criterion of permutation set of SCMAA failed to maximize the sums of distances between interfering dimensions of transmitted codewords on some resource nodes (detailed explanation is offered in fifth line of Section 3.3.2). Hence the quality of initial information

of MPA receiver was unsatisfactory. On the other side, the increase of differences between the reliabilities of detections on all undetected transmitted layers in each decision process was limited by the uniform characteristic of SCMAA, and the criterion of permutation set of SCMAA did not increase the variance of the set of absolute differences between the sums of distances between interfering dimensions of transmitted codewords multiplexed on all resource nodes (detailed analysis is offered in Section 3.3.2 and the sixth paragraph of Section 4.2). Hence the convergence reliability of the detected layer in each decision process was unsatisfactory at the MPA receiver of SCMAA.

Driven by these problems, this paper proposes a nonuniform code multiple access (NCMA) scheme. Compared with SCMAA, some major improvements made in the proposed NCMA scheme are as follows. (i) Different transmitted layers of NCMA are generated from different complex multidimension constellations, respectively, while all the transmitted layers of SCMAA are generated from the same complex multidimension constellation. Therefore, in NCMA, the number of nonzero elements of transmitted codewords multiplexed on each resource node is totally different or not exactly the same (detailed explanation is offered in Section 3.2), and the number of nonzero elements occupied by each transmitted layer is totally different. However, in SCMAA, the number of nonzero elements of transmitted codewords multiplexed on each resource node is the same and so is the number of nonzero elements occupied by each transmitted layer. (ii) A novel basic complex multidimension constellation design is proposed. Compared with the basic complex multidimension constellation design of SCMAA, the proposed basic complex multidimension constellation design can further increase the minimum intrapartition distance. (iii) This paper proposes a novel criterion of permutation set, which can maximize the sum of distances (detailed definition is offered in the fourth paragraph of Section 3.3) between interfering dimensions of transmitted codewords multiplexed on any resource node, where the number of nonzero elements of transmitted codewords is more than 1. (iv) This paper proposes an advanced MPA receiver. At the proposed MPA receiver, the detection order of transmitted layers is fixed, and the function of initial information is equal to the function of initial information at traditional MPA receiver (traditional MPA receiver is short for the MPA receiver of SCMAA) multiplied by an amplification factor. On the other side, the complexity of the proposed MPA receiver is less than that of traditional MPA receiver (detailed explanation is offered in the fourth paragraph of Section 4.2).

Section 2 introduces the system model of NCMA. The codebook design of NCMA is presented in Section 3. The proposed MPA receiver and the performance analysis of NCMA scheme are offered in Section 4. Finally, in Section 5, the block error rate (BLER) performance of NCMA is compared with that of SCMAA and SCMA according to simulations.

2. System Model

In NCMA system, there are J transmitted layers multiplexed on K resource nodes. Each transmitted layer has its dedicated

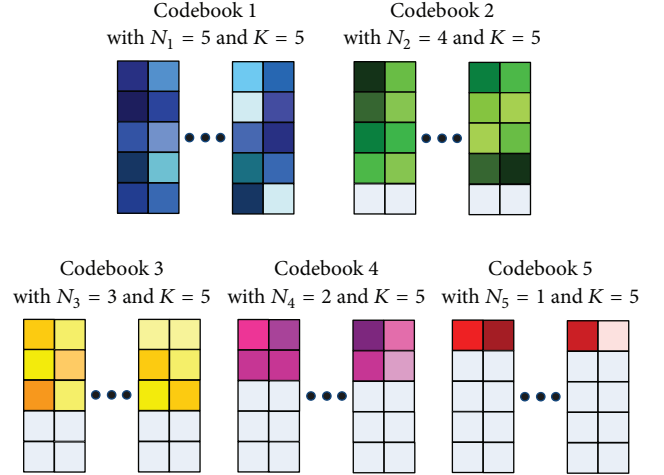


FIGURE 1: The codebooks of transmitted layers of NCMA with $N_1 = 5$, $N_2 = 4$, $N_3 = 3$, $N_4 = 2$, and $N_5 = 1$.

codebook. A codebook contains a plurality of K -dimension codewords. For layer j , a K -dimension codeword is generated by multiplying the binary mapping matrix V_j by a point from the complex N_j -dimension constellation C_j , and the size of C_j is M_j . V_j includes $K - N_j$ all-zero rows, and the rest can be expressed as identity matrix I_{N_j} after removing the all-zero rows from V_j . Hence each codeword of layer j includes N_j nonzero elements and $K - N_j$ zero elements. In NCMA system, different transmitted layers are generated from different complex multidimension constellations, respectively; that is, $C_i \neq C_j$, $N_i \neq N_j$, $i \neq j$, $\forall i, j = 1, \dots, J$. If $N_1 = 5$, $N_2 = 4$, $N_3 = 3$, $N_4 = 2$, and $N_5 = 1$, the codebooks of transmitted layers of NCMA are shown in Figure 1.

In order to improve spectral efficiency, more than one layer is multiplexed on limited resource nodes. In NCMA system, the received symbol after J layers multiplexing can be defined as

$$y = \sum_{j=1}^J \text{diag}(h_j) x_j + n_0, \quad (1)$$

where $h_j = (h_{1j}, h_{2j}, \dots, h_{Kj})^T$ is the channel vector of layer j , $x_j = (x_{1j}, x_{2j}, \dots, x_{Kj})^T$ is the codeword of layer j , $\text{diag}(h_j)$ is a diagonal matrix with elements from h_j , and n_0 is the white Gaussian noise vector.

In NCMA, the set of resource nodes occupied by layer j is determined by the indices of nonzero elements in f_j , $\forall j = 1, \dots, J$. f_j is a binary indicator vector, where the nonzero elements are determined by the indices of nonzero rows in V_j . As there are J transmitted layers in NCMA system, the structure of NCMA can be represented by a factor graph matrix $F = (f_1, \dots, f_J)$. In F , if $(F)_{kj} = 1$, layer node j and resource node k are connected. Figure 2 shows the factor graph representation of F with $N_1 = 5$, $N_2 = 4$, $N_3 = 3$, $N_4 = 2$, and $N_5 = 1$.

3. NCMA Codebook Design

Figure 3 shows the codebook design of NCMA with $N = 2$ and $K = 5$. According to Figure 3, we can conclude that the

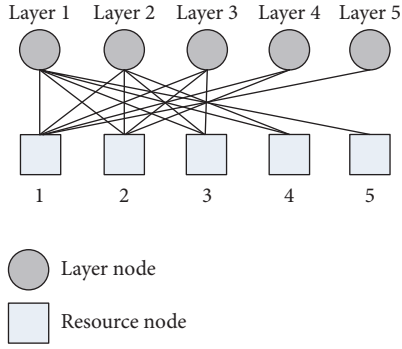


FIGURE 2: Factor graph of NCMA with $N_1 = 5$, $N_2 = 4$, $N_3 = 3$, $N_4 = 2$, and $N_5 = 1$.

codebook design of NCMA includes complex N -dimension constellation design (here N is short for N_j), permutation set, and mapping matrix. The complex N -dimension constellation design includes basic complex N -dimension constellation design, coordinate interleaving, and phase rotation. In the proposed codebook design of NCMA, coordinate interleaving and phase rotation are the same as the codebook design of SCMAA [7, 10, 11]. In the following, we will focus on the basic complex N -dimension constellation design, mapping matrix, and permutation set.

3.1. Basic Complex N -Dimension Constellation Design

3.1.1. The Basic Complex N -Dimension Constellation Design of SCMAA. The basic complex N -dimension constellation design of SCMAA was divided into two steps. First, the set of basic complex N -dimension signals was constructed by N -fold Cartesian product of a QAM signal set [12]. Then, in order to increase the minimum intrapartition distance, the set of basic complex N -dimension signals was divided into P partitions by Turbo Trellis Coded Modulation (Turbo TCM) technology [13, 14]. As Turbo TCM was applied in set partitioning, the minimum intrapartition distance was asymptotically suboptimal as the number of partitions increased.

3.1.2. The Basic Complex N -Dimension Constellation Design of NCMA. In order to further increase the minimum intrapartition distance, a novel basic complex N -dimension constellation design is proposed for NCMA. The proposed basic complex N -dimension constellation design is divided into three steps.

(i) We construct a real $2N$ -dimension constellation by sphere packing with the known densest lattice [15].

(ii) The real $2N$ -dimension constellation is divided into P partitions. The P partitions themselves will be translation-equivalent lattices; that is, each partition can be translated from any other partition. Hence they are all generated by the same set of basis vectors V_{per} , and the minimum intrapartition distance d_{min} is the same in each partition. If we draw spheres centered at points in each partition and the spheres just touch each other, we must choose the radius of the spheres to be $r = d_{\text{min}}/2$. Maximizing d_{min} for a given P

is equivalent to maximizing r for given $|\det V_{\text{per}}|$, where $P = |\det V_{\text{per}}|$, and $|\det V_{\text{per}}|$ is the absolute value of determinant of V_{per} . Hence the real $2N$ -dimension constellation partitioning is a sphere packing problem; that is, $V_{\text{per}} = a * V_{\text{gen}}^T$, where V_{gen}^T is the transpose of the generator matrix V_{gen} of the densest $2N$ -dimension lattice and a is a constant that is determined by P . For example, for a real 2-dimension constellation, the hexagonal lattice is the densest sphere packing in two dimensions, and therefore each partition is also hexagonal.

Hence $V_{\text{per}} = [v_1 \ v_2] = \begin{bmatrix} 2a & a \\ 0 & \sqrt{3}a \end{bmatrix}$, where $a = \sqrt{P/(2\sqrt{3})}$. The minimum intrapartition distance can be expressed as $d_{\text{min}} = \min(\|v_1\|, \|v_2\|, \|v_1 - v_2\|, \|v_1 + v_2\|) = \sqrt{2P/\sqrt{3}}$, and $d_{\text{min}} > d_{\text{min}}^T = \sqrt{P}$, where d_{min}^T is the maximum d_{min} of the basic complex 1-dimension constellation of SCMAA. It will do the same for other real multidimension constellations.

(iii) As a real $2N$ -dimension constellation point $s = [s_1, s_2, \dots, s_{2N}]$ is given, we can obtain a basic complex N -dimension constellation point $s_c = [s_1 + js_2, s_3 + js_4, \dots, s_{2N-1} + js_{2N}]$.

According to (i), (ii), and (iii), we can conclude that the proposed basic complex N -dimension constellation design can increase the minimum intrapartition distance compared with the basic complex N -dimension constellation design of SCMAA.

3.2. Mapping Matrix of NCMA. The nonuniform characteristic of NCMA is determined by a mapping matrix set $V = \{V_1, V_2, \dots, V_J\}$. The mapping matrix design rules of NCMA are as follows. (i) $V_j \in B^{K \times N_j}$, where B represents a binary matrix. (ii) $V_i \neq V_j, \forall i \neq j, i, j = 1, \dots, J$. (iii) $V_j^{[\Theta]} = I_{N_j}$, where $V_j^{[\Theta]}$ is V_j after removing its all-zero rows. The mapping properties of V are as follows.

(i) The number of nonzero elements of transmitted codewords multiplexed on each resource node is totally different or not exactly the same. Moreover, d_{1f} is the maximum in $\{d_{1f}, \dots, d_{kf}, \dots, d_{Kf}\}$, and $d_{Kf} = 1$. In other words, $1 \leq d_{kf} \leq d_{1f}$, where d_{kf} is the number of nonzero elements of transmitted codewords multiplexed on resource node k .

(ii) The number of nonzero elements occupied by each transmitted layer is totally different, and $n_{1f} > \dots > n_{jf} > \dots > n_{Jf}$, where n_{jf} is the number of nonzero elements occupied by layer $j, \forall j = 2, \dots, J - 1$.

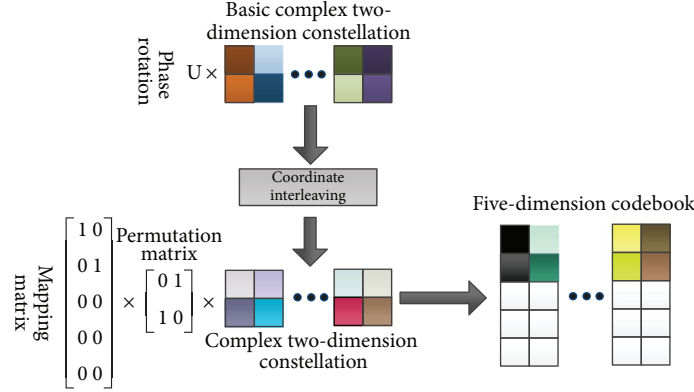
(iii) $K = N_1$, and $N_1 > \dots > N_j > \dots > N_j, \forall j = 2, \dots, J - 1$.

(iv) $J = d_{1f}$.

For example, if $N_1 = 5, N_2 = 4, N_3 = 3, N_4 = 2$, and $N_5 = 1$, there are five transmitted layers multiplexed on $K = N_1 = 5$ resource nodes, and therefore the factor graph matrix can be

expressed as $F_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$. In $F_1, d_{1f} > d_{2f} > d_{3f} > d_{4f} >$

d_{5f} . Hence the number of nonzero elements of transmitted codewords multiplexed on each resource node is totally different. For another example, if $N_1 = 4, N_2 = 3$, and $N_3 = 1$, there are three transmitted layers multiplexed on $K = N_1 = 4$ resource nodes, and therefore the factor graph matrix can

FIGURE 3: NCMA codebook design with $N = 2$ and $K = 5$.

be expressed as $F_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. In F_2 , $d_{1f} > d_{2f} = d_{3f} > d_{4f}$. Hence the number of nonzero elements of transmitted codewords multiplexed on each resource node is not exactly the same.

3.3. Permutation Set. For layer j , if the operator on constellation C_j is limited to permutation matrix π_j , the codeword can be defined as

$$x_j = q_j = V_j \pi_j z_j, \quad \forall j = 1, \dots, J, \quad (2)$$

where $z_j = (z_j^1, z_j^2, \dots, z_j^{N_j})^T$ represents an arbitrary alphabet of constellation C_j , $z_j^n \in {}^n C_j = \{c_{nm_j} = (c_{m_j})_n \mid \forall c_{m_j} \in C_j, m_j = 1, \dots, M_j\}$, and ${}^n C_j$ represents the n th dimension of constellation C_j . Under these conditions, the aggregate received symbol can be expressed as

$$p(z) = \sum_{j=1}^J q_j(z_j) = \sum_{j=1}^J V_j \pi_j z_j, \quad (3)$$

where $p(z) = (p_1(z), \dots, p_k(z), \dots, p_K(z))^T$ is a $K \times 1$ vector, $p_k(z) = d_{k1} z^{1,k} + d_{k2} z^{2,k} + \dots + d_{kN_s} z^{N_s,k}$ represents the interfering polynomial on resource node k , $z^{n,k}$ represents the n th dimension of any constellation on resource node k , $1 \leq N_s \leq N_{\max}$, N_{\max} is the maximum in $\{N_1, N_2, \dots, N_J\}$, and $\forall k = 1, \dots, K$. As the number of nonzero elements of transmitted codewords multiplexed on resource node k is d_{kf} , we can conclude that $\sum_{n=1}^{N_s} d_{kn} = d_{kf}, \forall k = 1, \dots, K$. For example, according to F_1 in Section 3.2, the interfering polynomial on resource node 2 can be expressed as $p_2(z) = 2z^{1,2} + 2z^{2,2}$. According to $p_2(z)$, we can conclude that there are four nonzero elements of transmitted codewords multiplexed on resource node 2. In the four nonzero elements, two of them come from ${}^1 C$, and the others come from ${}^2 C$, where ${}^1 C = \{{}^1 C_1, {}^1 C_2, {}^1 C_3, {}^1 C_4\}$ and ${}^2 C = \{{}^2 C_1, {}^2 C_2, {}^2 C_3, {}^2 C_4\}$. In summary, for a given mapping matrix set V , the set $d_{\text{set}}^k = \{d_{k1}, \dots, d_{kn}, \dots, d_{kN_s}\}$ depends on permutation set $\Pi = [\pi_j]_{j=1}^J, \forall k = 1, \dots, K$. Hence there is a one-to-one mapping between permutation set Π and $p(z)$. Permutation

set Π determines the sum of distances between interfering dimensions of transmitted codewords multiplexed on any resource node, where the number of nonzero elements of transmitted codewords is more than 1. If $d_{kf} > 1$, the sum of distances between interfering dimensions of transmitted codewords multiplexed on resource node k can be expressed as

$$\begin{aligned} E_r^k &= |x_{j1,n1}^{k,r} - x_{j2,n2}^{k,r}|^2 + \dots \\ &\quad + |x_{j1,n1}^{k,r} - x_{jd_{kf},nd_{kf}}^{k,r}|^2 + \dots \\ &\quad + |x_{j(d_{kf}-2),n(d_{kf}-2)}^{k,r} - x_{jd_{kf},nd_{kf}}^{k,r}|^2 \\ &\quad + |x_{j(d_{kf}-1),n(d_{kf}-1)}^{k,r} - x_{jd_{kf},nd_{kf}}^{k,r}|^2, \\ E_{\text{im}}^k &= |x_{j1,n1}^{k,\text{im}} - x_{j2,n2}^{k,\text{im}}|^2 + \dots \\ &\quad + |x_{j1,n1}^{k,\text{im}} - x_{jd_{kf},nd_{kf}}^{k,\text{im}}|^2 + \dots \\ &\quad + |x_{j(d_{kf}-2),n(d_{kf}-2)}^{k,\text{im}} - x_{jd_{kf},nd_{kf}}^{k,\text{im}}|^2 \\ &\quad + |x_{j(d_{kf}-1),n(d_{kf}-1)}^{k,\text{im}} - x_{jd_{kf},nd_{kf}}^{k,\text{im}}|^2, \\ n(p_k(z)) &= \sqrt{E_r^k + E_{\text{im}}^k}, \end{aligned} \quad (4)$$

where $x_{j,n}^{k,r}$ is the real part of the signal on the n th dimension of the codeword of layer j on resource node k , $x_{j,n}^{k,\text{im}}$ is the imaginary part of the signal on the n th dimension of the codeword of layer j on resource node k , and $n(p_k(z))$ is the sum of distances between interfering dimensions of transmitted codewords multiplexed on resource node k . As illustrated in the third paragraph of Section 3.3, there is a one-to-one mapping between permutation set Π and $p(z)$. Hence there is a one-to-one mapping between permutation set Π and $n(p(z))$, where $n(p(z)) = \{n(p_1(z)), n(p_2(z)), \dots, n(p_{K_s}(z))\}$, and K_s is the number of resource nodes where the number of nonzero elements of transmitted codewords is more than 1.

3.3.1. The Novel Criterion of Permutation Set of NCMA. In the NCMA scheme, a novel criterion of permutation set is proposed to maximize $n(p_k(z))$, and the proposed criterion is divided into two steps (the first step corresponds to formula (5), and the second step corresponds to formula (6)). First, formula (5) selects the permutation sets where $n(p_1(z)) + \dots + n(p_{K_s}(z))$ is maximum.

$$\begin{aligned} & \{\Pi^{1*}, \Pi^{2*}, \dots\} \\ & = \arg \max_{\Pi} (n(p_1(z)) + \dots + n(p_{K_s}(z))). \end{aligned} \quad (5)$$

There is more than one permutation set selected by formula (5); that is, $\Pi^* = \{\Pi^{1*}, \Pi^{2*}, \dots\}$. Then, among Π^* , formula (6) selects the most appropriate permutation set Π^{l**} , which can minimize the variance of all the elements in $n(p(z)) = \{n(p_1(z)), n(p_2(z)), \dots, n(p_{K_s}(z))\}$.

$$\Pi^{l**} = \arg \min_{\Pi^*} \text{var} (n(p(z))), \quad \Pi^{l*} \in \Pi^*, \quad (6)$$

where var is the variance function.

3.3.2. The Criterion of Permutation Set of SCMAA. The criterion of permutation set of SCMAA was divided into two steps [7]. First, the criterion of SCMAA selected the permutation sets where the minimum in corresponding $n(p(z))$ was maximum. Secondly, among the selected permutation sets, the criterion of SCMAA selected the most appropriate permutation set, which could maximize the variance of all the elements in $n(p(z))$. But the criterion of SCMAA did not maximize some elements in the set (the set $n^*(p(z))$ is the set $n(p(z))$ selected in the second step) $n^*(p(z))$. On the other side, the criterion of SCMAA did not increase the variance of all the elements in n_{set} , where $n_{\text{set}} = \{n_{1,2}, n_{1,3}, \dots, n_{K_s-1, K_s}\}$, $n_{k_1, k_2} = |n(p_{k_1}(z)) - n(p_{k_2}(z))|$, and $k_1 < k_2, \forall k_1 = 1, \dots, K_s - 1, \forall k_2 = 2, \dots, K_s$.

4. The Proposed MPA Receiver and the Performance Analysis of NCMA Scheme

4.1. The Proposed MPA Receiver of NCMA. In this paper, the proposed MPA receiver of NCMA uses an advanced min-sum algorithm. The structure of NCMA can be represented by a factor graph F with J layer nodes and K resource nodes. At the proposed MPA receiver, layer nodes can be seen as check nodes, resource nodes can be seen as variable nodes, and the process where messages are exchanged between variable nodes and check nodes is as follows.

The message exchanged from variable node k to check node j is given by

$$v_{k \rightarrow j}(x_j) = \gamma_k(x_j) + \sum_{i \in \Psi(k) \setminus j} \mu_{i \rightarrow k}(x_i), \quad (7)$$

$$\begin{aligned} & \gamma_k(x_j) \\ & = -\varepsilon_k \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{\|y_k - \sum_{i \in \Psi(k)} x_{i,k} h_k\|^2}{2\sigma^2} \right) \right), \end{aligned} \quad (8)$$

where y_k is the received symbol on resource node k , $v_{k \rightarrow j}(x_j)$ is the cost function where message is exchanged from variable node k to check node j when the value of check node j is x_j , $\gamma_k(x_j)$ is the function of initial information on variable node k when the value of check node j is x_j , ε_k is the amplification factor in $\gamma_k(x_j)$, $\varepsilon_k > 0$, σ^2 is noise power, $\mu_{i \rightarrow k}(x_i)$ is the cost function where message is exchanged from check node i to variable node k when the value of check node i is x_i , $\Psi(k) \setminus j$ represents the set of all check nodes connecting to variable node k except check node j , and $\exp(\cdot)$ is the exponential function.

The message exchanged from check node j to variable node k is given by

$$\mu_{j \rightarrow k}(x_j) = \min \left(\sum_{l \in \Phi(j) \setminus k} v_{l \rightarrow j}(x_j) \right), \quad (9)$$

where $\Phi(j) \setminus k$ represents the set of all variable nodes connecting to check node j except variable node k . After several iterations, the final cost function of check node j , when the value of check node j is x_j , is

$$\mu(x_j) = \sum_{l \in \Phi(j)} v_{l \rightarrow j}(x_j). \quad (10)$$

At the proposed MPA receiver, the process where messages are exchanged between variable nodes and check nodes is similar to that at traditional MPA receiver [16, 17]. However, on each resource node, the function of initial information at the proposed MPA receiver is equal to the function of initial information at traditional MPA receiver multiplied by the corresponding amplification factor. As the number of nonzero elements of transmitted codewords of NCMA multiplexed on each resource node is totally different or not exactly the same, the amplification factor on each resource node is totally different or not exactly the same. Moreover, if $d_{k_1, f}$ is less than $d_{k_2, f}$, ε_{k_1} is more than ε_{k_2} , where $k_1 \neq k_2, \forall k_1, k_2 = 1, 2, \dots, K$.

4.2. Performance Analysis of NCMA Scheme. In this paper, the NCMA scheme is proposed to improve the quality of initial information on each resource node and the convergence reliability of the detected layer in each decision process at the proposed MPA receiver. The performance analysis of the proposed NCMA scheme is presented in two aspects as follows.

(i) The quality of initial information on resource node k can be improved by enlarging the decision region of \widehat{y}_k [7], where \widehat{y}_k is the expected symbol on resource node $k, \forall k = 1, \dots, K$. On resource node k , if there are interfering nonzero elements of transmitted codewords, increasing $n(p_k(z))$ will enlarge the decision region of \widehat{y}_k . In the codebook design of NCMA, a novel criterion of permutation set is proposed. The proposed criterion of permutation set maximizes $n(p_1(z)) + \dots + n(p_{K_s}(z))$ and minimizes the variance of all the elements in $n(p(z)) = \{n(p_1(z)), n(p_2(z)), \dots, n(p_{K_s}(z))\}$ and therefore can maximize $n(p_k(z)), \forall k = 1, \dots, K_s$. On the other side, on resource node k , if there are no interfering nonzero elements

of transmitted codewords, the decision region of \widehat{y}_k can be enlarged by increasing the minimum intrapartition distance. If $N_1 = 5$, $N_2 = 4$, $N_3 = 3$, $N_4 = 2$, and $N_5 = 1$, the factor graph matrix of NCMA can be expressed as $F_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$. According to F_1 , we can conclude that there are no interfering nonzero elements of transmitted codewords multiplexed on resource node 5 in the first decision process. If the transmitted codeword of layer 1 has been detected in the first decision process, there will be no interfering nonzero elements of transmitted codewords multiplexed on resource node 4 in the second decision process. It will do the same for resource node 1, resource node 2, and resource node 3 in the other decision processes. In the codebook design of NCMA, a novel basic complex multidimension constellation design is proposed. As illustrated in Section 3.1.2, the proposed basic complex multidimension constellation design increases the minimum intrapartition distance compared with the basic complex multidimension constellation design of SCMAA.

(ii) In each decision process, the convergence reliability of the detected layer is related to the differences between the reliabilities of detections on all undetected layers and the differences between the reliabilities of detections on the codewords of each undetected layer [7]. Therefore, a novel mapping matrix and an advanced MPA receiver are proposed in the NCMA scheme.

According to the proposed mapping matrix of NCMA, we can conclude that the number of nonzero elements occupied by each transmitted layer is totally different, and the number of nonzero elements of transmitted codewords multiplexed on each resource node is totally different or not exactly the same. Benefiting from the nonuniform characteristic of NCMA, the differences between the reliabilities of detections on all undetected layers will be increased in each decision process, and the detection order of transmitted layers is fixed at the proposed MPA receiver; that is, layer 1 is detected in the first decision process, layer 2 is detected in the second decision process, ..., and layer J is detected in the J th decision process. Detailed analysis is shown as follows. According to F_1 in second paragraph of Section 4.2, we can conclude that $n_{1f} > n_{2f} > n_{3f} > n_{4f} > n_{5f}$ and $d_{1f} > d_{2f} > d_{3f} > d_{4f} > d_{5f}$. In formula (10), $\mu(x_j)$ is equal to $\sum_{l \in \Phi(j)} \nu_{l \rightarrow j}(x_j)$, and $\Phi(j)$ is determined by n_{jf} . The more n_{jf} is, the more detection information layer j obtains, $\forall j = 1, \dots, J$. On the other side, in formula (8), the value of $\|y_k - \sum_{i \in \psi(k)} x_{i,k} h_k\|^2$ is determined by $\psi(k)$, and $\psi(k)$ is determined by d_{kf} . The less d_{kf} is, the less the value of $\|y_k - \sum_{i \in \psi(k)} x_{i,k} h_k\|^2$ is, $\forall k = 1, \dots, K$. Hence the quality of initial information on resource node k can be improved by decreasing d_{kf} , $\forall k = 1, \dots, K$. As $d_{5f} < d_{kf}$ and there are no interfering nonzero elements of transmitted codewords multiplexed on resource node 5, the quality of initial information on resource node 5 obviously outperforms that on resource node k , $\forall k = 1, 2, 3, 4$. In the first decision process at the proposed MPA receiver, as $n_{1f} > n_{jf}$ and resource node 5 is only occupied by layer 1, layer 1 can obtain more reliable detection information than layer j ($\forall j = 2, 3, 4, 5$), and therefore layer 1 is detected. After

layer 1 has been detected, there will be no interfering nonzero elements of transmitted codewords multiplexed on resource node 4 and $d_{4f} < d_{kf}$, and therefore the quality of initial information on resource node 4 will obviously outperform that on resource node k , $\forall k = 1, 2, 3$. In the second decision process, as $n_{2f} > n_{jf}$ and resource node 4 is occupied by layer 2, layer 2 can obtain more reliable detection information than layer j ($\forall j = 3, 4, 5$), and therefore layer 2 is detected. It will do the same for layer 3, layer 4, and layer 5 in their corresponding decision processes. Therefore, for F_1 , layer 1 is detected in the first decision process, layer 2 is detected in the second decision process, layer 3 is detected in the third decision process, layer 4 is detected in the fourth decision process, and layer 5 is detected in the fifth decision process. In any other NCMA scheme with different parameters, the detection order of transmitted layers is similar to that of transmitted layers in the NCMA scheme with $N_1 = 5$, $N_2 = 4$, $N_3 = 3$, $N_4 = 2$, and $N_5 = 1$. In addition, in each decision process, the proposed MPA receiver of NCMA selects the codeword of which the value of final cost function is the least, after detecting the codewords of a given transmitted layer. However, in each decision process, traditional MPA receiver selects the codeword of which the value of final cost function is the least, after detecting the codewords of all the undetected transmitted layers. Therefore, the complexity of the proposed MPA receiver is less than that of traditional MPA receiver.

In each decision process at the proposed MPA receiver, the amplification factor in the function of initial information can increase the differences between the reliabilities of detections on the codewords of each undetected layer and therefore can improve the reliability of detection on each transmitted layer. Detailed analysis is shown as follows. As illustrated in the fourth paragraph of Section 4.2, the less d_{kf} is, the higher the quality of initial information on resource node k is, $\forall k = 1, \dots, K$. Therefore, in the process of detection on a transmitted codeword, we can prefer the information of such resource node occupied by the codeword, the interferences on which are less than those on another resource node. According to F_1 in second paragraph of Section 4.2, we can conclude that $d_{1f} > d_{2f} > d_{3f} > d_{4f} > d_{5f}$. As illustrated in Section 4.1, if d_{k_1f} is less than d_{k_2f} , ε_{k_1} is more than ε_{k_2} , where $k_1 \neq k_2$, $\forall k_1, k_2 = 1, 2, \dots, K$. Hence $\varepsilon_5 > \varepsilon_4 > \varepsilon_3 > \varepsilon_2 > \varepsilon_1$, and therefore the ratio of detection information on the resource nodes with less interferences to that on all the resource nodes in $\mu(x_1)$ will be increased. On the other side, ε_k can increase the difference between $\gamma_k(x_1^i)$ and $\gamma_k(x_1^j)$, and therefore the difference between $\mu(x_1^i)$ and $\mu(x_1^j)$ will be increased, where x_1^i and x_1^j are the codewords of layer 1, $\forall i \neq j, i, j = 1, \dots, M_1$, $\forall k = 1, \dots, 5$. All in all, the amplification factor can further increase the differences between the reliabilities of detections on the codewords of layer 1 and therefore improve the reliability of detection on layer 1. It will do the same for the other layers. For the factor graph of NCMA with other parameters, the amplification factor can also improve the reliability of detection on each transmitted layer.

For SCMAA, the convergence reliability of the detected layer in each decision process is unsatisfactory at traditional

MPA receiver. Detailed analysis is shown as follows. If $J = 6$ and $K = 4$, the factor graph matrix of SCMAA can be expressed as $F_S = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$. According to F_S , we can conclude that $n_{1f} = n_{2f} = n_{3f} = n_{4f} = n_{5f} = n_{6f}$ and $d_{1f} = d_{2f} = d_{3f} = d_{4f}$. Limited by the uniform characteristic of SCMAA, the differences between the reliabilities of detections on all undetected transmitted layers in each decision process cannot be obtained by increasing the differences between any two elements in $n_f = \{n_{1f}, n_{2f}, n_{3f}, n_{4f}, n_{5f}, n_{6f}\}$ and the differences between any two elements in $d_f = \{d_{1f}, d_{2f}, d_{3f}, d_{4f}\}$. On the other side, under some initial conditions (these initial conditions are as follows. (i) The value of x_j^S is expectation, where x_j^S is the value of layer node j of SCMAA, $\forall j = 1, 2, \dots, 6$. (ii) The initial values of $v_{k \rightarrow j}(x_j^S)$ and $\mu_{j \rightarrow k}(x_j^S)$ are 0, $\forall k = 1, 2, 3, 4, \forall j = 1, 2, \dots, 6$), the difference between $\mu(x_1^S)$ and $\mu(x_6^S)$ can be expressed as $\mu(x_1^S) - \mu(x_6^S) = |\gamma_2^S - \gamma_4^S| - |\gamma_1^S - \gamma_3^S|$, and the difference between $\mu(x_3^S)$ and $\mu(x_4^S)$ can be expressed as $\mu(x_3^S) - \mu(x_4^S) = |\gamma_3^S - \gamma_4^S| - |\gamma_1^S - \gamma_2^S|$, where $\mu(x_j^S)$ is the final cost function of layer node j when the value of layer node j of SCMAA is x_j^S and γ_k^S is the function of initial information on resource node k at traditional MPA receiver. Detailed derivation process of the difference between $\mu(x_1^S)$ and $\mu(x_6^S)$ refers to [7] and so is the difference between $\mu(x_3^S)$ and $\mu(x_4^S)$. At traditional MPA receiver, the larger the difference between any two elements in $\mu^S = \{\mu(x_1^S), \mu(x_2^S), \dots, \mu(x_6^S)\}$ is, the larger the differences between the reliabilities of detections on all undetected layers in each decision process. Moreover, in each decision process, the larger the differences between the reliabilities of detections on all undetected layers are, the higher the convergence reliability of the detected layer is [7]. As illustrated in Section 3.3.2, the criterion of permutation set of SCMAA increases neither the difference between $|n(p_2(z)) - n(p_4(z))|$ and $|n(p_1(z)) - n(p_3(z))|$ nor the difference between $|n(p_3(z)) - n(p_4(z))|$ and $|n(p_1(z)) - n(p_2(z))|$. That is, the criterion of permutation set of SCMAA increases neither the difference between $|\gamma_2^S - \gamma_4^S|$ and $|\gamma_1^S - \gamma_3^S|$ nor the difference between $|\gamma_3^S - \gamma_4^S|$ and $|\gamma_1^S - \gamma_2^S|$ (γ_k^S is determined by $n(p_k(z))$ [7], $\forall k = 1, \dots, 4$). Therefore, the criterion of permutation set of SCMAA will attenuate the convergence reliability of layer 1, layer 3, layer 4, and layer 6 in their corresponding decision processes at traditional MPA receiver. It will do the same in the process of detections on the transmitted layers of SCMAA scheme with other parameters. In summary, in each decision process, the increase of the differences between the reliabilities of detections on all undetected transmitted layers is limited by the uniform characteristic of SCMAA, and the criterion of SCMAA fails to increase the differences between the reliabilities of detections on some undetected layers. Hence the convergence reliability of the detected layer of SCMAA in each decision process is unsatisfactory.

According to (ii), we can conclude that, benefiting from the proposed mapping matrix and the proposed MPA receiver, NCMA can further improve the convergence

reliability of the detected layer in each decision process compared with SCMAA.

5. Simulation Results

In this section, simulations are based on long-term evolution (LTE) system [18], and the channel code uses Turbo code with the rate 1/2. In NCMA, SCMAA, and SCMA, the number of iterations is 4 at the proposed MPA receiver and traditional MPA receiver (traditional MPA receiver is applied in SCMAA and SCMA). For NCMA, the real 2-dimension constellation is constructed by sphere packing with A_2 , the real 4-dimension constellation is constructed by sphere packing with D_4 , the real 6-dimension constellation is constructed by sphere packing with E_6 , the real 8-dimension constellation is constructed by sphere packing with E_8 , and the real 10-dimension constellation is constructed by sphere packing with Λ_{10} [15]. For SCMAA and SCMA, the set of basic complex two-dimension signals is constructed by 2-fold Cartesian product of a QPSK set. As the spectral efficiency is 2 bits/tonne, NCMA uses the factor graph with $N_1 = 4, N_2 = 3$, and $N_3 = 1$, while SCMAA and SCMA use the factor graph (for SCMAA and SCMA, the factor graph is shown in [7]) with $J = 4$ and $K = 4$. As the spectral efficiency is 3 bits/tonne, NCMA uses the factor graph with $N_1 = 5, N_2 = 4, N_3 = 3, N_4 = 2$, and $N_5 = 1$, while SCMAA and SCMA use the factor graph with $J = 6$ and $K = 4$. In the following, NCMA with traditional MPA receiver is short for the NCMA scheme, where the proposed codebook design and traditional MPA receiver are applied, and NCMA with proposed MPA receiver is short for the NCMA scheme, where the proposed codebook design and the proposed MPA receiver are applied.

Figure 4 is the BLER performance of NCMA with traditional MPA receiver, SCMAA with traditional MPA receiver, and SCMA with traditional MPA receiver over AWGN channel with spectral efficiency 2 bits/tonne. As can be observed in Figure 4, the BLER performance of NCMA with traditional MPA receiver outperforms that of SCMAA with traditional MPA receiver, while the BLER performance of SCMAA with traditional MPA receiver outperforms that of SCMA with traditional MPA receiver. NCMA with traditional MPA receiver has 1.1 dB gain over SCMAA with traditional MPA receiver. Figure 5 is the BLER performance of NCMA with traditional MPA receiver, SCMAA with traditional MPA receiver, and SCMA with traditional MPA receiver over AWGN channel with spectral efficiency of 3 bits/tonne. As can be observed in Figure 5, the BLER performance of NCMA with traditional MPA receiver outperforms that of SCMAA with traditional MPA receiver, while the BLER performance of SCMAA with traditional MPA receiver outperforms that of SCMA with traditional MPA receiver. NCMA with traditional MPA receiver has 1.4 dB gain over SCMAA with traditional MPA receiver. Simulation results show that the proposed codebook design of NCMA can improve the performance of traditional MPA receiver compared with the codebook design of SCMAA over AWGN channel.

Figure 6 is the BLER performance of NCMA with proposed MPA receiver, SCMAA with traditional MPA receiver,

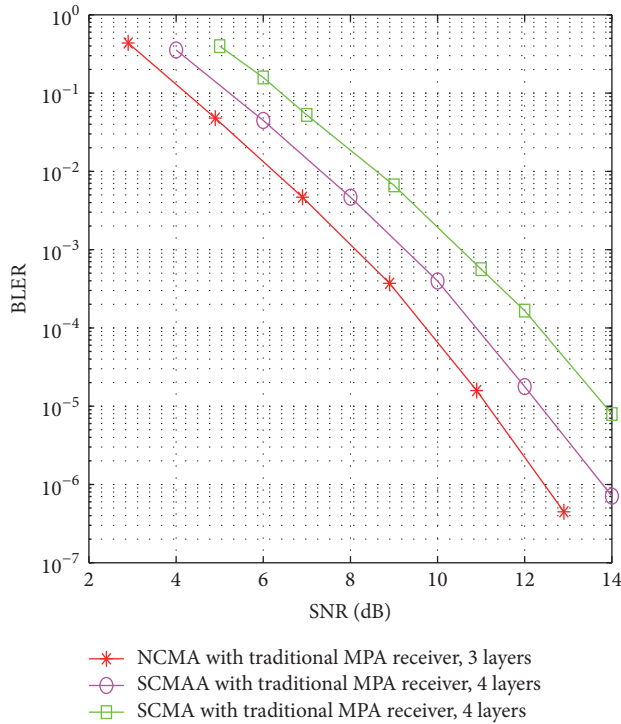


FIGURE 4: NCMA with traditional MPA receiver versus SCMAA with traditional MPA receiver and SCMA with traditional MPA receiver over AWGN channel with 2 bits/tonne.

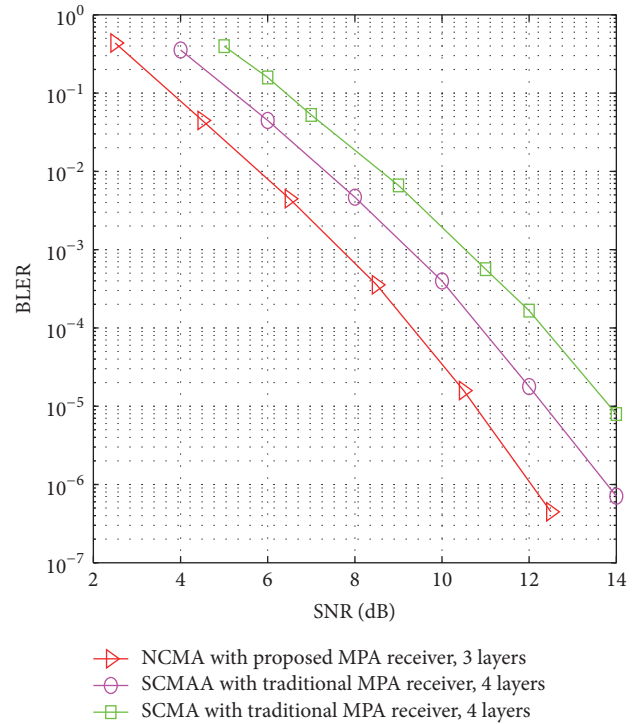


FIGURE 6: NCMA with proposed MPA receiver versus SCMAA with traditional MPA receiver and SCMA with traditional MPA receiver over AWGN channel with 2 bits/tonne.

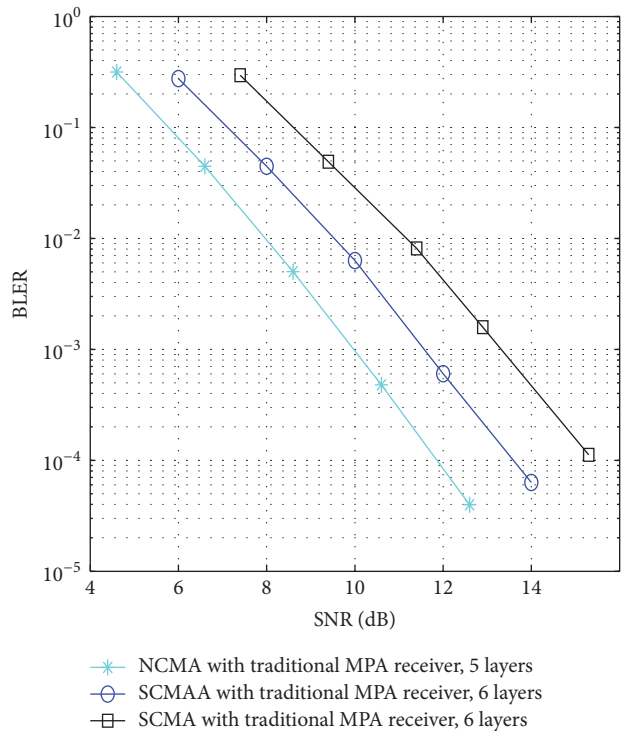


FIGURE 5: NCMA with traditional MPA receiver versus SCMAA with traditional MPA receiver and SCMA with traditional MPA receiver over AWGN channel with 3 bits/tonne.

and SCMA with traditional MPA receiver over AWGN channel with spectral efficiency of 2 bits/tonne. As can be observed in Figure 6, the BLER performance of NCMA with proposed MPA receiver outperforms that of SCMAA with traditional MPA receiver, while the BLER performance of SCMAA with traditional MPA receiver outperforms that of SCMA with traditional MPA receiver. NCMA with proposed MPA receiver has 1.5 dB gain over SCMAA with traditional MPA receiver. As can be observed in Figures 4 and 6, the BLER performance of NCMA with proposed MPA receiver outperforms that of NCMA with traditional MPA receiver. Figure 7 is the BLER performance of NCMA with proposed MPA receiver, SCMAA with traditional MPA receiver, and SCMA with traditional MPA receiver over AWGN channel with spectral efficiency of 3 bits/tonne. As can be observed in Figure 7, the BLER performance of NCMA with proposed MPA receiver outperforms that of SCMAA with traditional MPA receiver, while the BLER performance of SCMAA with traditional MPA receiver outperforms that of SCMA with traditional MPA receiver. NCMA with proposed MPA receiver has 1.9 dB gain over SCMAA with traditional MPA receiver. As can be observed in Figures 5 and 7, the BLER performance of NCMA with proposed MPA receiver outperforms that of NCMA with traditional MPA receiver. Simulation results show that the proposed MPA receiver can further improve the convergence reliability of the detected layer in each decision process compared with traditional MPA receiver over AWGN channel.

Figure 8 is the capacity of NCMA with proposed MPA receiver, SCMAA with traditional MPA receiver, and SCMA

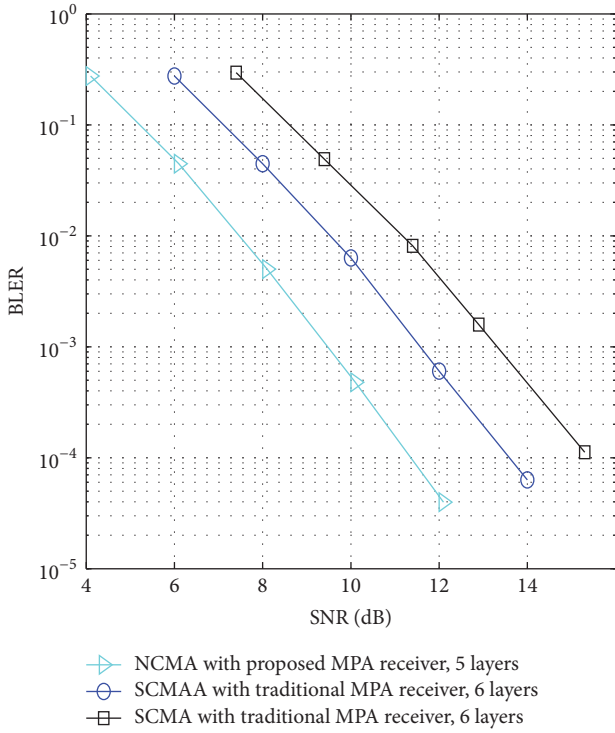


FIGURE 7: NCMA with proposed MPA receiver versus SCMAA with traditional MPA receiver and SCMA with traditional MPA receiver over AWGN channel with 3 bits/tone.

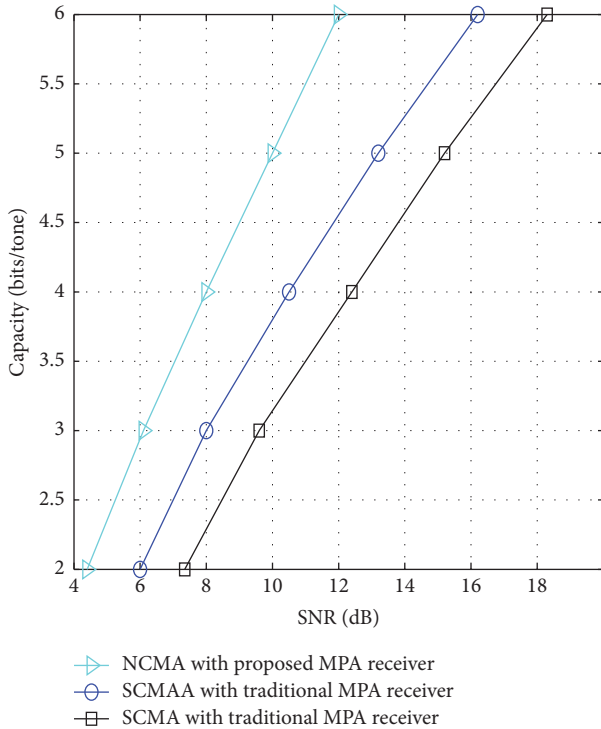


FIGURE 8: Capacity of NCMA with proposed MPA receiver, SCMAA with traditional MPA receiver, and SCMA with traditional MPA receiver over AWGN channel.

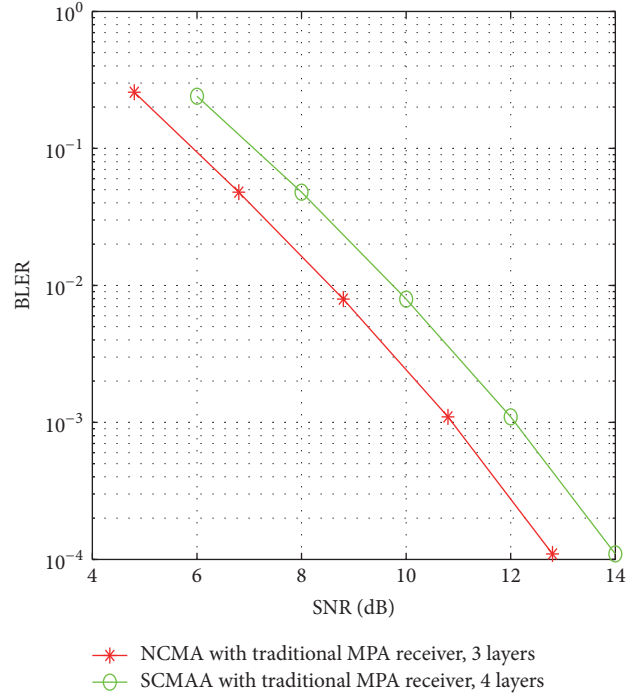


FIGURE 9: NCMA with traditional MPA receiver versus SCMAA with traditional MPA receiver over fading channel with 2 bits/tonne.

with traditional MPA receiver over AWGN channel. For each target spectral efficiency, the minimum SNR is selected to guarantee the appropriate performance for each waveform. As can be observed in Figure 8, we can conclude that, compared with SCMAA with traditional MPA receiver and SCMA with traditional MPA receiver, the gain of NCMA with proposed MPA receiver is obvious, and it grows as the SNR increases.

In Figures 9, 10, 11, and 12, the simulations are based on downlink LTE system, and all transmitted layers are multiplexed on orthogonal frequency division multiple access (OFDMA) tones in a pedestrian B (PB) fading channel with speed of 3 km/h [18]. The carrier frequency is 2 GHz and the frequency spacing is 15 KHz. A data payload occupies 6 LTE resource blocks (RBs). Figure 9 is the BLER performance of NCMA with traditional MPA receiver and SCMAA with traditional MPA receiver over fading channel with spectral efficiency of 2 bits/tonne. As can be observed in Figure 9, the BLER performance of NCMA with traditional MPA receiver outperforms that of SCMAA with traditional MPA receiver, and NCMA with traditional MPA receiver has 1.2 dB gain over SCMAA with traditional MPA receiver. Figure 10 is the BLER performance of NCMA with traditional MPA receiver and SCMAA with traditional MPA receiver over fading channel with spectral efficiency of 3 bits/tonne. As can be observed in Figure 10, the BLER performance of NCMA with traditional MPA receiver outperforms that of SCMAA with traditional MPA receiver, and NCMA with traditional MPA receiver has 1.8 dB gain over SCMAA with traditional MPA receiver. Simulation results show that the proposed

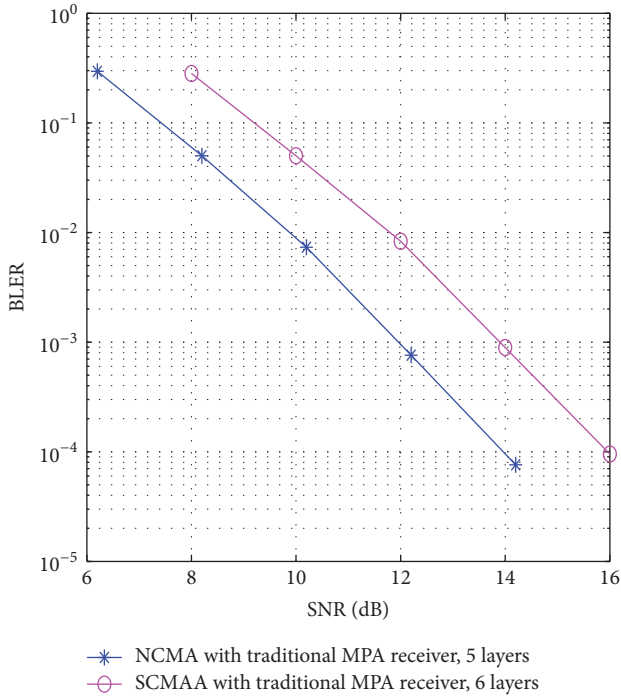


FIGURE 10: NCMA with traditional MPA receiver versus SCMAA with traditional MPA receiver over fading channel with 3 bits/ton.

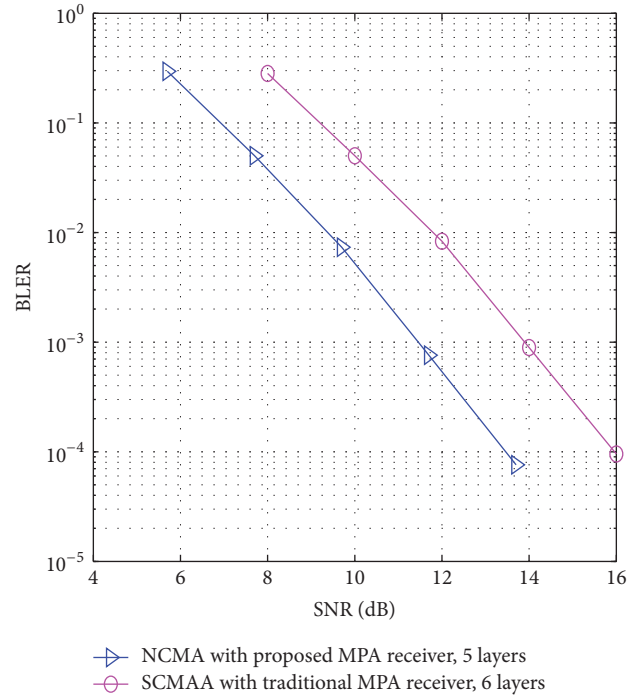


FIGURE 12: NCMA with proposed MPA receiver versus SCMAA with traditional MPA receiver over fading channel with 3 bits/ton.

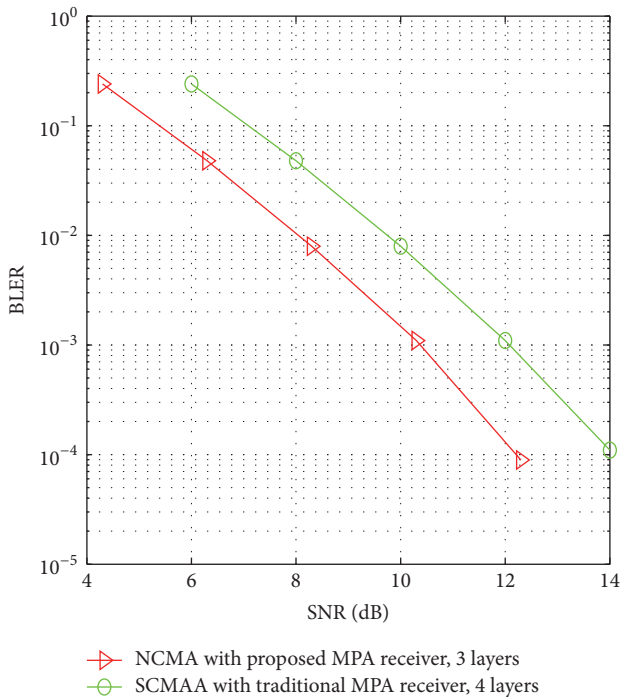


FIGURE 11: NCMA with proposed MPA receiver versus SCMAA with traditional MPA receiver over fading channel with 2 bits/ton.

codebook design of NCMA can improve the performance of traditional MPA receiver compared with the codebook design of SCMAA over fading channel.

Figure 11 is the BLER performance of NCMA with proposed MPA receiver and SCMAA with traditional MPA receiver over fading channel with spectral efficiency of 2 bits/ton. As can be observed in Figure 11, the BLER performance of NCMA with proposed MPA receiver outperforms that of SCMAA with traditional MPA receiver, and NCMA with proposed MPA receiver has 1.7 dB gain over SCMAA with traditional MPA receiver. As can be observed in Figures 9 and 11, the BLER performance of NCMA with proposed MPA receiver outperforms that of NCMA with traditional MPA receiver. Figure 12 is the BLER performance of NCMA with proposed MPA receiver and SCMAA with traditional MPA receiver over fading channel with spectral efficiency of 3 bits/ton. As can be observed in Figure 12, the BLER performance of NCMA with proposed MPA receiver outperforms that of SCMAA with traditional MPA receiver, and NCMA with proposed MPA receiver has 2.3 dB gain over SCMAA with traditional MPA receiver. As can be observed in Figures 10 and 12, the BLER performance of NCMA with proposed MPA receiver outperforms that of NCMA with traditional MPA receiver. Simulation results show that the proposed MPA receiver can further improve the convergence reliability of the detected layer in each decision process compared with traditional MPA receiver over fading channel.

6. Conclusions

This paper proposes a NCMA scheme. In the NCMA codebook design, different transmitted layers are generated from different complex multidimension constellations, respectively, and the proposed basic complex multidimension

constellation design increases the minimum intrapartition distance compared with the basic complex multidimension constellation design of SCMA. Then the proposed criterion of permutation set maximizes the sum of distances between interfering dimensions of transmitted codewords multiplexed on any resource node, where the number of nonzero elements of transmitted codewords is more than 1. On the other side, in each decision process, the proposed mapping matrix of NCMA and the proposed MPA receiver increase the differences between the reliabilities of detections on all undetected layers and the differences between the reliabilities of detections on the codewords of each undetected layer. In summary, benefiting from the proposed codebook design and the proposed MPA receiver, NCMA is superior to SCMA in the interlayer interference cancellation.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61501056), the Fundamental Research Funds for the Central Universities, and National Science and Technology Major Project of China (no. 2017ZX03001022).

References

- [1] HUAWEI Technologies Co. Ltd., *5G: A technology vision*, HUAWEI Technol. Co., Ltd., Shenzhen, China, 2013, http://www.huawei.com/ilink/en/download/HW_314849.
- [2] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 332–336, IEEE, London, UK, September 2013.
- [3] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proceedings of the 80th IEEE Vehicular Technology Conference, VTC 2014-Fall*, Canada, September 2014.
- [4] H. Nikopour and M. Baligh, "Systems and Methods for Sparse Code Multiple Access," *United States, US 2014/0140360 A1*, Article ID 0140360, 2014.
- [5] H. Nikopour, E. Yi, A. Bayesteh et al., "SCMA for downlink multiple access of 5G wireless networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '14)*, pp. 3940–3945, Austin, Tex, USA, December 2014.
- [6] J. Van De Beek and B. M. Popović, "Multiple access with low-density signatures," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, pp. 1–6, December 2009.
- [7] C. Yan, G. Kang, and N. Zhang, "A Dimension Distance-Based SCMA Codebook Design," *IEEE Access*, vol. 5, pp. 5471–5479, 2017.
- [8] Y. Ding, "Constellation Mapping of MPSK in BICM-ID," *Communication Technology*, vol. 41, no. 9, pp. 72–74, 2008.
- [9] J. Xiangdong, Y. Ouyang, and W. Xie, "Research on Decoding Algorithm for LDPC-COFDM Wireless Communication System," *Communications Technology*, vol. 5, pp. 12–15, May 2007.
- [10] B. D. Jelicic and S. Roy, "Design of Trellis Coded QAM for Flat Fading and AWGN Channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 192–201, 1995.
- [11] J. Boutros and E. Viterbo, "Signal space diversity: a power- and bandwidth-efficient diversity technique for the Rayleigh fading channel," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 44, no. 4, pp. 1453–1467, 1998.
- [12] G. D. Forney and L.-F. Wei, "Multidimensional Constellations-Part I: Introduction. Figures of Merit, and Generalized Cross Constellations," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 6, pp. 877–892, 1989.
- [13] L. Rong, Q. Zhuang, and J. Yin, "Turbo TCM 8CPFSK Research and Realization Based on Turbo TCM," *Journal of China Academy of Electronics and Information Technology*, vol. 2, no. 4, pp. 427–438, 2007.
- [14] K. V. Ravi, Tam Soh Khum, and H. K. Garg, "Performance of turbo TCM in wideband CDMA indoor mobile applications," in *Proceedings of the 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2000)*, pp. 898–902, September 2000.
- [15] J. H. Conway and N. J. Sloane, *Sloane, Sphere Packings, Lattices and Groups*, Springer-Verlag, 2016.
- [16] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.
- [17] B. Xiao, K. Xiao, S. Zhang, Z. Chen, B. Xia, and H. Liu, "Iterative detection and decoding for SCMA systems with LDPC codes," in *Proceedings of the International Conference on Wireless Communications and Signal Processing, WCSP 2015*, chn, October 2015.
- [18] S. Sesia, I. Toufik, and M. Baker, "LTE-the UMTS long term evolution," *Wiley Online Library*, 2015.

Research Article

On the Performance of Security-Based Nonorthogonal Multiple Access in Coordinated Multipoint Networks

Yue Tian ¹, Xianling Wang ¹ and Zhanwei Wang ²

¹Fujian Key Laboratory of Communication Network and Information Processing, School of Opto-Electronic and Communication Engineering, Xiamen University of Technology, Xiamen, China

²School of Information Engineering, Zhengzhou University, Zhengzhou, China

Correspondence should be addressed to Yue Tian; yue.tian.xmut@outlook.com

Received 10 December 2017; Revised 12 February 2018; Accepted 4 March 2018; Published 4 April 2018

Academic Editor: Imran S. Ansari

Copyright © 2018 Yue Tian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The conventional nonorthogonal multiple access (NOMA) strategy has secrecy challenge in coordinated multipoint (CoMP) networks. Under the secrecy considerations, this paper focuses on the security-based NOMA system, which aims to improve the physical layer security issues of conventional NOMA in the coordinated multipoint (NOMA-CoMP) networks. The secrecy performance of S-NOMA in CoMP, that is, the secrecy sum-rate and the secrecy outage probability, is analysed. In contrast to the conventional NOMA (C-NOMA), the results show that the proposed S-NOMA outperforms C-NOMA in terms of the secrecy outage probability and security-based effective sum-rate.

1. Introduction

As a promising candidate for the multiple access schemes in the fifth-generation (5G) mobile system, nonorthogonal multiple access (NOMA) has received widespread attention [1]. In contrast to the conventional orthogonal multiple access, NOMA provides higher capacity and better energy efficiency and supports massive connectivity by enabling users to use the same time, frequency, and code resources for information conveying [2–4]. Coordinated multipoint (CoMP) is one of the promising enhancements in LTE-A, owing to its ability to improve the coverage of high data rate, increase the system throughput, and control the co-channel interference [5, 6]. However, there was a challenge in the downlink of the CoMP network; that is, if an access point allocates a channel to a cell edge user, this channel cannot be used by other users at the same time [7]. Thus, the spectral effectiveness of the CoMP system degrades when the cell edge users increase in number. Recently, studies showed that, by introducing NOMA in coordinated multipoint (CoMP, i.e., a key enhancement for LTE-Advance) networks, not only can the resource efficiency of the whole network be further enhanced [7], but also the complexity of NOMA can be reduced by using proper user-scheduling strategy [8]. Apart from [7, 8], there are

several research contributions in the context of improving the cooperative networks performance by using NOMA [9–11]. In [9], the researchers investigated the outage performance of the relay-aided NOMA downlink and compared it to the conventional OMA strategy. In [10], Tian et al. conceived a user-relay-based multitier NOMA strategy, which aimed to improve the coverage of CoMP network. Reference [11] investigated the performance of multicell MIMO-NOMA networks, applying coordinated beamforming for dealing with the intercell interference in order to enhance the cell-edge users' throughput.

However, it should be noted that, due to the fact that the multiple users' signals in NOMA are sharing the same resource channels, the NOMA-CoMP network is confronted with a security issue; that is, the eavesdroppers in the NOMA-CoMP network could listen to the legal users' messages, which creates a secrecy challenge in the physical layer communications.

The physical layer security issue in wireless channel was proposed by Wyner in 1975 [12] and has been researched in diverse scenarios [13–16]. In spite of tremendous research in NOMA, very few existing NOMA researches focus on the security issues in NOMA transmissions. In [17], the authors investigated the maximization of the secrecy sum-rate of

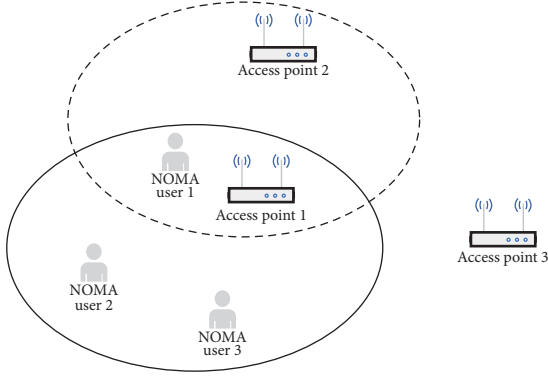


FIGURE 1: System model of downlink CoMP network.

NOMA in a single-input-single-output (SISO) network. In [18], an optimal design of decoding order for NOMA is proposed under the secrecy considerations; the authors showed that the proposed decoding strategy is an optimal solution for NOMA when considering the secrecy outage constraints. Reference [19] investigated the physical layer security of NOMA in the large-scale networks; the authors showed that the security issues of NOMA can be alleviated by generating a protected zone for the legal users and by creating artificial noise at the transmitters. In this paper, to enhance the physical layer security of NOMA-CoMP system, we proposed a security-based NOMA (S-NOMA) strategy, which combines the coordinated user-scheduling strategy and minimise-information-leakage-based joint transmission strategy in the downlink of a CoMP network. In contrast to the conventional NOMA (C-NOMA), we show that the proposed S-NOMA outperforms C-NOMA in terms of the secrecy outage probability and security-based effective sum-rate.

2. System Model

In this paper, we consider that the proposed CoMP network includes B APs with M antennas each and K single-antenna users, where the sets are defined as $\mathbf{B} = \{1, 2, \dots, B\}$, $\mathbf{M} = \{1, 2, \dots, M\}$, and $\mathbf{K} = \{1, 2, \dots, K\}$. In such CoMP network, define \mathbf{S}_i as the set of APs whose service area can cover the user i ; the set of users in the coverage area of AP b is given by \mathbf{W}_b ; for example, in Figure 1, the circle with the solid line denotes AP 1's service area, and the APs that can cover user 1 are in the circle with the dotted line.

Here we assume that, for any user $i \in \mathbf{K}$, it can be served and coordinated by more than one AP, that is, $\text{card}(\mathbf{S}_i) > 1$. In such NOMA-CoMP system, the observation at the user i can be expressed as

$$y_i = \sum_{b \in \mathbf{S}_i} \sum_{m \in \mathbf{M}} h_{i,b}^m \left(\sum_{j \in \mathbf{W}_b} \sqrt{a_j} P s_j \right) + n_i, \quad (1)$$

where a_i denotes the superposition coding (SC) power allocation coefficient of the signal s_i , P denotes normalized transmission power, and n_i denotes the additive white Gaussian

noise at the user i . The point-to-point channel from the m th antenna at an AP b to a user i is given by $\mathbf{h}_{i,b}^m$. For any AP $b \in \mathbf{B}$ and user $i \in \mathbf{K}$, the channel $h_{i,b}^m$ is considered as a Rayleigh fading channel, where $h_{i,b}^m = \sqrt{\alpha_{i,b}^m} g_{i,b}^m$; the factor $g_{i,b}^m$ is the independent and identically distributed circular symmetric complex Gaussian random variable (RV) with zero mean and variance σ_i^2 , representing fast fading; the factor $\alpha_{i,b}^m$ denotes the slow fading. The MISO channel from the AP b to the user i is denoted by $\mathbf{h}_{i,b}$, where $\mathbf{h}_{i,b} \in \mathbb{C}^{1 \times M}$.

3. Security-Based NOMA Strategy in CoMP Networks

During the broadcast transmissions in NOMA-CoMP [8], the set of users, whose observation contains the message of the user i , is denoted by

$$\mathbf{O}_i = \bigcup_{i \in \mathbf{S}_i} (\mathbf{W}_i). \quad (2)$$

Define \mathbf{E}_i as a subset of \mathbf{O}_i , where the users in such subset can be trusted by the user i . Let the complementary set \mathbf{E}_i^c denote the users that cannot be trusted by the user i ; that is, the users in the set \mathbf{E}_i^c are suspected as the eavesdropper to the user i . In the conventional NOMA-CoMP, the message of user i can also be obtained by the users in \mathbf{E}_i^c . To reduce the risk of information leakage, that is, to avoid the users in set \mathbf{E}_i^c monitoring user i 's message, a security-based NOMA (S-NOMA) strategy is proposed.

The basic idea of the S-NOMA strategy is to minimise the leakage of the signals from a target user to its untrusted users during the NOMA transmission process. Define ϵ_i as a channel quality threshold value at the user i ; the S-NOMA strategy can be implemented via the following processes.

Step 1 (security-based AP selection strategy and beamforming design). Assume that the power allocation range at transmitters is from P_{\min} to P_{\max} , $\forall b \in \mathbf{B}$ and $\forall j \in \mathbf{E}_i^c$, if $\epsilon_j > P_{\max} \|\mathbf{h}_{i,b}\|^2 \geq \epsilon_i$, add the node's index to \mathbf{S}'_i , where $\|\cdot\|$ denotes the Frobenius norm. \mathbf{S}'_i is user i 's preferred AP set considering the security problem. If AP b belongs to \mathbf{S}'_i , it means that such AP cannot be seen by the untrusted user i ; then it can be selected to transmit signal to the target user i , and this AP can be considered as an unconditioned secure AP to the user i . Therefore, if $\text{card}(\mathbf{S}'_i) \geq 1$, user i can be served by at least one unconditioned secure AP; if $\text{card}(\mathbf{S}'_i) < 1$, that means user i cannot be served by the unconditioned secure AP; then the minimise-information-leakage-based strategy should be used to prevent the untrusted users of user i .

Let \mathbf{v}_i denote the precoding matrix from the APs in set \mathbf{S}'_i to user i . If $\text{card}(\mathbf{S}'_i) \geq 1$, set $\mathbf{v}_i = a_i \mathbf{I}_i$, where a_i denotes the power allocation coefficient to user i (i.e., the NOMA superposition coding (SC) coefficient to user i [4]) and \mathbf{I}_i denotes an all-ones matrix, where $\mathbf{I}_i \in \mathbb{C}^{M \times 1}$. If $\text{card}(\mathbf{S}'_i) = 0$, then set $\mathbf{S}'_i = \mathbf{S}_i$ and let $\mathbf{v}_i = a_i \cdot \phi_{i,b}$, where $\phi_{i,b} \in \mathbb{C}^{M \times 1}$ denotes a minimum-leakage-based precoding vector that aims to minimise information leakage from the

AP b to user i . Let SLNR denote the value of signal-to-leakage-plus-noise ratio (SLNR), which is given by $\text{SLNR}_{i,b} = \|\mathbf{h}_{i,b}\phi_{i,b}\|^2 \{\text{tr}[\phi_{i,b}(\phi_{i,b})^H] + \sum_{j \in E_i^c} \|\mathbf{h}_{i,b}\phi_{i,b}\|^2\}^{-1}$ [20]; then, for the AP b in set \mathbf{S}_i , $\phi_{i,b}$ can be achieved as

$$\phi_{i,b} = \arg \max_{\phi_{i,b} \in \mathbb{C}^{M \times 1}} (\text{SLNR}_{i,b}), \quad (3)$$

where $\text{tr}(\cdot)$ indicates the trace.

Step 2 (superposition coding design). Considering that the precoding matrix is normalised, then, by considering the fairness of cell-edge users and the complexity of successive-interference-cancellation (SIC) strategy in NOMA-CoMP [8], the order of SC coefficients is designed based on the following conditions: (a) if $\text{card}(\mathbf{S}'_1) \geq \dots \geq \text{card}(\mathbf{S}'_K)$, then $a_1 \geq \dots \geq a_K$; (b) for the case that $\text{card}(\mathbf{S}'_1) = \dots = \text{card}(\mathbf{S}'_K)$, the order of SC coefficients will be sorted based on $\hat{\mathbf{h}}_k$, where $\hat{\mathbf{h}}_k = \sum_{b \in \mathbf{S}'_k} \mathbf{h}_{k,b}$; that is, for the case that $\text{card}(\mathbf{S}'_1) = \dots = \text{card}(\mathbf{S}'_K)$, if $\|\hat{\mathbf{h}}_1\| \leq \dots \leq \|\hat{\mathbf{h}}_K\|$, then $a_1 \geq \dots \geq a_K$.

Step 3 (NOMA broadcasting and SIC decoding). Define s_i as user i 's desired message. $\forall b \in \mathbf{B}$, AP b broadcasts a combined signal $\sum_{i \in \mathbf{W}_b} \sqrt{a_i} P s_i$, where P is a normalised transmit power. $\forall i \in \mathbf{K}$, user i decodes its observations based on SIC.

4. Performance Analysis of S-NOMA in CoMP Network

The physical layer reaches to secure transmissions when the capacity of the legal user channel is higher than that of the eavesdropper channel [12]; that is, the data can be transmitted at a rate equal to the capacity difference between the legal user and eavesdropper channel capacity. Via the S-NOMA strategy, the SINR at user i , γ_i , is given by

$$\gamma_i = \frac{\sum_{b \in \mathbf{S}'_i} \|\mathbf{h}_{i,b} \mathbf{v}_i\|^2}{\sum_{b \in \mathbf{S}'_i} \|\mathbf{h}_{i,b}\|^2 \sum_{m \in \mathbf{O}_i} \|\mathbf{v}_m\|^2 + \beta_i}, \quad (4)$$

where $a_m < a_i$, $\beta_i = |\varphi_i|^2 + P^{-1}\sigma_i^2$, and φ_i is the interference from the APs in the set $\{\mathbf{B} \setminus \mathbf{S}'_i\}$.

Remark 1. It should be noted that φ_i can be very small, as the power range of φ_i is limited by the threshold value; that is, for all $b' \in \{\mathbf{B} \setminus \mathbf{S}'_i\}$, there is $P_{\max} \|h_{i,b'}\|^2 < \epsilon_i$.

$$\begin{aligned} \mathbb{P}_i^s &= 1 - \int_0^{\theta_j} \int_0^{((\theta_i+1)x+\theta_i)/2} \frac{x^{Y_j/2-1} e^{-x/2} t^{Y_i/2-1}}{2^{Y_j/2} \Gamma(Y_j/2) \Gamma(Y_i/2)} e^{-t} dt dx \\ &\approx 1 - \frac{\Gamma(Y_i/2) \theta_j^{Y_j/2} (\theta_j + \theta_j \theta_i)^{Y_i/2} (\Gamma(Y_i + Y_j/2) - \Gamma(Y_i + Y_j/2, \theta_j + \theta_j \theta_i/2))}{2^{Y_j/2} \Gamma(Y_j/2) e^{\theta_i/2} (\theta_j (2 + \theta_i))^{(Y_i+Y_j)/2}}. \end{aligned} \quad (11)$$

For the nonideal case, the message of user q , where $q \in \{\mathbf{W}_b \cap E_i, b \in \mathbf{S}'_i\}$ and $a_q > a_i$, may not be successfully

The secrecy rate of user i can be expressed as

$$I_i = \log_2(1 + \gamma_i) - \log_2(1 + \gamma_{j'}), \quad (5)$$

where j' denotes the user with a maximum data rate in set \mathbf{E}_i^c , given as

$$j' = \arg \max_{j \in \mathbf{E}_i^c} [\log_2(1 + \gamma_j)]. \quad (6)$$

The channels from the APs in set \mathbf{S}'_i to user i , that is, $\sum_{b \in \mathbf{S}'_i} \|\mathbf{h}_{i,b}\|^2$, satisfy the chi-square distribution, where the probability density function (PDF) is given by

$$f_i(x) = \frac{x^{Y_i/2-1} e^{-x/2}}{2^{Y_i/2} \Gamma(Y_i/2)}, \quad (7)$$

where $Y_i = \text{card}(\mathbf{S}'_i)$ and $\Gamma(\cdot)$ is the gamma function. Define R_i as the target rate at user i , and let $\theta_i = 2^{R_i} - 1$.

Then security outage will happen when $I_i < R_i$, which means that the security capacity of user i does not meet the requirement. Here we define the security outage event as $E_i = \{I_i < R_i\}$, where

$$\begin{aligned} E_i &= \{I_i < R_i\} = \left\{ \frac{\gamma_i - \gamma_{j'}}{1 + \gamma_{j'}} < \theta_i \right\} \\ &= \{\gamma_i - (\theta_i + 1) \gamma_{j'} < \theta_i\}. \end{aligned} \quad (8)$$

Here we firstly consider an ideal case, where, for all $q \in \{W_l : l \in \mathbf{S}'_i\}$ and $a_q > a_i$, the message of a user q can be successfully detected by user i ; then, according to (5) and (8), the secrecy outage probability (SOP) at user i is given by [19]

$$\mathbb{P}_i^s = \Pr(E_i) = \int_0^\infty f_{\gamma_{j'}}(x) F_{\gamma_i}((\theta_i + 1)x + \theta_i) dx, \quad (9)$$

where the CDF $F_{\gamma_i}(x)$ is given by

$$F_{\gamma_i}(x) = \Gamma^{-1}\left(\frac{Y_i}{2}\right) \int_0^{x/2} t^{Y_i/2-1} e^{-t} dt. \quad (10)$$

Then the SOP in such NOMA-CoMP system can be derived as

detected by the target user i . The generalized SOP for the nonideal case is given by

$$\mathbb{P}_i = 1 - (1 - \mathbb{P}_i^o)(1 - \mathbb{P}_i^s) = \mathbb{P}_i^o + \mathbb{P}_i^s - \mathbb{P}_i^o\mathbb{P}_i^s, \quad (12)$$

where \mathbb{P}_i^o denotes the normal outage probability that user i cannot successfully detect its own message.

Let $R'_{i,q}$ denote user i 's target data rate when detecting the message of user q . Then the outage event may happen when $R'_{i,q} < R_i$. Here we define $E'_{i,q}$ as such event, where

$$E'_{i,q} = \left\{ \frac{\sum_{b \in \mathbf{S}'_i} \|\mathbf{h}_{i,b}\|^2 \mathbf{v}_q}{\sum_{b \in \mathbf{S}'_i} \|\mathbf{h}_{i,b}\|^2 \sum_{m \in \mathbf{O}_i} \|\mathbf{v}_m\|^2 + \beta_i} < \theta_q \right\} \quad (13)$$

$$\stackrel{(a)}{=} \left\{ \sum_{b \in \mathbf{S}'_i} \left\| \sqrt{d_{i,b}} \mathbf{h}_{i,b} \right\|^2 < u_n \right\},$$

where $d_i = (1 - \theta_q \sum_{m \in \mathbf{O}_i} (\|\mathbf{v}_q\|^2 / \|\mathbf{v}_m\|^2))$ and $u_n = \theta_n (\beta_k \|\mathbf{v}_n\|^2)^{-1}$. Note that step (a) follows the condition that, for all $b \in \mathbf{S}'_i$, there exist $d_i > 0$.

Let $\|\mathbf{H}_{i,b}\|^2 = \|\sqrt{d_{i,b}} \alpha_{i,b} \mathbf{g}_{i,b}\|^2$; $\|\mathbf{H}_{i,b}\|^2$ can be regarded as the generalized chi-square distribution [21, 22] with variance $d_{i,b} \sigma_{i,b}^2$.

The PDF of the unordered generalized chi-square random variable $\|\mathbf{H}_{i,b}\|^2$ is given by

$$f_i(x) = \frac{(\gamma_{i,b})^{Y_i} x^{Y_i-1} \exp(-\gamma_{i,b}x)}{\Gamma(Y_i)}, \quad (14)$$

where $\Gamma(\cdot)$ denotes the gamma function. Define

$$L_b = \text{card} \{ \mathbf{W}_b \cap E_i, b \in \mathbf{S}'_i \}, \quad (15)$$

and assume that s_i 's decoding order at user i is $l_b(i)$.

Define

$$\phi = \max \{ u_1, \dots, u_i \}. \quad (16)$$

Based on the high-order statistics in [20], the normal outage probability of user i , \mathbb{P}_i^o , can be derived as

$$\mathbb{P}_i^o = \int_0^\phi \frac{L_b! f(x) (F(x))^{l_b(i)-1} (1-F(x))^{L_b-l_b(i)}}{(l_b(i)-1)! (L_b-l_b(i))!} \mathbf{d}x \quad (17)$$

$$\stackrel{(b)}{=} \sum_{j=0}^{L_b-l_b(i)} \frac{(-1)^j (F_i(\phi))^{l_b(i)+j} L_b!}{(L_b-l_b(i))! (l_b(i)-j)! j! \gamma_{i,b} (l_b(i)+j)},$$

where step (b) follows the power series of exponential functions. The cumulative distribution function (CDF) $F_i(\phi)$ can be derived by integration of PDF $f_i(x)$ as

$$F_i(\phi) = \int_0^\phi \frac{(d_{i,b} \sigma_{i,b}^2)^{Y_i} x^{Y_i-1} \exp(-d_{i,b} \sigma_{i,b}^2 x)}{\Gamma(Y_i)} \mathbf{d}x \quad (18)$$

$$\stackrel{(c)}{=} 1 - \sum_{t=0}^{Y_i-1} \frac{(d_{i,b} \sigma_{i,b}^2 \phi)^t \exp(-d_{i,b} \sigma_{i,b}^2 \phi)}{\Gamma(t+1)},$$

where step (c) follows the power series of exponential functions. The similar derivation process of (17) and (18) can be seen in our previous research outputs [8].

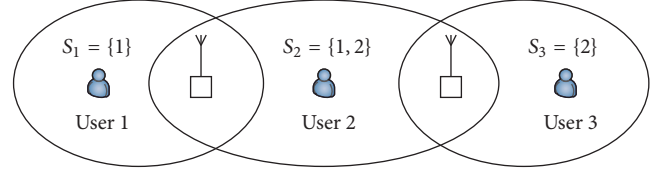


FIGURE 2: Service area of the two APs, that is, $S_1 = \{1\}$, $S_2 = \{1, 2\}$, and $S_3 = \{2\}$.

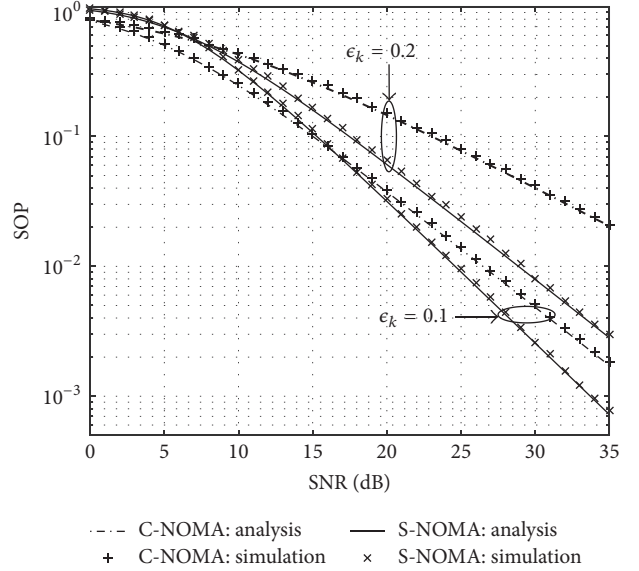


FIGURE 3: Secrecy outage probability of NOMA-CoMP as a function of SNR.

5. Performance Evaluation

In this section, to evaluate the performance of S-NOMA, we consider a CoMP network that contains two APs with two antennas each ($M = 2$) and three users. Let \mathbf{B} and \mathbf{K} denote the APs and users, respectively.

In such CoMP network, assume that the locations of the three users are randomly distributed in the service area of the two APs; that is, for a user i , if $\mathbf{S}'_i = \{1\}$ or $\mathbf{S}'_i = \{2\}$, then $\|\mathbf{h}_{i,b}\|^2 > \epsilon_i > \|\mathbf{h}_{i,b_0}\|^2$, where $b \in \mathbf{S}'_i$ and $b_0 \in \{\mathbf{B} \setminus b\}$; if $\mathbf{S}'_i = \{1, 2\}$, then, for all $b \in \mathbf{B}$, there exist $\|\mathbf{h}_{i,b}\|^2 > \epsilon_i$. Assume that, in this CoMP network, user 2 can be trusted by user 3 but will not be trusted by user 1; user 1 and user 3 are not trusted by each other; that is, $\mathbf{E}_1^c = \{2, 3\}$, $\mathbf{E}_2^c = \{1\}$, and $\mathbf{E}_3^c = \{1\}$. The service area of the two APs is considered in Figure 2.

Let l_i denote the SIC decoding order of s_i ; then the SC coefficient allocated to s_i is given by

$$a_i = (\text{card}(\mathbf{K}) - l_i + 1) c^{-1}, \quad (19)$$

where c denotes a constant to ensure that $\sum_{i \in \{\mathbf{W}_b, b \in \mathbf{B}\}} a_i = 1$. Figure 3 provides a comparison of the SOP between the conventional NOMA (C-NOMA) and the S-NOMA with different threshold values; that is, $\epsilon_k = 0.1$ and $\epsilon_k = 0.2$. The target data rate of each user is set to $R_k = 1$ bit per channel

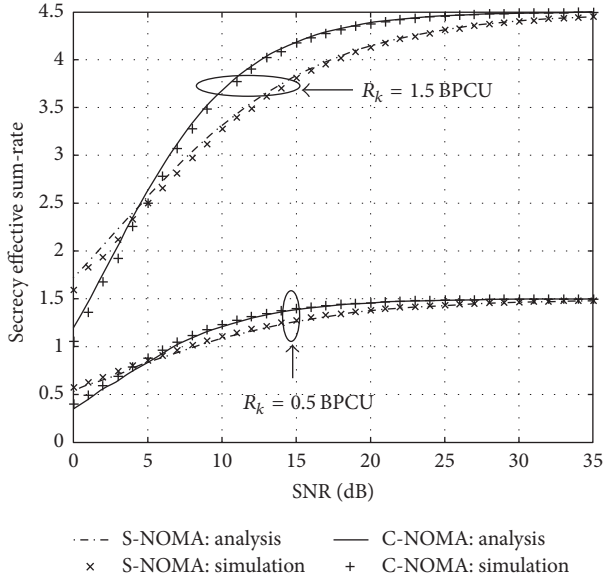


FIGURE 4: Secrecy effective sum-rate as a function of SNR under different data rate, R .

use (BPCU). In contrast to the C-NOMA, the proposed S-NOMA strategy shows better SOP performance, especially at a high value of ϵ_k . In Figure 3, it should be noted that when SNR is low, the noise will impact on the SLNR-based precoding vector; therefore the S-NOMA's performance is somehow lower than the C-NOMA's; when SNR is high, the performance of S-NOMA is significantly improved.

Figure 4 compares the security-based effective sum-rate (SESR) between S-NOMA and C-NOMA with different data rate. Let $\mathbb{P}_i(R_k)$ denote the SOP at user k under data rate R_k ; the SESR is defined as

$$R_{\text{eff}} = \sum_{k \in \mathbf{K}} R_k \mathbb{P}_i(1 - R_k). \quad (20)$$

The threshold value is set to $\epsilon_k = 0.1$. It can be concluded that when SNR is low (less than 5 dB), the effective sum-rate of S-NOMA is lower than the C-NOMA, as the outage probability of S-NOMA is worse than C-NOMA in this case. The performance gain that is provided by the S-NOMA becomes significant when SNR is larger than 5 dB, especially under the higher target data rate. When SNR is larger than 30 dB, the effective sum-rates of the two schemes are very close, because the outage probability is very small under the high SNR.

Remark 2. If a user m is regarded as an untrusted user of user n , then the data rate of detecting the message of user n at user m will be considered as the eavesdropping data rate. Therefore, the capacity of C-NOMA (the secrecy capacity) is smaller than the CoMP system capacity when considering the physical layer security issue. In contrast to the C-NOMA, the S-NOMA strategy reduces the leakage of information to the untrusted users; therefore the secrecy capacity of S-NOMA is more close to the upper bond capacity (the system capacity), so S-NOMA has a better performance than C-NOMA in

terms of secrecy outage probability and secrecy effective sum-rate.

6. Conclusion

This paper focuses on the performance of NOMA-CoMP under secure considerations. We proposed a security-based NOMA strategy, which aims to improve the physical layer security issues in the conventional NOMA-CoMP networks. The secrecy performance of the proposed S-NOMA in CoMP, that is, the secrecy sum-rate and the secrecy outage probability, is analysed and evaluated. In contrast to the conventional NOMA, the results show that the proposed S-NOMA has advantages over C-NOMA, especially when the target transmission data rate is high.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province, China (Grant no. 2016J01323).

References

- [1] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.
- [2] 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.
- [3] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource Allocation in NOMA based Fog Radio Access Networks," *IEEE Wireless Communications*, 2018.
- [4] 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 the IEEE 77th Vehicular Technology Conference (VTC '13)*, pp. 1–5, Dresden, Germany, June 2013.
- [5] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-advanced," *IEEE Wireless Communications Magazine*, vol. 17, no. 3, pp. 26–34, 2010.
- [6] V. Jungnickel, K. Manolakis, W. Zirwas et al., "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 44–51, 2014.
- [7] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Communications Letters*, vol. 18, no. 2, pp. 313–316, 2014.
- [8] Y. Tian, A. R. Nix, and M. Beach, "On the Performance of Opportunistic NOMA in Downlink CoMP Networks," *IEEE Communications Letters*, vol. 20, no. 5, pp. 998–1001, 2016.
- [9] J. Men and J. Ge, "Non-Orthogonal Multiple Access for Multiple-Antenna Relaying Networks," *IEEE Communications Letters*, vol. 19, no. 10, pp. 1686–1689, 2015.

- [10] Y. Tian, A. R. Nix, and M. Beach, "On the Performance of Multi-tier NOMA Strategy in Coordinated Multi-Point Networks," *IEEE Communications Letters*, 2017.
- [11] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Communications Letters*, vol. 21, no. 1, pp. 84–87, 2017.
- [12] A. D. Wyner, "The wire-tap channel," *Bell Labs Technical Journal*, vol. 54, no. 8, pp. 1355–1387, 1975.
- [13] Z. Ding, K. K. Leung, D. L. Goeckel, and D. Towsley, "On the application of cooperative transmission to secrecy communications," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 359–368, 2012.
- [14] Y. Zou, X. Wang, W. Shen, and L. Hanzo, "Security versus reliability analysis of opportunistic relaying," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 6, pp. 2653–2661, 2014.
- [15] A. Mukherjee and A. L. Swindlehurst, "Robust beamforming for security in MIMO wiretap channels with imperfect CSI," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 351–361, 2011.
- [16] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource Allocation in NOMA based Fog Radio Access Networks," in *IEEE Wireless Communications*, Early Access, 2018.
- [17] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy Sum Rate Maximization in Non-orthogonal Multiple Access," *IEEE Communications Letters*, vol. 20, no. 5, pp. 930–933, 2016.
- [18] B. He, A. Liu, N. Yang, and V. K. N. Lau, "On the Design of Secure Non-Orthogonal Multiple Access Systems," *IEEE Communications Letters*, vol. 35, no. 10, pp. 2196–2206, 2017.
- [19] Y. Liu, Z. Qin, M. ElKashlan, Y. Gao, and L. Hanzo, "Enhancing the Physical Layer Security of Non-Orthogonal Multiple Access in Large-Scale Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1656–1672, 2017.
- [20] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1711–1721, 2007.
- [21] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Acquiring partial CSI for spatially selective transmission by instantaneous channel norm feedback," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1188–1204, 2008.
- [22] J. E. Gentle, *Computational Statistics*, Springer, New York, NY, USA, 2009.

Research Article

An Efficient SCMA Codebook Optimization Algorithm Based on Mutual Information Maximization

Chao Dong , Guili Gao, Kai Niu, and Jiaru Lin

Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Chao Dong; dongchao@bupt.edu.cn

Received 21 November 2017; Revised 2 February 2018; Accepted 26 February 2018; Published 1 April 2018

Academic Editor: Imran S. Ansari

Copyright © 2018 Chao Dong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An efficient codebook optimization algorithm is proposed to maximize mutual information in sparse code multiple access (SCMA). At first, SCMA signal model is given according to superposition modulation structure, in which the channel matrix is column-extended. The superposition model can well describe the relationship between the codebook matrix and received signal. Based on the above model, an iterative codebook optimization algorithm is proposed to maximize mutual information between discrete input and continuous output. This algorithm can efficiently adapt to multiuser channels with arbitrary channel coefficients. The simulation results show that the proposed algorithm has good performance in both AWGN and non-AWGN channels. In addition, message passing algorithm (MPA) works well with the codebook optimized according to the proposed algorithm.

1. Introduction

Nonorthogonal multiple access (NOMA) [1–3] has attracted much attention from both industry and academia and has become the key technique in 5G. NOMA can accommodate more users than orthogonal multiple access (OMA) and improve the spectral efficiency. At the same time, the interuser interference is introduced in NOMA and more efficient detection is required at the receiver.

Code-domain NOMA is considered as an important candidate technique [3]. At first, low density spreading CDMA (LDS-CDMA), which belongs to code-domain NOMA, is proposed in [4], where the spreading sequence is sparse and message passing algorithm (MPA) is introduced to lower multiuser detection complexity. In [5], message passing algorithm (MPA) is proven to be optimal when the length of low density spreading sequence tends to be infinity. Later, in [6], LDS-OFDM is proposed, where the iteration number of MPA is optimized according to extrinsic information transfer (EXIT) chart [7]. In [8], LDPC codes and LDS constitute three-layer factor graph and the joint belief propagation algorithm exhibits good performance. In [9], the progressive edge growth (PEG) algorithm is introduced to improve LDS pattern and the multiuser detection performance becomes better.

Recently, sparse code multiple access (SCMA) [10, 11] has been proposed to combine low density spreading (LDS) with codebook design. In [10], the codebook design is implemented by rotating the phases of different users' constellation points. At the receiver, message passing algorithm (MPA) is applied and SCMA exhibits better performance than LDS-CDMA [10]. Afterwards, several methods to improve MPA in SCMA are proposed in [12–14]. Furthermore, in the recent paper [15], SCMA detection can be seen as a specific tree search problem and sphere decoding algorithm shows much lower complexity with negligible performance loss. Inspired by transmit signal optimization in two-user multiple access channels [16], SCMA codebook design is mainly focused on phase rotation optimization in [17, 18]. In addition, the product distance and cutoff-rate are introduced to optimize SCMA codebook in [19] and [20], respectively.

In this paper, SCMA signal model is given according to superposition modulation structure, where the channel matrix is column-extended. The proposed model can well describe the relationship between the codebook of each user and received signal. Therefore, the codebook optimization algorithm is derived according to maximizing mutual information between the discrete input and continuous output. Our analysis shows that the proposed algorithm can

efficiently adapt to multiuser channels with random channel coefficients.

This paper is organized as follows. In Section 2, SCMA signal model is represented as superposition modulation structure, where the channel matrix is extended in the column space to explicitly demonstrate the relationship between the codebook and received signal. Afterwards, the principles of codebook optimization to maximize mutual information are described in Section 3. The concrete expressions of mutual information and its gradient are given. Subsequently, Karush-Kuhn-Tucker (KKT) conditions are introduced to realize codebook optimization. In Section 4, the implementation steps of the proposed iterative codebook optimization algorithm are elaborated. The simulation results are given in Section 5. Section 6 draws the conclusions.

In the following parts, lower and upper boldface letters denote the vector and matrix, respectively. For the matrix \mathbf{A} , \mathbf{A}^{-1} , \mathbf{A}^T , and \mathbf{A}^H denote its inverse, transpose, and Hermitian, respectively. The row vector $\mathbf{e}_{K,j}$ denotes the j th row of the $K \times K$ identity matrix \mathbf{I}_K . The matrix $\bar{\mathbf{A}} = \text{blkdiag}\{\mathbf{A}_1, \dots, \mathbf{A}_N\}$ denotes the block diagonal matrix, in which \mathbf{A}_i is the submatrix on the i th diagonal block. The operator \otimes denotes Kronecker product. In addition, $E_{\mathbf{x}}[\cdot]$ denotes the expectation over the random variable \mathbf{x} .

2. SCMA Signal Model with Superposition Modulation

In this section, SCMA signal model is given according to superposition modulation structure. The analysis shows that the channel matrix is column-extended in the proposed structure. In addition, the relationship between the codebook and received signal is given. The analysis in this section lays foundation for the codebook optimization. For clearness, the typical SCMA signal model is detailed in the first subsection.

2.1. Typical SCMA Model. A typical SCMA factor graph is given in Figure 1. In this paper, K denotes the number of multiple access users and N denotes the number of subchannels. In factor graph, the user node is usually called “variable node” and the subchannel is usually called “function node.” The variable node degree d_v denotes the number of subchannels occupied by one user. Meanwhile, the function node degree d_f denotes the number of users carried by one subchannel.

In the factor graph shown in Figure 1, d_v is equal to 2 and d_f is equal to 3. The load is equal to K/N . The mapping between variable nodes and function nodes in Figure 1 is given by

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}. \quad (1)$$

The matrix \mathbf{F} has K columns and N rows. It can be seen that the j th row of \mathbf{F} denotes the mapping from all variable nodes to the j th function node. Similarly, the i th column of

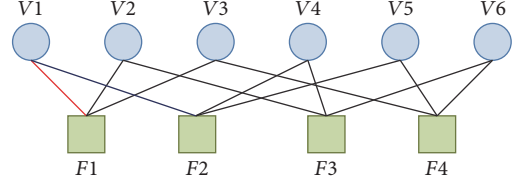


FIGURE 1: Factor graph with $K = 6$ and $N = 4$.

\mathbf{F} denotes the mapping from all function nodes to the i th variable node. In addition, the number of nonzero elements in each row is equal to d_f and the number of nonzero elements in each column is equal to d_v .

The matrix \mathbf{F} denotes the channel matrix in AWGN scenario. The existing codebook designs in [10, 11, 17, 18] are mainly based on \mathbf{F} in (1). However, in the wireless communication, the channel response amplitudes and phases of K users are usually different from each other. In the following, $h_{j,i}$ denotes the channel response of the i th user on the j th subchannel. Therefore, the channel matrix \mathbf{H}_S corresponding to the factor graph in Figure 1 can be given by

$$\mathbf{H}_S = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & 0 & 0 & 0 \\ h_{2,1} & 0 & 0 & h_{2,4} & h_{2,5} & 0 \\ 0 & h_{3,2} & 0 & h_{3,4} & 0 & h_{3,6} \\ 0 & 0 & h_{4,3} & 0 & h_{4,5} & h_{4,6} \end{bmatrix}. \quad (2)$$

Based on the channel matrix in (2), SCMA codebook optimization proposed in [10] can be denoted by

$$\mathbf{G}^* = \arg \max_{\mathbf{G}} m(S(\mathbf{H}_S, \mathbf{G}, N, K, d_v, M)), \quad (3)$$

where M denotes the modulator order of each user, $S(\cdot)$ denotes the matrix function related to the variables in (3), and $m(\cdot)$ gives performance measure in codebook optimization. In this paper, our main focus is on the factor graph in Figure 1 with $K = 6$, $N = 4$, $d_f = 3$, $d_v = 2$, and $M = 4$.

In the next subsection, the superposition modulation SCMA signal model is carefully analyzed.

2.2. Superposition Modulation Model. It can be seen from Figure 1 that user 1 is connected to function nodes $F1$ and $F2$. In SCMA, the signal of user 1 mapped to $F1$ and $F2$ is given by $x_{1,1}$ and $x_{2,1}$, respectively. For clearness, the above two signal elements are collected to generate the following signal vector:

$$\mathbf{x}_1 = [x_{1,1}, x_{2,1}]^T. \quad (4)$$

In SCMA model with $M = 4$, the signal elements $x_{1,1}$ and $x_{2,1}$ of user 1 carry the same two information bits. In [10, 11], $x_{1,1}$ and $x_{2,1}$ are generated by phase rotation of the chosen mother constellation.

In this paper, the superposition modulation structure is introduced. In [21, 22], the superposition modulation is proven to be an efficient modulation scheme to approach

the channel capacity. Based on the above analysis, \mathbf{x}_1 can be rewritten as follows:

$$\mathbf{x}_1 = \begin{bmatrix} x_{1,1} \\ x_{2,1} \end{bmatrix} = \begin{bmatrix} g_{1,1}^{(1)} & g_{1,2}^{(1)} \\ g_{2,1}^{(1)} & g_{2,2}^{(1)} \end{bmatrix} \cdot \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix} = \mathbf{G}_1 \mathbf{b}_1, \quad (5)$$

where the superscript (1) denotes the user index, \mathbf{G}_1 denotes the codebook matrix of user 1, and the bit vector \mathbf{b}_1 contains the two information bits of user 1.

By extending the above model to $K = 6$ users, the transmit signal vector of user i is given by

$$\mathbf{x}_i = \mathbf{G}_i \mathbf{b}_i, \quad 1 \leq i \leq K. \quad (6)$$

$$\overline{\mathbf{H}}_S = \begin{bmatrix} h_{1,1} & 0 & h_{1,2} & 0 & h_{1,3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & h_{2,1} & 0 & 0 & 0 & 0 & h_{2,4} & 0 & h_{2,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & h_{3,2} & 0 & 0 & 0 & h_{3,4} & 0 & 0 & h_{3,6} & 0 \\ 0 & 0 & 0 & 0 & 0 & h_{4,3} & 0 & 0 & 0 & h_{4,5} & 0 & h_{4,6} \end{bmatrix}. \quad (8)$$

For example, in the first column of \mathbf{H}_S , the channel coefficients corresponding to $x_{1,1}$ and $x_{2,1}$ are $h_{1,1}$ and $h_{2,1}$, respectively. To match the two-dimensional transmission vector \mathbf{x}_1 shown in (5), the first column of \mathbf{H}_S should be extended to generate the following two columns:

$$\begin{aligned} \mathbf{h}_{1,1} &= [h_{1,1}, 0, 0, 0]^T, \\ \mathbf{h}_{2,1} &= [0, h_{2,1}, 0, 0]^T. \end{aligned} \quad (9)$$

Based on the above analysis, the column-extended channel matrix $\overline{\mathbf{H}}_S$ is given by the long expression in (8). It can be seen from (8) that $\overline{\mathbf{H}}_S$ is obtained by dividing each column of \mathbf{H}_S into two columns in order to match the two-dimensional modulation symbol vector of each user.

Based on $\overline{\mathbf{H}}_S$ in (8), the received signal is rewritten as

$$\mathbf{y} = \overline{\mathbf{H}}_S \mathbf{x} + \mathbf{n} = \overline{\mathbf{H}}_S \overline{\mathbf{G}} \mathbf{b} + \mathbf{n}, \quad (10)$$

where \mathbf{n} is N -dimensional additive white Gaussian noise vector with distribution $\mathcal{CN}(0, \sigma_n^2 \mathbf{I}_N)$.

According to block diagonal property of $\overline{\mathbf{G}}$, the received signal \mathbf{y} in (10) can be rewritten as the superposition results of $K = 6$ users' signals:

$$\mathbf{y} = \overline{\mathbf{H}}_S \overline{\mathbf{G}} \mathbf{b} + \mathbf{n} = \sum_{i=1}^K \mathbf{H}_i \mathbf{G}_i \mathbf{b}_i + \mathbf{n}, \quad (11)$$

Therefore, the overall transmitted vector can be obtained by stacking \mathbf{x}_1 to \mathbf{x}_6 :

$$\begin{aligned} \mathbf{x} &= [\mathbf{x}_1^T, \mathbf{x}_2^T, \mathbf{x}_3^T, \mathbf{x}_4^T, \mathbf{x}_5^T, \mathbf{x}_6^T]^T = \overline{\mathbf{G}} \mathbf{b} \\ &= \text{blkdiag} \{ \mathbf{G}_1, \dots, \mathbf{G}_6 \} \times \mathbf{b}, \end{aligned} \quad (7)$$

where $\overline{\mathbf{G}}$ denotes the block diagonal codebook matrix and \mathbf{b} collects all the users' information bits with $\mathbf{b} = [\mathbf{b}_1^T, \mathbf{b}_2^T, \mathbf{b}_3^T, \mathbf{b}_4^T, \mathbf{b}_5^T, \mathbf{b}_6^T]^T$.

In order to adapt to the transmit signal expression in (7), the channel expression \mathbf{H}_S shown in (2) should be extended in the column space. It is noted that the dimension of \mathbf{x} is equal to $d_v \times K = 12$. Therefore, the column number of channel matrix should be extended from 6 to 12.

where \mathbf{H}_i is the equivalent channel matrix of user i . For example, the equivalent channel matrix \mathbf{H}_1 corresponding to user 1 is given by

$$\mathbf{H}_1 = \begin{bmatrix} h_{1,1} & 0 \\ 0 & h_{2,1} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \quad (12)$$

Based on the above analysis, the received signal \mathbf{y} is connected to the codebook matrix \mathbf{G}_i of each user with the help of the column-extended channel matrix \mathbf{H}_i , $1 \leq i \leq K$. Therefore, the codebook optimization can be implemented according to various criteria. In the next section, the codebook matrix \mathbf{G}_i , $1 \leq i \leq K$, is optimized according to maximizing mutual information between the bit vector \mathbf{b} and the received signal \mathbf{y} shown in (11).

3. Principle of SCMA Codebook Optimization

In this section, the codebook optimization to maximize mutual information is carefully analyzed. We assume that the number of users and the channel responses are known by the transmitter. In practical wireless communication systems, this assumption is possible for the downlink transmission with channel state information feedback but not possible for the uplink.

At first, the concrete expression of mutual information between the bit vector \mathbf{b} and the continuous received signal \mathbf{y} is given. Afterwards, KKT conditions based on the gradient of mutual information are introduced to realize codebook

optimization. It is shown that the gradient of mutual information with respect to codebook matrix \mathbf{G}_i , $1 \leq i \leq K$, depends on the mean squared error matrix \mathbf{E}_b . For clearness, the details of calculating mean squared error matrix \mathbf{E}_b are given in Appendix A.

3.1. Detailed Expression of Mutual Information. Similar to that in [23, 24], mutual information between discrete input \mathbf{b} and continuous output \mathbf{y} can be given by

$$\begin{aligned} I(\mathbf{b}; \mathbf{y}) &= H(\mathbf{b}) - H(\mathbf{b} | \mathbf{y}) = K \log_2 M \\ &- \sum_{\bar{m}=1}^{M^K} \int_{\mathbf{y}} p(\mathbf{b}_{\bar{m}}, \mathbf{y}) \log \frac{p(\mathbf{y})}{p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})} d\mathbf{y} = K \\ &\cdot \log_2 M - \sum_{\bar{m}=1}^{M^K} \int_{\mathbf{y}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) \\ &\cdot \log \frac{p(\mathbf{y})}{p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})} d\mathbf{y}. \end{aligned} \quad (13)$$

In this paper, all possible input vectors are assumed to have equal probability. The input constellation alphabet size is equal to M^K and $p(\mathbf{b}_{\bar{m}}) = 1/M^K$. When signal-to-noise ratio (SNR) tends to be infinity, the mutual information is not larger than the entropy $H(\mathbf{b})$, which is equal to $K \log_2 M$. The subscript \bar{m} denotes the index of the constellation point from 1 to M^K . With additive white Gaussian noise, the conditional probability distribution function $p(\mathbf{y} | \mathbf{b}_{\bar{m}})$ is given by

$$p(\mathbf{y} | \mathbf{b}_{\bar{m}}) = \frac{1}{(\pi\sigma_n^2)^N} \exp\left(-\frac{\|\mathbf{y} - \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}}\|^2}{\sigma_n^2}\right). \quad (14)$$

In addition, the probability distribution function $p(\mathbf{y})$ in (13) can be given by

$$\begin{aligned} p(\mathbf{y}) &= \sum_{\bar{k}=1}^{M^K} p(\mathbf{b}_{\bar{k}}) p(\mathbf{y} | \mathbf{b}_{\bar{k}}) \\ &= \sum_{\bar{k}=1}^{M^K} \frac{1}{M^K (\pi\sigma_n^2)^N} \exp\left(-\frac{\|\mathbf{y} - \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{k}}\|^2}{\sigma_n^2}\right), \end{aligned} \quad (15)$$

where the subscript \bar{k} denotes the constellation point index.

When the bit vector $\mathbf{b}_{\bar{m}}$ is transmitted, the received signal is given by $\mathbf{y} = \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n}$. In this case, the unknown contained in \mathbf{y} is only additive white Gaussian noise vector \mathbf{n} . Therefore, the integral of \mathbf{y} can be expressed as the integral of \mathbf{n} . Consequently, in the \bar{m} th integral of the summation in (13), \mathbf{y} is replaced by $\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n}$:

$$\begin{aligned} &\int_{\mathbf{y}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) \log \frac{p(\mathbf{y})}{p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})} d\mathbf{y} \\ &= \int_{\mathbf{n}} p(\mathbf{b}_{\bar{m}}) p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n} | \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}}) \end{aligned}$$

$$\begin{aligned} &\cdot \log \frac{p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n})}{p(\mathbf{b}_{\bar{m}}) p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n} | \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}})} d\mathbf{n} \\ &= \int_{\mathbf{n}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{n}) \log \frac{p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n})}{p(\mathbf{b}_{\bar{m}}) p(\mathbf{n})} d\mathbf{n} \\ &= \int_{\mathbf{n}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{n}) \\ &\cdot \log \frac{\sum_{\bar{k}=1}^{M^K} p(\mathbf{b}_{\bar{k}}) p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n} | \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{k}})}{p(\mathbf{b}_{\bar{m}}) p(\mathbf{n})} d\mathbf{n}, \end{aligned} \quad (16)$$

where $p(\mathbf{n}) = 1/(\pi\sigma_n^2)^N \times \exp(-\|\mathbf{n}\|^2/\sigma_n^2)$. For the first equation, we assume that the channel matrix and codebook matrix are perfectly known. With the expression of $p(\mathbf{y} | \mathbf{b}_{\bar{m}})$ in (14), the second equation is achieved. In the third equation, $p(\mathbf{y} | \mathbf{b}_{\bar{k}})$ is replaced by $p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n} | \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{k}})$, whose expression is given by

$$\begin{aligned} &p(\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}} + \mathbf{n} | \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{k}}) \\ &= \frac{1}{(\pi\sigma^2)^N} \exp\left(-\frac{\|\bar{\mathbf{H}}_S \bar{\mathbf{G}} (\mathbf{b}_{\bar{m}} - \mathbf{b}_{\bar{k}}) + \mathbf{n}\|^2}{\sigma^2}\right). \end{aligned} \quad (17)$$

Based on the above analysis, the integral value in (16) depends on the Euclidean distances between $\bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b}_{\bar{m}}$ and all the other received signal constellations.

Under the equal probability input assumption, the mutual information in (13) can be rewritten as

$$\begin{aligned} I(\mathbf{b}; \mathbf{y}) &= K \log_2 M \\ &- \frac{1}{M^K} \sum_{\bar{m}=1}^{M^K} \int_{\mathbf{n}} p(\mathbf{n}) \log \sum_{\bar{k}=1}^{M^K} \exp(-q_{\bar{m},\bar{k}}) d\mathbf{n} \\ &= K \log_2 M \\ &- \frac{1}{M^K} \sum_{\bar{m}=1}^{M^K} E_{\mathbf{n}} \left[\log \sum_{\bar{k}=1}^{M^K} \exp(-q_{\bar{m},\bar{k}}) \right], \end{aligned} \quad (18)$$

where $q_{\bar{m},\bar{k}}$ is given by

$$q_{\bar{m},\bar{k}} = \frac{\|\bar{\mathbf{H}}_S \bar{\mathbf{G}} (\mathbf{b}_{\bar{m}} - \mathbf{b}_{\bar{k}}) + \mathbf{n}\|^2 - \|\mathbf{n}\|^2}{\sigma_n^2}. \quad (19)$$

It should be noted that \bar{m} and \bar{k} are both constellation point indexes and they are independent of each other.

From the above analysis, it can be seen that mutual information is the function of codebook matrices. In the following subsection, the gradient of mutual information with respect to codebook matrix of each user is analyzed and the KKT conditions are introduced to maximize mutual information.

3.2. *KKT Conditions When Maximizing Mutual Information.* To optimize mutual information, the gradient of mutual information with respect to \mathbf{G}_i , $1 \leq i \leq K$, is calculated.

According to the results in [25, 26], the gradient with respect to the overall block diagonal codebook matrix $\overline{\mathbf{G}}$ is given. Applying Theorem 2 in [26], we have

$$\begin{aligned} \nabla_{\overline{\mathbf{G}}} I(\mathbf{b}; \mathbf{y}) &= \frac{\partial}{\partial \overline{\mathbf{G}}} I(\mathbf{b}; \mathbf{y}) = \frac{1}{\ln 2} \overline{\mathbf{H}}_S^H (\sigma_n^2 \mathbf{I}_N)^{-1} \overline{\mathbf{H}}_S \overline{\mathbf{G}} \mathbf{E}_b \\ &= \frac{1}{\ln 2 \cdot \sigma_n^2} \overline{\mathbf{H}}_S^H \overline{\mathbf{H}}_S \overline{\mathbf{G}} \mathbf{E}_b, \end{aligned} \quad (20)$$

where the factor $1/\ln 2$ is appended because the natural logarithm is applied in [26]. The matrix \mathbf{E}_b denotes the mean squared error matrix. In [26], it is proven that the above gradient expression holds for the linear received signal model in (10) regardless of the structure of the channel matrix $\overline{\mathbf{H}}_S$ and the codebook matrix $\overline{\mathbf{G}}$.

In SCMA, we assume that the codebook matrix of each user satisfies individual power constraint. This requires the gradient with respect to each user's codebook matrix \mathbf{G}_i , $1 \leq i \leq K$. Based on the fact that \mathbf{G}_i is $d_v \times d_v$ submatrix on the i th diagonal block of $\overline{\mathbf{G}}$, the gradient with respect to \mathbf{G}_i is given by

$$\begin{aligned} \nabla_{\mathbf{G}_i} I(\mathbf{b}; \mathbf{y}) &= \frac{\partial}{\partial \mathbf{G}_i} I(\mathbf{b}; \mathbf{y}) \\ &= (\mathbf{e}_{K,i} \otimes \mathbf{I}_{d_v}) \frac{\partial}{\partial \overline{\mathbf{G}}} I(\mathbf{b}; \mathbf{y}) (\mathbf{e}_{K,i} \otimes \mathbf{I}_{d_v})^H \\ &= \frac{1}{\ln 2 \cdot \sigma_n^2} (\mathbf{e}_{K,i} \otimes \mathbf{I}_{d_v}) \overline{\mathbf{H}}_S^H \overline{\mathbf{H}}_S \overline{\mathbf{G}} \mathbf{E}_b (\mathbf{e}_{K,i} \otimes \mathbf{I}_{d_v})^H, \end{aligned} \quad (21)$$

where $\mathbf{e}_{K,i}$ is the i th row of the $K \times K$ identity matrix \mathbf{I}_N . From (19), it can be seen that $\nabla_{\mathbf{G}_i} I(\mathbf{b}; \mathbf{y})$ can be easily calculated from the result of $\nabla_{\overline{\mathbf{G}}} I(\mathbf{b}; \mathbf{y})$.

In order to maximize mutual information between \mathbf{b} and \mathbf{y} , the optimization problem is given by

$$\begin{aligned} \max_{\mathbf{G}_i, 1 \leq i \leq K} \quad & I(\mathbf{b}; \mathbf{y}) \\ \text{s.t.} \quad & \text{tr}(\mathbf{G}_i \mathbf{G}_i^H) \leq P_i, \quad 1 \leq i \leq K. \end{aligned} \quad (22)$$

Unfortunately, $I(\mathbf{b}; \mathbf{y})$ is not a convex function of the codebook matrix \mathbf{G}_i , $1 \leq i \leq K$, and it is difficult to calculate its globally optimal solution. An efficient method to solve this kind of problem is to find locally optimal solution according to KKT conditions [22]. Therefore, we have the following lemma.

Lemma 1. *With the power constraint of each user, the KKT conditions corresponding to problem (22) are given by*

$$\begin{aligned} \lambda_i \mathbf{G}_i &= \nabla_{\mathbf{G}_i} I(\mathbf{b}; \mathbf{y}), \\ \lambda_i &\geq 0, \\ \text{tr}(\mathbf{G}_i \mathbf{G}_i^H) &\leq P_i, \\ \lambda_i [\text{tr}(\mathbf{G}_i \mathbf{G}_i^H) - P] &= 0. \end{aligned} \quad (23)$$

Proof. According to the result in [22], the Lagrangian corresponding to problem (22) is given by

$$L(\lambda_i, \mathbf{G}_i) = -I(\mathbf{b}; \mathbf{y}) + \sum_{i=1}^K \lambda_i [\text{tr}(\mathbf{G}_i \mathbf{G}_i^H) - P_i], \quad (24)$$

where λ_i is the Lagrangian dual variable corresponding to the i th user's power constraint. By making the gradient of (24) with respect to \mathbf{G}_i equal to zero, the first equation in (23) is achieved. Afterwards, by adding the power constraint and nonnegative Lagrangian dual variable constraint, the KKT conditions shown in (23) are obtained. \square

Depending on the KKT conditions, the line search method shown in [22] can be applied to optimize the codebook matrix. It should be noted that mutual information shown in (18) contains rather complex integrals and it is difficult to achieve its closed-form expression. In Section 4, the calculation of mutual information is achieved by Monte Carlo simulations and the iterative codebook optimization algorithm is proposed.

In addition, it can be seen that when calculating the gradient with respect to \mathbf{G}_i in (21), the expression of \mathbf{E}_b is required. The details of deriving the expression of \mathbf{E}_b are given in Appendix A. It can be seen that \mathbf{E}_b also contains very complex integrals and its value is obtained by Monte Carlo simulations.

4. Iterative Codebook Optimization Algorithm

In Section 3, the KKT conditions do not give explicit method to find the optimal codebook matrix. In this section, inspired by the line search method in [22], the iterative codebook optimization algorithm is proposed, where the codebook optimization is implemented by searching the suitable update step size along the direction of the gradient.

In the first subsection, the line search applied in the iterative codebook optimization algorithm is described. Afterwards, the steps of the proposed algorithm are elaborated. Because the mutual information and mean squared error do not have closed-form expressions, the optimization is implemented based on their Monte Carlo simulation results.

4.1. Line Search Optimization Method. Based on the line search method in [22], the codebook matrix of each user should be updated along the direction of the gradient. During optimization, the update step size should be optimized to make sure that mutual information after codebook updating is nondecreasing. In this paper, the backtracking line search method [22] is introduced to determine the step size.

There are two nested loops in the proposed algorithm. The outer-loop index denotes the iteration number and the inner loop index denotes the user index from 1 to K .

In the n th outer loop, the expression of $I(\mathbf{b}; \mathbf{y})$ after the $(i-1)$ th user's updating is denoted by

$$\begin{aligned} I^{(n,i-1)}(\mathbf{b}; \mathbf{y}) \\ = f(\widehat{\mathbf{G}}_1^{(n)}, \dots, \widehat{\mathbf{G}}_{i-1}^{(n)}, \mathbf{G}_i^{(n)}, \mathbf{G}_{i+1}^{(n)}, \dots, \mathbf{G}_k^{(n)}), \end{aligned} \quad (25)$$

Input: Randomly select codebook matrix $\mathbf{G}_i^{(1)}$, $1 \leq i \leq K$, $\text{tr}(\mathbf{G}_i \mathbf{G}_i^H) = P_i$

(1) Initialization: $\overline{\mathbf{G}}^{(1,0)} = \text{blkdiag}[\mathbf{G}_1^{(1)}, \dots, \mathbf{G}_k^{(1)}]$

(2) Outer loop: for $n = 1 : 1 : N_{\text{ite}}$

(3) Inner loop: for $i = 1 : 1 : K$

(a) Perform monte-carlo simulations to calculate $I^{(n,i-1)}(\mathbf{b}, \mathbf{y})$ and $\mathbf{E}_b^{(n,i-1)}$

(b) Calculate the gradient $\nabla_{\mathbf{G}_i^{(n)}} I^{(n,i-1)}(\mathbf{b}, \mathbf{y})$ according to (28)

Do

(c) Update $\nabla \mathbf{G}_i^{(n)}$ according to (27)

(d) Calculate $\widehat{\mathbf{G}}_i^{(n)}$ to satisfy the power constraint according to (29)

(e) Perform monte-carlo simulations to calculate $I^{(n,i)}(\mathbf{b}, \mathbf{y})$ according to (30)

(f) Update step size $t = t \times \beta$

While $I^{(n,i)}(\mathbf{b}, \mathbf{y}) < I^{(n,i-1)}(\mathbf{b}, \mathbf{y}) + \alpha t \times \|\nabla_{\mathbf{G}_i^{(n)}} I^{(n,i-1)}(\mathbf{b}, \mathbf{y})\|_F^2$

(g) Generate the updated codebook matrix $\overline{\mathbf{G}}^{(n,i)} = \text{blkdiag}[\widehat{\mathbf{G}}_1^{(n)}, \dots, \widehat{\mathbf{G}}_{i-1}^{(n)}, \widehat{\mathbf{G}}_i^{(n)}, \dots, \mathbf{G}_k^{(n)}]$

(4) End Inner loop

(h) Generate $\overline{\mathbf{G}}^{(n+1,0)} = \overline{\mathbf{G}}^{(n,K)}$

(5) End Outer loop

ALGORITHM 1: Concrete process of iterative codebook optimization algorithm.

where mutual information is considered as the function of codebook matrix of each user and the superscript $(n, i - 1)$ of $I(\mathbf{b}, \mathbf{y})$ denotes the outer loop and inner loop index pair. The codebook matrix corresponding to $I^{(n,i-1)}(\mathbf{b}, \mathbf{y})$ is given by

$$\overline{\mathbf{G}}^{(n,i-1)} = \text{blkdiag}[\widehat{\mathbf{G}}_1^{(n)}, \dots, \widehat{\mathbf{G}}_{i-1}^{(n)}, \mathbf{G}_i^{(n)}, \dots, \mathbf{G}_k^{(n)}], \quad (26)$$

where the matrices from $\widehat{\mathbf{G}}_1^{(n)}$ to $\widehat{\mathbf{G}}_{i-1}^{(n)}$ denote the codebooks that have been updated in the n th outer loop.

In addition, $\overline{\mathbf{G}}^{(n,0)} = \text{blkdiag}[\mathbf{G}_1^{(n)}, \dots, \mathbf{G}_k^{(n)}]$ denotes the initial codebook matrix in the n th iteration; the corresponding mutual information is $I^{(n,0)}(\mathbf{b}, \mathbf{y}) = f(\mathbf{G}_1^{(n)}, \dots, \mathbf{G}_k^{(n)})$.

Based on the gradient expression in (21), the line search result is given by

$$\nabla \mathbf{G}_i^{(n)} = \mathbf{G}_i^{(n)} + t \nabla_{\mathbf{G}_i^{(n)}} I^{(n,i-1)}(\mathbf{b}, \mathbf{y}), \quad (27)$$

where t is the step size and the expression of $\nabla_{\mathbf{G}_i^{(n)}} I^{(n,i-1)}(\mathbf{b}, \mathbf{y})$ is given by

$$\begin{aligned} \nabla_{\mathbf{G}_i} I^{(n,i-1)}(\mathbf{b}, \mathbf{y}) &= \frac{1}{\ln 2 \cdot \sigma_n^2} (\mathbf{e}_{K,i} \otimes \mathbf{I}_{d_v}) \\ &\times \overline{\mathbf{H}}_S^H \overline{\mathbf{H}}_S \overline{\mathbf{G}}^{(n,i-1)} \mathbf{E}_b^{(n,i-1)} \\ &\times (\mathbf{e}_{K,i} \otimes \mathbf{I}_{d_v})^H, \end{aligned} \quad (28)$$

where the mean squared error matrix $\mathbf{E}_b^{(n,i-1)}$ is calculated based on the codebook matrix $\overline{\mathbf{G}}^{(n,i-1)}$.

In addition, the codebook matrix of each user should satisfy the power constraint. Assuming that the maximum transmit power of user i is equal to P_i , the normalized codebook matrix is given by

$$\widehat{\mathbf{G}}_i^{(n)} = \frac{\sqrt{P_i} \times \nabla \mathbf{G}_i^{(n)}}{\|\nabla \mathbf{G}_i^{(n)}\|_F}. \quad (29)$$

Afterwards, the i th $d_v \times d_v$ diagonal block $\mathbf{G}_i^{(n)}$ is replaced by $\widehat{\mathbf{G}}_i^{(n)}$ and the updated mutual information is calculated according to

$$I^{(n,i)}(\mathbf{b}, \mathbf{y}) = f(\widehat{\mathbf{G}}_1^{(n)}, \dots, \widehat{\mathbf{G}}_{i-1}^{(n)}, \widehat{\mathbf{G}}_i^{(n)}, \mathbf{G}_{i+1}^{(n)}, \dots, \mathbf{G}_k^{(n)}). \quad (30)$$

Based on the backtracking line search method [22], the following constraint should be satisfied to make sure that the updated mutual information is nondecreasing:

$$I^{(n,i)}(\mathbf{b}, \mathbf{y}) > I^{(n,i-1)}(\mathbf{b}, \mathbf{y}) + \alpha t \|\nabla_{\mathbf{G}_i^{(n)}} I^{(n,i-1)}(\mathbf{b}, \mathbf{y})\|_F^2, \quad (31)$$

where α is the predetermined parameter and always belongs to the interval $(0, 0.3)$ [22].

If the above constraint is not satisfied, the calculations in (27)–(30) are repeated to update $I^{(n,i)}(\mathbf{b}, \mathbf{y})$ and the “backtracking” is performed with updated size $t = t \times \beta$, where $\beta \in (0, 0.8)$ is the predetermined parameter [22]. Afterwards, the constraint in (31) is retested.

In the next subsection, the detailed steps of the proposed iterative codebook optimization algorithm are given.

4.2. Concrete Steps of Iterative Codebook Optimization Algorithm. From the analysis in Section 3 and Appendix A, it is shown that both $I(\mathbf{b}, \mathbf{y})$ and \mathbf{E}_b contain rather complex integrals and it is difficult to derive their closed-form expressions. Therefore, in the proposed algorithm, the calculation of $I(\mathbf{b}, \mathbf{y})$ and \mathbf{E}_b is realized according to the Monte Carlo simulations, which should cover all the M^K constellation points. It is believed that the computational complexity is proportional to M^K .

According to the above analysis, concrete steps of the proposed algorithm are given in Algorithm 1. The parameter N_{ite} is the number of outer loops.

It should be noted that the performance of backtracking line search method depends on the initial values of the

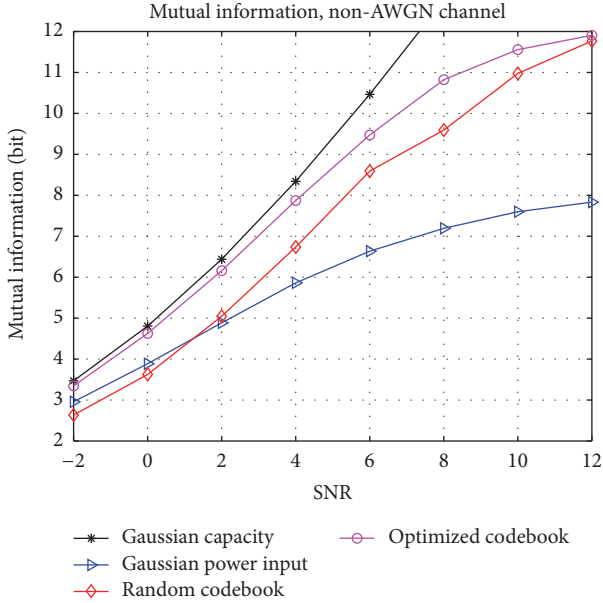


FIGURE 2: Mutual information performance in non-AWGN channel. The SCMA structure is given in Figure 1 with $K = 6$ and $N = 4$. The channel responses are given in Appendix B.

codebook matrices. Therefore, in simulations, the iterative optimization shown in Algorithm 1 should be repeated multiple times with different initial codebook matrices.

In order to evaluate upper bound of the proposed algorithm, Gaussian channel capacity with the same channel coefficient matrix should be calculated. According to [27], under Gaussian input assumption, the iterative water-filling algorithm is able to find the globally optimal power allocation result, which achieves Gaussian capacity bound. This can be seen as the upper bound of the proposed iterative codebook optimization algorithm.

5. Simulation Results

In this section, the simulation results are given. With the factor graph in Figure 1 and $M = 4$, mutual information between the information bit vector \mathbf{b} and received signal \mathbf{y} is bounded by $H(\mathbf{b}) = K \log_2 M = 12$ bit. The codebook matrix of each user should satisfy the power constraint, $\text{tr}(\mathbf{G}_i \mathbf{G}_i^H) \leq P_i$, $1 \leq i \leq K$. In the following simulations, we set $P_i = N/K = 2/3$, $1 \leq i \leq K$. Simulation results in non-AWGN and AWGN channels are given in Sections 5.1 and 5.2, respectively.

5.1. Non-AWGN Channel Simulation Results. In Figure 2, mutual information achieved by the proposed iterative codebook optimization algorithm in non-AWGN channel is shown. The responses of non-AWGN channel are given in Appendix B. In addition, the channel setting makes sure that the channel power satisfies the following constraint:

$$\text{tr}(\mathbf{H}_s \mathbf{H}_s^H) = Nd_f = Kd_v = 12. \quad (32)$$

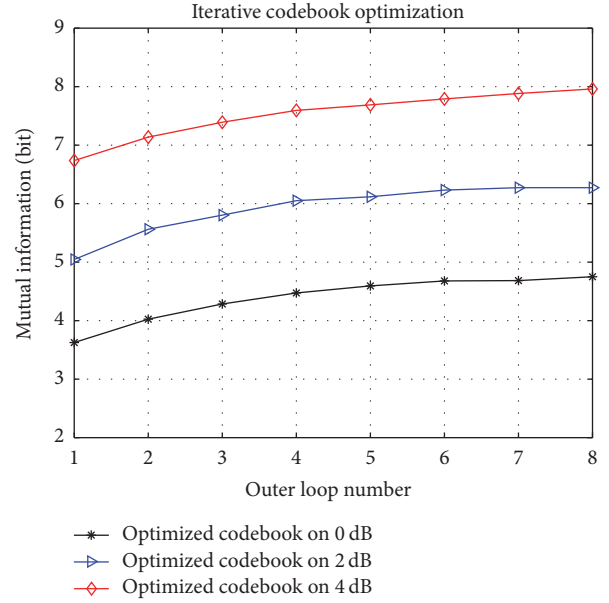


FIGURE 3: Convergence performance of the proposed iterative optimization algorithm in non-AWGN channel. The SNR is set equal to 0 dB, 2 dB, and 4 dB, respectively.

According to the analysis in Section 4, the performance of the proposed iterative codebook optimization algorithm depends on values of the initial codebook matrices. Therefore, the codebook optimization result is chosen from 20 realizations with different initial codebook matrices.

In Figure 2, the result of the proposed iterative codebook optimization algorithm is denoted by “optimized codebook.” The Gaussian capacity bound with the same channel responses according to [27] is denoted by “Gaussian capacity.” In addition, we introduce the scheme called “Gaussian power input.” In this setting, the codebook matrix \mathbf{G}_i , $1 \leq i \leq K$, is squared root of the power distribution matrix obtained from iterative water-filling algorithm in [27]. With above \mathbf{G}_i , mutual information between discrete input \mathbf{b} and continuous output \mathbf{y} is calculated and denoted by “Gaussian power input” in Figure 2. From the analysis in [27], iterative water-filling algorithm also requires channel state information. In addition, the result of random codebook satisfying the power constraint is denoted by “random codebook.” Figure 2 demonstrates that the proposed iterative codebook optimization algorithm can approach Gaussian capacity bound in low and medium SNR regime. Due to the inability to track the channel responses, the performance of “random codebook” is worse than that of “optimized codebook.” When SNR is lower than 1 dB, the result of “Gaussian power input” is better than that of “random codebook.” However, when SNR increases, “Gaussian power input” method fails to approach the performance of “optimized codebook.” This indicates that iterative water-filling algorithm with Gaussian input assumption cannot be directly applied in the discrete input channel even with perfect channel state information.

Furthermore, in Figure 3, the convergence of the proposed iterative codebook optimization algorithm is shown.

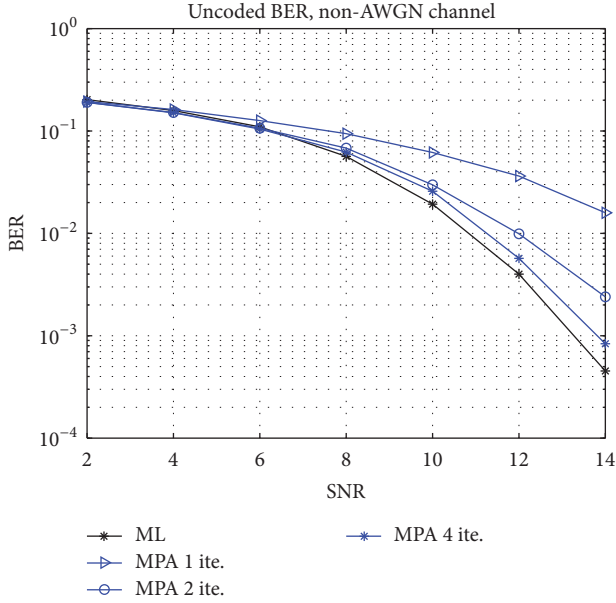


FIGURE 4: Uncoded BER performance of maximum likelihood detection algorithm (ML) and message passing algorithm (MPA) in non-AWGN channel.

The maximum number of outer loops in the proposed algorithm is set as 8. In addition, the initial value of step size parameter t is set as 1. During iterative codebook optimization, the parameter α is set as 0.1 and β is set as 0.5. It can be seen that, after 6 iterations, the increment of mutual information becomes marginal. This means that the proposed algorithm tends to converge after limited outer loops.

In the following, the optimized codebook with mutual information equal to 6 bits is applied. The concrete codebook expressions are given in Appendix B. In Figure 4, the uncoded bit error rate (uncoded BER) results of maximum likelihood algorithm (ML) and message passing algorithm (MPA) are given. It can be seen that MPA can approach the performance of ML detection after 4 iterations. When BER is equal to 10^{-3} , the loss of MPA with 4 iterations is only about 0.6 dB. This indicates that MPA works well with the optimized codebook.

In Figure 5, the coded bit error rate (coded BER) with the optimized codebook matrix is given. Turbo code in LTE [28] is applied and the information bit length is equal to 1024. Because the codebook in Appendix B is optimization result when mutual information is equal to 6 bits, the channel code rate is set as 0.5. The inner iteration number of Turbo decoding is equal to 7. In multiuser detection, the iteration number of MPA is equal to 4. Two channel coding schemes are involved in Figure 5. In scheme 1, each user in SCMA has its own channel coding block. Figure 5 shows that the best user is about 3 dB better than the worst user. In addition, in scheme 1, the average bit error rate is limited by the worst user. In scheme 2, the channel coding across all $K = 6$ users is introduced. According to the statement in [29], coding across channels with different reliabilities can achieve better coded

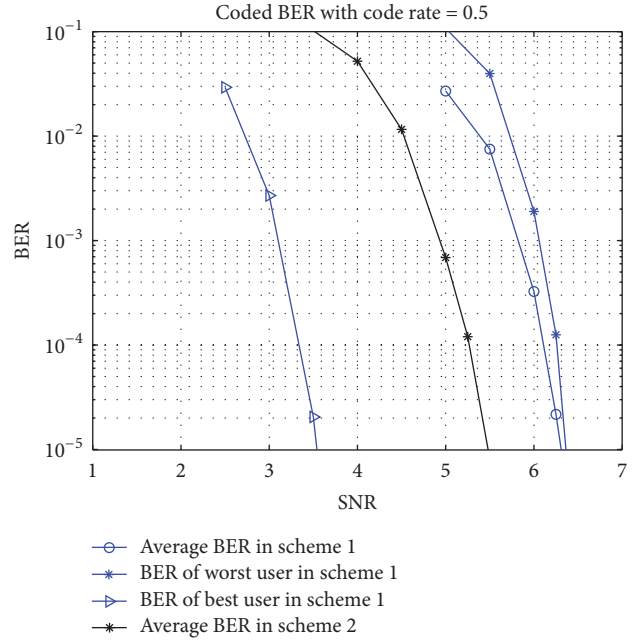


FIGURE 5: Coded BER performance with the optimized codebook matrix in non-AWGN channel. Two channel coding schemes are involved.

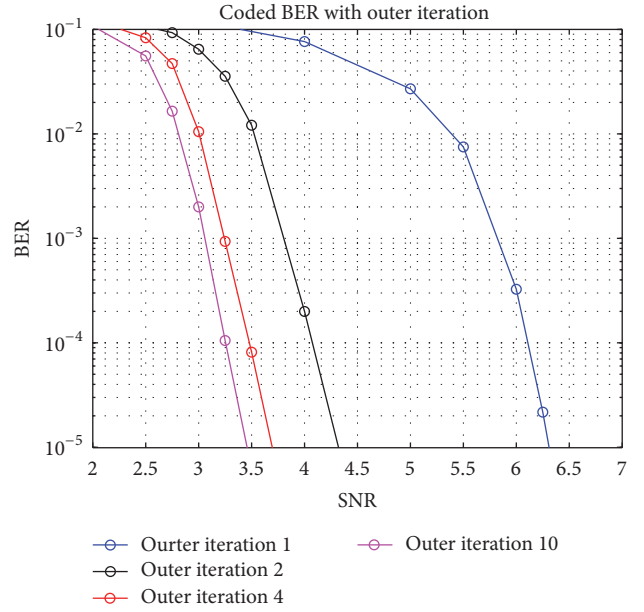


FIGURE 6: Coded BER performance with outer iteration between channel decoding and message passing algorithm (MPA) in non-AWGN channel. The SCMA structure is given in Figure 1 with $K = 6$ and $N = 4$.

BER performance. In Figure 5, it is shown that the average bit error rate of scheme 2 is about 1 dB better than scheme 1.

In addition, the performance of outer-loop iteration between channel decoder and message passing algorithm (MPA) with scheme 1 is given in Figure 6. In scheme 1, each user has its own channel coding block. Similar to that in

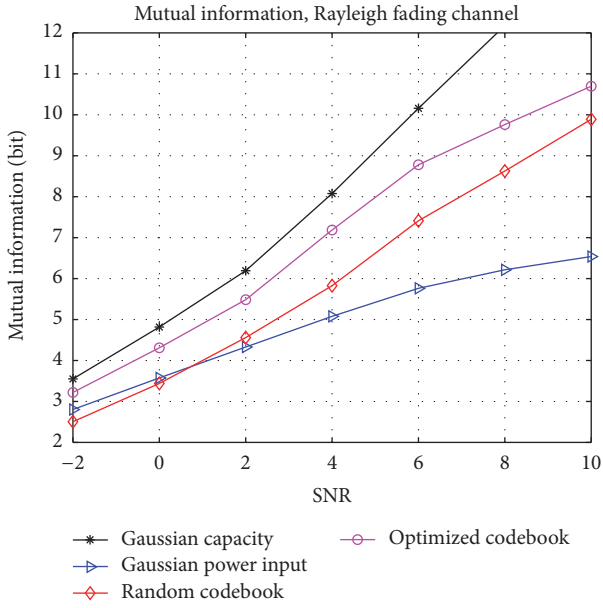


FIGURE 7: Mutual information performance averaging over 1000 Rayleigh fading channels. The SCMA structure is given in Figure 1 with $K = 6$ and $N = 4$.

Figure 5, the information bit length is equal to 1024 and the channel coding rate is equal to 0.5. The inner iteration number of Turbo decoding is equal to 7 and the iteration number of MPA is equal to 4. Because channel decoding feedback provides high-reliability extrinsic information for MPA, the outer-loop iteration can greatly improve the receiver performance. After 10 outer-loop iterations, the performance improvement is about 3 dB when BER is equal to 10^{-5} .

In order to improve the credibility, we further give the simulation results averaging over 1000 Rayleigh fading channels in Figure 7. The curve legends in Figure 7 are the same as that in Figure 2. The simulation results show that the performance of “optimized codebook” is better than that of “random codebook” and “Gaussian power input.” Compared with Gaussian capacity upper bound, the loss of “optimized codebook” is not very large in low and medium SNR regime. When SNR is lower than 0 dB, the result of “Gaussian power input” is better than that of “random codebook.” With the increase of SNR, “Gaussian power input” is unable to approach the performance of “optimized codebook.” The above analysis shows that when averaging over many Rayleigh channels, the proposed optimization algorithm still has better performance.

5.2. AWGN Channel Simulation Results. In this subsection, simulation results in AWGN channel are given. Figure 8 demonstrates mutual information for the factor graph in Figure 1 in AWGN channel. The result of the proposed iterative codebook optimization algorithm is denoted by “optimized codebook.” The Gaussian capacity bound is denoted by “Gaussian capacity.” In addition, the result of the existing codebook proposed by Huawei Corporation in [30] is denoted by “Huawei codebook.” It can be seen

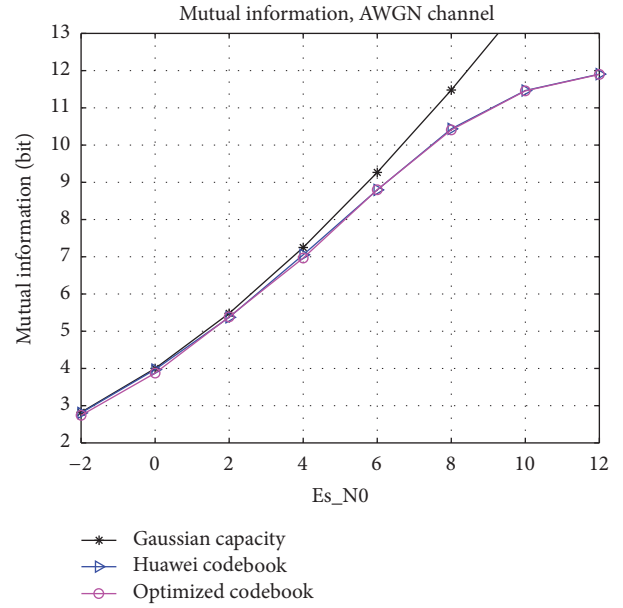


FIGURE 8: Mutual information performance in AWGN channel. The SCMA structure is given in Figure 1 with $K = 6$ and $N = 4$.

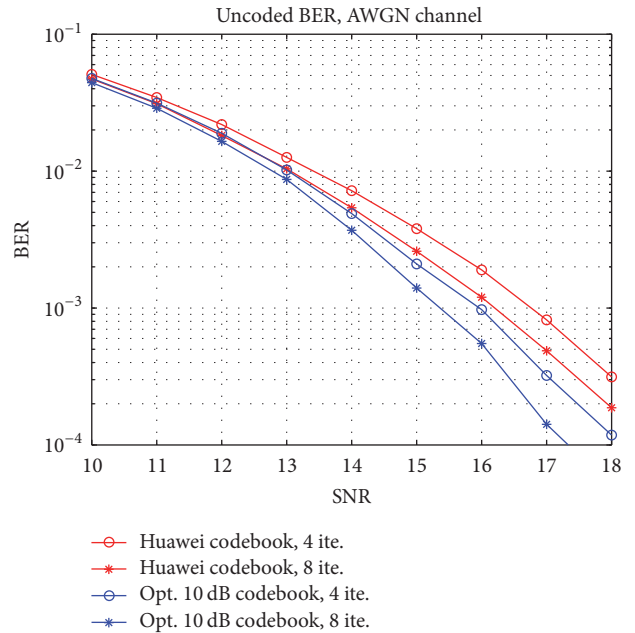


FIGURE 9: Uncoded BER performance of message passing algorithm (MPA) in AWGN channel. The SCMA structure is given in Figure 1 with $K = 6$ and $N = 4$.

that the proposed algorithm can achieve the same mutual information performance as “Huawei codebook.” In low and medium SNR regime, the proposed algorithm can approximate “Gaussian capacity” bound with small performance loss.

Furthermore, the uncoded bit error rate (uncoded BER) of the optimized codebook in AWGN channel is given in Figure 9. The message passing algorithm (MPA) is performed at the receiver. The codebook matrices are optimization

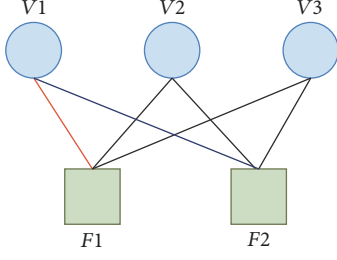


FIGURE 10: Factor graph with $K = 3$ and $N = 2$.

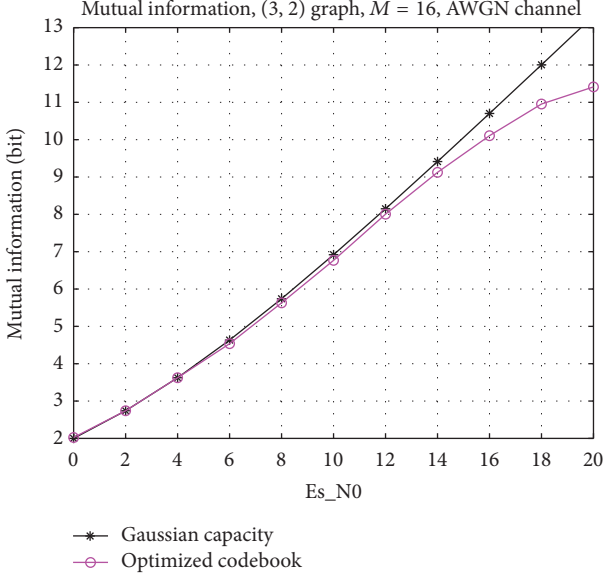


FIGURE 11: Mutual information performance with $M = 16$ in AWGN channel. The SCMA structure is given in Figure 10 with $K = 3$ and $N = 2$.

results of the proposed algorithm when SNR = 10 dB, whose expressions are detailed in Appendix C. Compared with “Huawei codebook” in [30], “optimized codebook” according to the proposed algorithm has better performance. With 8 iterations of MPA, “optimized codebook” has 1 dB performance gain over “Huawei codebook.” For clearness, Huawei codebook in [30] is rewritten according to superposition modulation matrices and its concrete expressions are given in Appendix C.

The above simulations’ results are all based on the factor graph in Figure 1 with $M = 4$. In the following simulation, the codebook design is extended to the case with $M = 16$. Considering the codebook optimization complexity, our focus is on the factor graph with 2 subchannels and 3 users, whose structure is shown in Figure 10.

The proposed column-extended channel model can well describe the codebook optimization problem with $M = 16$. The detailed signal model analysis with $M = 16$ is given in Appendix D. Figure 11 demonstrates the simulation result of (3, 2) factor graph with $M = 16$ in AWGN channel. The optimized codebook can efficiently approach Gaussian capacity upper bound. When SNR is lower than

6 dB, the performance loss between the optimized codebook and upper bound is negligible.

6. Conclusion

In this paper, an efficient SCMA codebook optimization algorithm is proposed according to maximizing mutual information between the discrete input and continuous output. Firstly, SCMA signal model is given based on the superposition modulation structure, which can well represent the relationship between the codebook matrix and received signal. Based on the superposition model, the iterative codebook optimization algorithm is proposed, where the line search method is applied to find locally optimal codebooks. It is shown that the superposition model can be applied in multiuser channel with random channel coefficients. In AWGN channel, the proposed optimization codebook can approach Gaussian capacity upper bound in low and medium SNR regime. In non-AWGN channel, the performance loss compared with upper bound is not very large. In addition, with the optimized codebook, message passing algorithm (MPA) at the receiver exhibits good performance.

Appendix

A. Details of Mean Squared Error

Based on the result in [31], mean squared error matrix denotes the error correlation between the transmit bit vector \mathbf{b} and the detection result $\hat{\mathbf{b}}(\mathbf{y})$. Therefore, we have

$$\mathbf{E}_{\mathbf{b}} = E_{\mathbf{b}, \mathbf{y}} \left[(\mathbf{b} - \hat{\mathbf{b}}(\mathbf{y})) (\mathbf{b} - \hat{\mathbf{b}}(\mathbf{y}))^H \right], \quad (\text{A.1})$$

where $\hat{\mathbf{b}}(\mathbf{y})$ is achieved by calculating the conditional mean of the transmit bit vector based on the received signal \mathbf{y} and it is denoted by

$$\begin{aligned} \hat{\mathbf{b}}(\mathbf{y}) &= \sum_{\bar{m}=1}^{M^K} \mathbf{b}_{\bar{m}} p(\mathbf{b}_{\bar{m}} | \mathbf{y}) \\ &= \frac{\sum_{\bar{m}=1}^{M^K} \mathbf{b}_{\bar{m}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})}{\sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})}. \end{aligned} \quad (\text{A.2})$$

Furthermore, expression (A.1) can be rewritten as follows:

$$\begin{aligned} \mathbf{E}_{\mathbf{b}} &= E_{\mathbf{b}, \mathbf{y}} \left[(\mathbf{b} - \hat{\mathbf{b}}(\mathbf{y})) (\mathbf{b} - \hat{\mathbf{b}}(\mathbf{y}))^H \right] = \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \\ &\quad \cdot \int_{\mathbf{y}} (\mathbf{b}_{\bar{m}} - \hat{\mathbf{b}}(\mathbf{y})) \times (\mathbf{b}_{\bar{m}} - \hat{\mathbf{b}}(\mathbf{y}))^H p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y} \\ &= \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \int_{\mathbf{y}} (\mathbf{b}_{\bar{m}} \mathbf{b}_{\bar{m}}^H - \mathbf{b}_{\bar{m}} \hat{\mathbf{b}}^H(\mathbf{y}) \\ &\quad - \hat{\mathbf{b}}(\mathbf{y}) \mathbf{b}_{\bar{m}}^H + \hat{\mathbf{b}}(\mathbf{y}) \hat{\mathbf{b}}^H(\mathbf{y})) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y}. \end{aligned} \quad (\text{A.3})$$

There are four parts included in the integral of the above expression and the derivation details of each part are given as follows.

For the first part, we have

$$\begin{aligned}
& \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \int_{\mathbf{y}} \mathbf{b}_{\bar{m}} \mathbf{b}_{\bar{m}}^H p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y} \\
&= \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \mathbf{b}_{\bar{m}} \mathbf{b}_{\bar{m}}^H \int_{\mathbf{y}} p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y} \\
&= \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \mathbf{b}_{\bar{m}} \mathbf{b}_{\bar{m}}^H = \mathbf{I}_{K \times d_v}.
\end{aligned} \tag{A.4}$$

In the above expression, the second equation holds because $p(\mathbf{y} | \mathbf{b}_{\bar{m}})$ shown in (14) is Gaussian distributed probability density function with $\int_{\mathbf{y}} p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y} = 1$.

For the second part, we have

$$\begin{aligned}
& \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \int_{\mathbf{y}} \mathbf{b}_{\bar{m}} \hat{\mathbf{b}}^H(\mathbf{y}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y} \\
&= \int_{\mathbf{y}} \sum_{\bar{m}=1}^{M^K} \mathbf{b}_{\bar{m}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) \hat{\mathbf{b}}^H(\mathbf{y}) d\mathbf{y} \\
&= \int_{\mathbf{y}} \frac{\sum_{\bar{m}=1}^{M^K} \mathbf{b}_{\bar{m}} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})}{\sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}})} \\
&\quad \times \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) \times \hat{\mathbf{b}}^H(\mathbf{y}) d\mathbf{y} \\
&= \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \int_{\mathbf{y}} \hat{\mathbf{b}}(\mathbf{y}) \hat{\mathbf{b}}^H(\mathbf{y}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y}.
\end{aligned} \tag{A.5}$$

In the above expression, the third equation is achieved based on the expression of $\hat{\mathbf{b}}(\mathbf{y})$ shown in (A.2).

It can be seen that the third part and the fourth part have the same result as (A.5). Therefore, the mean squared error matrix in (A.3) is rewritten as

$$\begin{aligned}
\mathbf{E}_{\mathbf{b}} &= \mathbf{I}_{K \times d_v} \\
&\quad - \sum_{\bar{m}=1}^{M^K} p(\mathbf{b}_{\bar{m}}) \int_{\mathbf{y}} \hat{\mathbf{b}}(\mathbf{y}) \hat{\mathbf{b}}^H(\mathbf{y}) p(\mathbf{y} | \mathbf{b}_{\bar{m}}) d\mathbf{y}.
\end{aligned} \tag{A.6}$$

With equal probability input assumption, the expression of $\mathbf{E}_{\mathbf{b}}$ can be further denoted by

$$\begin{aligned}
\mathbf{E}_{\mathbf{b}} &= \mathbf{I}_{K \times d_v} - \frac{1}{M^K} \\
&\quad \cdot \sum_{\bar{m}=1}^{M^K} E_{\mathbf{n}} \left[\frac{\left(\sum_{\bar{k}=1}^{M^K} \mathbf{b}_{\bar{k}} u_{\bar{m},\bar{k}} \right) \left(\sum_{\bar{k}=1}^{M^K} \mathbf{b}_{\bar{k}} u_{\bar{m},\bar{k}} \right)^H}{\left(\sum_{\bar{k}=1}^{M^K} u_{\bar{m},\bar{k}} \right)^2} \right],
\end{aligned} \tag{A.7}$$

where the variable $u_{\bar{m},\bar{k}}$ is given by

$$u_{\bar{m},\bar{k}} = \exp \left(- \frac{\| \bar{\mathbf{H}}_S \bar{\mathbf{G}} (\mathbf{b}_{\bar{m}} - \mathbf{b}_{\bar{k}}) + \mathbf{n} \|^2}{\sigma_n^2} \right). \tag{A.8}$$

The above analysis shows that it is difficult to derive the closed-form expression of $\mathbf{E}_{\mathbf{b}}$. During the implementation of iterative codebook optimization algorithm in Section 4, $\mathbf{E}_{\mathbf{b}}$ is achieved from Monte Carlo simulations.

B. Details of Non-AWGN Channel Response and Codebook Expressions

The channel responses applied in non-AWGN scenario are given by

$$\begin{aligned}
h_{1,1} &= 0.4843 - 1.1249i, \\
h_{2,1} &= 0.5868 - 0.3945i, \\
h_{1,2} &= 0.5700 + 0.5846i, \\
h_{3,2} &= 0.9879 - 0.5978i, \\
h_{1,3} &= -0.6148 - 0.6748i, \\
h_{4,3} &= 0.8837 + 0.6211i, \\
h_{2,4} &= -0.1626 + 0.8983i, \\
h_{3,4} &= -1.0336 - 0.3137i, \\
h_{2,5} &= 1.1138 - 0.3047i, \\
h_{4,5} &= 0.7967 - 0.1786i, \\
h_{3,6} &= 0.3878 + 0.5912i, \\
h_{4,6} &= 1.2039 - 0.2250i.
\end{aligned} \tag{B.1}$$

The optimized codebook matrices from \mathbf{G}_1 to \mathbf{G}_6 with mutual information equal to 6 bits are given by

$$\begin{aligned}
\mathbf{G}_1^{(\text{opt})} &= \begin{bmatrix} 0.2570 + 0.5092i & 0.4398 - 0.3495i \\ 0.0385 + 0.1257i & -0.0789 + 0.0480i \end{bmatrix}, \\
\mathbf{G}_2^{(\text{opt})} &= \begin{bmatrix} -0.2314 + 0.0851i & 0.1009 - 0.0359i \\ 0.1741 - 0.4712i & -0.3640 - 0.4568i \end{bmatrix}, \\
\mathbf{G}_3^{(\text{opt})} &= \begin{bmatrix} -0.1860 - 0.0235i & 0.1222 + 0.5095i \\ 0.0277 + 0.5366i & -0.1556 + 0.2099i \end{bmatrix}, \\
\mathbf{G}_4^{(\text{opt})} &= \begin{bmatrix} 0.3295 - 0.1191i & 0.3341 - 0.2843i \\ -0.2860 + 0.3745i & 0.0675 - 0.3533i \end{bmatrix}, \\
\mathbf{G}_5^{(\text{opt})} &= \begin{bmatrix} -0.5955 - 0.1326i & -0.3274 + 0.3694i \\ 0.1337 + 0.0159i & -0.1658 + 0.0717i \end{bmatrix}, \\
\mathbf{G}_6^{(\text{opt})} &= \begin{bmatrix} -0.1840 - 0.0873i & -0.1115 - 0.0126i \\ 0.0684 + 0.4833i & -0.5262 - 0.3122i \end{bmatrix}.
\end{aligned} \tag{B.2}$$

C. Details of AWGN Channel Codebook Expressions

In AWGN channel, the optimized codebook matrices for factor graph in Figure 1 when SNR =10 dB are given by

$$\begin{aligned}
 \mathbf{G}_1^{(10\text{ dB})} &= \begin{bmatrix} -0.4537 - 0.2942i & -0.1114 - 0.2503i \\ 0.2563 + 0.0679i & -0.2026 - 0.4334i \end{bmatrix}, \\
 \mathbf{G}_2^{(10\text{ dB})} &= \begin{bmatrix} -0.4935 + 0.1406i & -0.1302 - 0.1910i \\ 0.0727 - 0.1944i & -0.4370 - 0.3404i \end{bmatrix}, \\
 \mathbf{G}_3^{(10\text{ dB})} &= \begin{bmatrix} 0.1123 - 0.3513i & -0.1685 + 0.4982i \\ -0.4298 + 0.0382i & -0.2537 - 0.0589i \end{bmatrix}, \\
 \mathbf{G}_4^{(10\text{ dB})} &= \begin{bmatrix} -0.1376 + 0.1682i & 0.3903 - 0.3857i \\ -0.4810 + 0.1877i & -0.2117 + 0.0831i \end{bmatrix}, \\
 \mathbf{G}_5^{(10\text{ dB})} &= \begin{bmatrix} -0.0548 - 0.2603i & -0.5657 - 0.0541i \\ 0.3334 + 0.2855i & -0.2423 + 0.1469i \end{bmatrix}, \\
 \mathbf{G}_6^{(10\text{ dB})} &= \begin{bmatrix} -0.0130 + 0.3272i & -0.0089 - 0.3914i \\ 0.0410 - 0.4939i & -0.0347 - 0.3992i \end{bmatrix}.
 \end{aligned} \tag{C.1}$$

In addition, Huawei codebook proposed in [30] can be given by the following superposition modulation matrices:

$$\begin{aligned}
 \mathbf{G}_1^{(\text{HW})} &= \begin{bmatrix} 0.2269 - 0.1648i & 0.4083 - 0.2965i \\ 0.3132 - 0.3958i & -0.1740 + 0.2199i \end{bmatrix}, \\
 \mathbf{G}_2^{(\text{HW})} &= \begin{bmatrix} -0.2804 & -0.5047 \\ -0.4083 - 0.2965i & 0.2269 + 0.1648i \end{bmatrix}, \\
 \mathbf{G}_3^{(\text{HW})} &= \begin{bmatrix} -0.0122 - 0.5045i & 0.0068 + 0.2803i \\ 0.2269 - 0.1648i & 0.4083 - 0.2965i \end{bmatrix}, \\
 \mathbf{G}_4^{(\text{HW})} &= \begin{bmatrix} -0.2804 & -0.5047 \\ 0.3132 - 0.3958i & -0.1740 + 0.2199i \end{bmatrix}, \\
 \mathbf{G}_5^{(\text{HW})} &= \begin{bmatrix} -0.4083 - 0.2965i & 0.2269 + 0.1648i \\ -0.2804 & -0.5047 \end{bmatrix}, \\
 \mathbf{G}_6^{(\text{HW})} &= \begin{bmatrix} -0.2804 & -0.5047 \\ -0.0122 - 0.5045i & 0.0068 + 0.2803i \end{bmatrix}.
 \end{aligned} \tag{C.2}$$

D. Signal Model of Figure 9 with $M = 16$

Based on the factor graph in Figure 9, the mapping matrix between the user nodes and subchannels is given by

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

In AWGN scenario, the channel matrix \mathbf{H}_S is equal to above F. After column extension, the following $\bar{\mathbf{H}}_S$ is achieved:

$$\bar{\mathbf{H}}_S = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{D.2}$$

With $M = 16$, the bit vector of the k th user, $1 \leq k \leq 3$, is given by

$$\mathbf{b}_k = [b_1^{(k)}, b_2^{(k)}, b_3^{(k)}, b_4^{(k)}]^T. \tag{D.3}$$

The corresponding codebook \mathbf{G}_k , $1 \leq k \leq 3$, is a 2×4 matrix. Consequently, the overall block diagonal codebook matrix is given by

$$\bar{\mathbf{G}} = \text{blkdiag} \{ \mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3 \}. \tag{D.4}$$

Based on the above analysis, the received signal is given by

$$\mathbf{y} = \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b} + \mathbf{n}, \tag{D.5}$$

where \mathbf{b} collects all the users' information bits with $\mathbf{b} = [\mathbf{b}_1^T, \mathbf{b}_2^T, \mathbf{b}_3^T]^T$.

In addition, the multiuser access model can be further denoted by

$$\mathbf{y} = \bar{\mathbf{H}}_S \bar{\mathbf{G}} \mathbf{b} + \mathbf{n} = \sum_{i=1}^3 \mathbf{H}_i \mathbf{G}_i \mathbf{b}_i + \mathbf{n}, \tag{D.6}$$

where \mathbf{H}_i is column-extended result of the i th column of \mathbf{H}_S and it is given by

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{D.7}$$

According to the above expression, the proposed iterative codebook optimization algorithm can be implemented.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61601047, 61671080, and 61771066).

References

- [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] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

- [3] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and Multiple Access for 5G Networks," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 629–646, 2018.
- [4] 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.
- [5] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 3, pp. 421–431, 2008.
- [6] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *Proceedings of the 2010 IEEE 71st Vehicular Technology Conference, VTC 2010-Spring*, Taiwan, May 2010.
- [7] R. Razavi, M. Al-Imari, M. A. Imran, R. Hoshyar, and D. Chen, "On receiver design for uplink low density signature OFDM (LDS-OFDM)," *IEEE Transactions on Communications*, vol. 60, no. 11, pp. 3409–3508, 2012.
- [8] L. Wen, R. Razavi, M. A. Imran, and P. Xiao, "Design of Joint Sparse Graph for OFDM System," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1823–1836, 2015.
- [9] M.-C. Chang and Y. T. Su, "Overloaded multiple access systems: A generalized model and a low-complexity multiuser decoder," in *Proceedings of the 9th International Symposium on Turbo Codes and Iterative Information Processing, ISTC 2016*, pp. 231–235, France, September 2016.
- [10] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 332–336, IEEE, London, UK, September 2013.
- [11] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proceedings of the 80th IEEE Vehicular Technology Conference, VTC 2014-Fall*, Canada, September 2014.
- [12] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proceedings of the IEEE International Conference on Communications, ICC 2015*, pp. 2918–2923, UK, June 2015.
- [13] H. Mu, Z. Ma, M. Alhaji, P. Fan, and D. Chen, "A fixed low complexity message pass algorithm detector for Up-Link SCMA system," *IEEE Wireless Communications Letters*, vol. 4, no. 6, pp. 585–588, 2015.
- [14] B. Xiao, K. Xiao, S. Zhang, Z. Chen, B. Xia, and H. Liu, "Iterative detection and decoding for SCMA systems with LDPC codes," in *Proceedings of the International Conference on Wireless Communications and Signal Processing, WCSP 2015*, China, October 2015.
- [15] F. Wei and W. Chen, "Low Complexity Iterative Receiver Design for Sparse Code Multiple Access," *IEEE Transactions on Communications*, vol. 65, no. 2, pp. 621–634, 2017.
- [16] J. Harshan and B. S. Rajan, "On two-user Gaussian multiple access channels with finite input constellations," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 57, no. 3, pp. 1299–1327, 2011.
- [17] M. Cheng, Y. Wu, and Y. Chen, "Capacity analysis for non-orthogonal overloading transmissions under constellation constraints," in *Proceedings of the International Conference on Wireless Communications and Signal Processing, WCSP 2015*, China, October 2015.
- [18] S. Zhang, K. Xiao, B. Xiao et al., "A capacity-based codebook design method for sparse code multiple access systems," in *Proceedings of the 8th International Conference on Wireless Communications and Signal Processing, WCSP 2016*, China, October 2016.
- [19] J. Bao, Z. Ma, G. K. Karagiannidis, M. Xiao, and Z. Zhu, "Joint Multiuser Detection of Multidimensional Constellations over Fading Channels," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 161–172, 2017.
- [20] J. Bao, Z. Ma, Z. Ding, G. K. Karagiannidis, and Z. Zhu, "On the design of multiuser codebooks for uplink SCMA Systems," *IEEE Communications Letters*, vol. 20, no. 10, article no. A42, pp. 1920–1923, 2016.
- [21] X. Ma and L. Ping, "Coded modulation using superimposed binary codes," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 50, no. 12, pp. 3331–3343, 2004.
- [22] L. Ping, J. Tong, X. Yuan, and Q. Guo, "Superposition coded modulation and iterative linear MMSE detection," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 6, pp. 995–1004, 2009.
- [23] C. Xiao, Y. R. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3301–3314, 2011.
- [24] M. Wang, W. Zeng, and C. Xiao, "Linear precoding for MIMO multiple access channels with finite discrete inputs," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 3934–3942, 2011.
- [25] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.
- [26] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 52, no. 1, pp. 141–154, 2006.
- [27] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 50, no. 1, pp. 145–152, 2004.
- [28] Multiplexing and channel coding, Release 8, 2009 3GPP TS 36.212.
- [29] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge, UK, 2005.
- [30] <http://www.innovateasia.com/5g/en/gp2.html>, SCMA Codebooks, (Jun. 2015).
- [31] A. Lozano, A. M. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 52, no. 7, pp. 3033–3051, 2006.