Cluster-Free NOMA Communications Toward Next Generation Multiple Access

Xiaoxia Xu, Yuanwei Liu, Senior Member, IEEE, Xidong Mu, Member, IEEE, Qimei Chen, Member, IEEE, and Zhiguo Ding, Fellow, IEEE

Abstract—A generalized downlink multi-antenna non-orthogonal multiple access (NOMA) transmission framework is proposed with the novel concept of cluster-free successive interference cancellation (SIC). In contrast to conventional NOMA approaches, where SIC is successively carried out within the same cluster, the key idea is that the SIC can be flexibly implemented between any arbitrary users to achieve efficient interference elimination. Based on the proposed framework, a sum rate maximization problem is formulated for jointly optimizing the transmit beamforming and the SIC operations between users, subject to the SIC decoding conditions and users’ minimal data rate requirements. To tackle this highly-coupled mixed-integer nonlinear programming problem, an alternating direction method of multipliers-successive convex approximation (ADMM-SCA) algorithm is developed. The original problem is first reformulated into a tractable biconvex augmented Lagrangian (AL) problem by handling the non-convex terms via SCA. Then, this AL problem is decomposed into two subproblems that are iteratively solved by the ADMM to obtain the stationary solution. Furthermore, to reduce the computational complexity and alleviate the parameter initialization sensitivity of ADMM-SCA, a Matching-SCA algorithm is proposed. The intractable binary SIC operations are solved through an extended many-to-many matching, which is jointly combined with an SCA process to optimize the transmit beamforming. The proposed Matching-SCA can converge to an enhanced exchange-stable matching that guarantees the local optimality. Numerical results demonstrate that: i) the proposed Matching-SCA algorithm achieves comparable performance and a faster convergence compared to ADMM-SCA; ii) the proposed generalized framework realizes scenario-adaptive communications and outperforms traditional multi-antenna NOMA approaches in various communication regimes.

Index Terms—Next-generation multiple access (NGMA), non-orthogonal multiple access (NOMA), multiple antennas, successive interference cancellation (SIC).

I. INTRODUCTION

WIRELESS communications are currently undergoing an unprecedented revolution. It is predicted by Cisco that the number of wireless-enabled devices will increase to more than 40 billion by 2023 [1]. Furthermore, the types of future wireless-enabled devices will vary from smart phones to connected cars, wearables, sensors, collaborative robots, and so on. Due to the explosive demands of wireless traffics and the emergence of various innovative wireless applications, next-generation wireless network, also referred to as the sixth generation (6G), is evolving towards a new era of the Internet of Everything (IoE) [2]. Driven by this exciting vision, 6G is expected to embrace broadband-hungry transmissions, pervasive access, and extremely massive connectivity in diverse and heterogeneous communication scenarios [3]. To meet these challenges, the realization of 6G requires a fully integrated and a seamless convergence of different multiple access technologies, namely next generation multiple access (NGMA) [4]. As a promising multiple access technology, power-domain non-orthogonal multiple access (NOMA)1 [5], [6] has become an indispensable component of NGMA. By exploiting the signal superposition at transmitters and the successive interference cancellation (SIC) at receivers, NOMA enables users served by the same time/frequency/code resource block to be further multiplexed and distinguished in the power domain. Hence, it can dramatically enhance the network capacity and user connections, as well as reducing the outage probability [7].

On the road from NOMA to NGMA, the integration of NOMA and multiple-antenna multiple-output-multiples-input (MIMO) technologies has been regarded as one key aspect [8]. On the one hand, multiple-antenna technologies can enable spatial-domain multi access (SDMA) and provide additional spatial degrees of freedom (DoFs) to assist NOMA communications. On the other hand, NOMA opens up new dimensions and opportunities for resource reuse, which is capable of

1For the sake of expression, we refer to the power-domain NOMA as NOMA in this paper.
increasing the affordable traffic loadings of multiple-antenna communication systems [9]. Therefore, multi-antenna NOMA provides a promising way to significantly improve spectral efficiency and connection density for next-generation wireless systems [8], [10].

A. Prior Works

In the past few years, extensive literatures have been devoted to the development of multi-antenna NOMA systems. Existing multi-antenna NOMA systems can be loosely classified into two categories, namely beamformer-based NOMA and cluster-based NOMA, which differ in the strategies of both multi-antenna beamforming and SIC operation designs [11].

1) Studies on Beamformer-Based NOMA: Beamformer-based NOMA [12], [13], [14], [15] directly serves different users via distinct beamforming vectors, whose beamforming strategy is similar to conventional multiple-antenna communication systems. Meanwhile, by carrying out SIC between the multiplexed users, the spatial interference that cannot be effectively mitigated by beamforming can be further suppressed leveraging NOMA. Based on a minimization-maximization algorithm, the authors of [12] optimized the beamformer to maximize the sum rate for a multi-user downlink multiple-input single-output NOMA (MISO-NOMA) system. Simulation results signified that beamformer-based NOMA outperforms the traditional multiple-antenna communication systems in the severely overloaded systems, where the transmit antenna number is much larger than the user number. Additionally, the authors of [13] investigated the optimal power allocation in a two-user downlink MIMO-NOMA system, which can achieve the capacity region of the MIMO broadcast channel under the derived channel state information (CSI) condition. The authors of [14] derived the condition of quasi-degraded channels, based on which a low-complexity precoding scheme was proposed for multi-user MISO-NOMA transmissions to approach the rate region of the dirty paper coding. By considering both perfect and imperfect CSI cases, the authors of [15] further proposed low-complexity beamforming and user selection schemes to improve the sum rate and the outage probability of beamformer-based NOMA systems.

2) Studies on Cluster-Based NOMA: Different from the beamformer-based NOMA, cluster-based NOMA [10], [16], [17], [18] typically divides the highly channel correlated users into the same cluster, where each cluster shares the same beamforming vector. While the inter-cluster interference is mitigated via beamforming, the intra-cluster interference is suppressed by carrying out SIC within each cluster [10]. In [16], the authors analysed the performance of the cluster-based NOMA system, which analytically demonstrated the superiority of cluster-based NOMA over the MIMO and orthogonal multiple-access (MIMO-OMA) system in terms of both sum channel capacity and ergodic sum capacity. The authors of [17] investigated an uplink millimeter-wave (mmWave) massive MIMO-NOMA system with hybrid analog-digital beamforming, where user clustering was obtained by considering both users’ channel correlations and gain difference, and the power allocation is designed to maximize the energy efficiency. In [18], the authors proposed a two-stage cluster grouping algorithm for an angle-domain mmWave MIMO-NOMA system, and investigated the max-min power control to enhance user fairness. Considering multi-cell MISO-NOMA system, the authors of [19] proposed a distributed user grouping, beamforming and power control algorithm for power consumption minimization. Furthermore, the author of [20] investigated two different NOMA beamforming schemes, where the NOMA user shares the spatial beam with legacy SDMA users or exploits a dedicated beam. The optimal solution for both schemes are analyzed, and the studies showed that sharing spatial beam can significantly reduce the computational complexity at the expense of a slight performance loss.

B. Motivations and Contribution

Note that SIC plays an important role in NOMA and the design of SIC operations between users is crucial for the eventual performance achieved by NOMA. As discussed above, current multi-antenna NOMA approaches [8], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] generally assume that the SIC is sequentially carried out within the same cluster, namely cluster-specific SIC, thus leading to both benefits and drawbacks. To be more specific, on the one hand, beamformer-based NOMA assigns all users to a single cluster, which is shown to be capable of achieving the same performance as the dirty paper coding scheme in some specific scenarios [14]. However, given the sequential nature of cluster-specific SIC, users in higher SIC decoding orders have to implement a large number of SIC operations before decoding their own signals, thus leading to a high system complexity. Moreover, beamformer-based NOMA also encounters the SIC overuse issue, especially when users’ channels are low-correlated [4]. This is because the SIC decoding conditions can impose undesired spatial interference to low channel-correlated users, even if this interference could have been eliminated via the spatial multiplexing. On the other hand, cluster-based NOMA partially alleviates the SIC overuse issue by dividing users into different clusters, where the inter-cluster and intra-cluster interference can be mitigated via spatially separated beamforming and SIC, respectively. Therefore, it can support a large number of users with a moderate SIC complexity. However, cluster-based NOMA relies on the assumption that the users in the same cluster have high channel correlations while the users of different clusters experiencing low channel correlations, which may not always hold due to the randomness of wireless channels.

It can be observed that both beamformer-based NOMA and cluster-based NOMA are scenario-centric, whose effectiveness relies on specific scenarios, and thus cannot deal with heterogeneous scenarios for next-generation wireless networks. Against this background and to pave the way to NGMA, this paper proposes a novel generalized downlink multi-antenna NOMA transmission framework with the concept of cluster-free SIC. It enables SIC to be flexibly implemented over any arbitrary channel-correlated users to achieve efficient interference elimination, thus breaking the constraints of the existing.
cluster-specific multi-antenna NOMA approaches. Mathematically, it provides a generalized modelling, which not only unifies the existing approaches but also provides more flexible transmission options, thus overcoming the shortcoming of existing approaches. This enables a paradigm of scenario-adaptive multi-antenna NOMA for NGMA. The contributions of this paper can be summarized as follows.

- We propose a novel generalized downlink multi-antenna NOMA transmission framework with the concept of cluster-free SIC, which enables flexible SIC operations between users to facilitate efficient interference elimination. The proposed framework can overcome shortcomings of traditional methods and empower a scenario-adaptive multi-antenna NOMA paradigm. We formulate a sum rate maximization problem for jointly optimizing the transmit beamforming and the SIC operations subject to SIC decoding conditions and users’ data rate constraints.
- We develop an alternating direction method of multipliers-successive convex approximation (ADMM-SCA) algorithm to tackle the formulated mixed-integer nonlinear programming (MINLP) problem, which is highly coupled and non-convex. The original problem is first reformulated into a tractable augmented Lagrangian (AL) problem, where the non-convex terms are handled by invoking the SCA method. The obtained biconvex AL problem is then decomposed into two convex subproblems, which are iteratively solved by ADMM to obtain a stationary solution.
- We propose a Matching-SCA algorithm to further reduce the computational complexity and overcome the parameter initialization sensitivity of ADMM-SCA. The SIC operations between users are modelled as a two-sided many-to-many matching with externality. Then, we extend the conventional swap-based matching to efficiently solve the SIC operation problem, while employing the SCA to jointly optimize the corresponding transmit beamforming. The proposed Matching-SCA can converge to an enhanced exchange-stable matching, which guarantees the local optimality.
- Numerical results demonstrate that the proposed Matching-SCA algorithm results in comparable performance and a faster convergence compared to the ADMM-SCA algorithm, especially in the overloaded regime. It is also shown that the proposed generalized multi-antenna NOMA framework is capable of achieving efficient SIC operations and scenario-adaptive communications, which outperforms traditional transmission schemes regardless of system loadings and users’ channel correlations.

C. Organization and Notation

The rest of this paper is organized as follows. Section II presents the generalized downlink multi-antenna NOMA transmission framework and formulates the sum rate maximization problem. In Section III, an ADMM-SCA algorithm is developed for solving the formulated problem. Furthermore, a low-complexity and fast-convergent Matching-SCA algorithm is proposed in Section IV by extending the conventional many-to-many matching theory. Section V presents numerical results to demonstrate efficiencies of the proposed algorithms, and Section VI finally concludes the paper.

Notation: Vectors and matrices are denoted by bold-face letters. \( \Re(x) \) represents the real part of a complex variable \( x \). \( \| x \| \) denotes the Euclidean norm of a vector \( x \). \( x^T \) and
\( x^H \) denote the transpose and Hermitian conjugate of vector \( x \), respectively. \( I_{N \times N} \) indicates an identity matrix of size \( N \). \( I_{M \times N} \) denotes an \( M \times N \) all-ones matrix.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Generalized Cluster-Free Multi-Antenna NOMA Framework

We consider a downlink multi-antenna NOMA system, as shown in Fig. 1. There exists an \( M \)-antenna base station (BS) serving \( K \) single-antenna users randomly distributed within its coverage, indexed by \( K = \{1, 2, \ldots, K\} \). Define \( h_k \in \mathbb{C}^{M \times 1} = L^{-1/2}(d_k)h_k \) as the channel vector from BS to user \( k \), where \( L_k(d_k) \) denotes the large-scale fading determined by the distance \( d_k \) between BS and user \( k \), and \( h_k \) indicates the small-scale fading. In contrast to cluster-based NOMA which assigns a single beamforming vector for each cluster, the proposed cluster-free framework assigns a dedicated transmit beamforming vector \( w_k \in \mathbb{C}^{N \times 1} \) for each user \( k \). Denote the transmit beamforming matrix by \( W = [w_1, w_2, \ldots, w_K] \in \mathbb{C}^{N \times K} \). For each user \( k \in K \), the received signal can be expressed as

\[
y_k(W) = h_k^H w_k s_k + \sum_{u \neq k} h_k^H w_u s_u + z_k, \tag{1}
\]

where \( s_k \) denotes the data signal of user \( k \) with normalized power, i.e., \( \mathbb{E}\{s_k s_k^H\} = 1 \). Moreover, \( z_k \sim \mathcal{CN}(0, \sigma^2) \) denotes the additive white Gaussian noise (AWGN), which can be modelled as the circularly symmetric i.i.d zero-mean complex Gaussian variables.

To efficiently mitigate the inter-user interference suffered by each user, the proposed framework introduces a novel cluster-free SIC concept, which differs from conventional methods in that it allows SIC to be flexibly implemented between any two channel-correlated users without the predefined user clusters. Mathematically, we define the binary indicator \( \alpha_{ik} \), \( \forall i, k \in K \), which specifies whether the SIC operation is carried out at user \( i \) to decode the signal of user \( k \). Specifically, \( \alpha_{ik} = 1 \) indicates that user \( i \) will first employ the SIC to decode the signal of user \( k \) before decoding its own signal for eliminating interference from user \( k \), and \( \alpha_{ik} = 0 \) otherwise. As it is generally impossible to mutually implement the SIC decoding at both users, we have

\[
\alpha_{ik} + \alpha_{ki} \leq 1, \quad \forall i, k \in K, \quad i \neq k. \tag{2}
\]

B. SIC Decoding Order

Once the SIC operation \( \alpha \) is given, the SIC decoding order at each user \( i \) can be also determined. For the sake of illustration, we start by considering a three-user network with \( K = \{i, u, k\} \), which can be further generalized to the multi-user network. Considering \( \alpha_{ik} + \alpha_{ki} \leq 1 \), \( \forall i \neq k \), in constraint (2), if user \( i \) would perform SIC decoding to at least one user, i.e., \( \exists k, k \neq i, \alpha_{ik} = 1 \), then the possible configurations of SIC operations can be summarized into the following three cases, where the decoding order at user \( i \) is determined accordingly.

(i) Cluster-free case: As shown in Fig. 2(a), if user \( i \) would perform SIC decoding to both users \( u \) and \( k \), but no SIC operation would be performed between \( u \) and \( k \), i.e., \( \alpha_{ik} = \alpha_{iu} = 1 \) and \( \alpha_{uk} = \alpha_{ku} = 0 \), then users \( i, k, \) and \( u \) would not act as a NOMA cluster. In this case, since the decoding order of \( u \) and \( k \) at user \( i \) cannot be directly determined by \( \alpha \), we would further decide it based on the commonly adopted ascending-channel-gain ordering, which has been demonstrated to be a rational and effective predefined decoding order \([14], [21]\). Specifically, if \(|h_u|^2 \leq |h_k|^2\), the decoding order at \( i \) would be \( u \rightarrow k \rightarrow i \). Otherwise if \(|h_k|^2 \leq |h_u|^2\), the decoding order at \( i \) would be \( k \rightarrow u \rightarrow i \).

(ii) Three-user cluster case: As shown in Fig. 2(b), if user \( i \) would perform SIC decoding to both users \( u \) and \( k \), and there exists either \( \alpha_{ku} = 1 \) or \( \alpha_{uk} = 1 \), then users \( i, u, \) and \( k \) form a typical three-user SIC cluster. Following the principle of conventional NOMA methods, if \( \alpha_{ik} = \alpha_{iu} = \alpha_{ku} = 1 \), the decoding order at user \( i \) would be \( u \rightarrow k \rightarrow i \). Similarly, if \( \alpha_{ik} = \alpha_{ui} = \alpha_{uk} = 1 \), the decoding order at user \( i \) would be \( k \rightarrow u \rightarrow i \).

(iii) Two-user cluster case: As shown in Fig. 2(c), if user \( i \) would perform SIC decoding to only one user \( k \) while taking other users’ interference as noise, then user \( i \) and user \( k \) form a tow-user cluster. Thus, the SIC decoding order at user \( i \) can be directly obtained as \( k \rightarrow i \).

Given \( \alpha \) and the corresponding decoding order, user \( i \) will sequentially decode the signals of each user \( k \) that satisfies \( \alpha_{ik} = 1 \). Once user \( k \)'s signal is decoded, user \( i \) can remove the interference from user \( k \) when decoding the remaining users’ signals. When \( \alpha_{ik} = 1 \), to successfully implement SIC for interference elimination, the following SIC decoding constraint should be satisfied \([11], [16]\)

\[
R_{i \rightarrow k}(\alpha, W) \geq \alpha_{ik} R_{k \rightarrow i}(\alpha, W), \quad \forall i, k \in K, \quad i \neq k, \tag{3}
\]

where \( R_{i \rightarrow k}(\alpha, W) \) is the achievable rate for user \( i \) to decode user \( k \)'s signal, and \( R_{k \rightarrow i}(\alpha, W) \) denotes the achievable rate for user \( k \) to decode its own signal. The SIC decoding constraint (3) indicates that the achievable rate \( R_{i \rightarrow k} \) for decoding user \( k \)'s signal at user \( i \) should be maintained at a sufficiently high level to successfully eliminate the interfering signal from user \( k \).

C. Rate Modelling

The achievable rate of the proposed framework can be modelled as follows.

1) Communication Rate Modelling: When user \( k \) decodes its own signal, the observed interference \( \text{Interf}_{k \rightarrow k} \) after SIC operations can be expressed as

\[
\text{Interf}_{k \rightarrow k}(\alpha, W) = \sum_{u \neq k} (1 - \alpha_{ku}) |h_k^H w_u|^2 + \sigma^2, \quad \forall k \in K, \tag{4}
\]

where \( \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]^T \) being the SIC operation vector for user \( k \). Therefore, the signal-to-interference-plus-noise
ratio (SINR) SINR\textsubscript{i→k} for user k to decode its own signal can be given by

\[
\text{SINR}_{i \rightarrow k} (\alpha, W) = \frac{|h_i^H w_k|^2}{\text{Intf}_{i \rightarrow k} (\alpha, W)}, \quad \forall k \in K. \tag{5}
\]

As a result, the achievable data rate \(R_{i \rightarrow k} (\alpha, W)\) for user k to decode its own signal can be computed as

\[
R_{i \rightarrow k} (\alpha, W) = \log_2 (1 + \text{SINR}_{i \rightarrow k} (\alpha, W)), \quad \forall k \in K.
\]

2) SIC Decoding Rate Modelling: For ease of expression, we sort users in ascending order of their channel gains, i.e., \(\|h_i\|^2 \leq \|h_k\|^2, \forall i < k\). Define Intf\textsubscript{i→k} (\(\alpha, W\)), \(\forall i, k \in K\), \(i \neq k\), as the observed interference when user k decoding the signal of user i. From the analysis of Section II-B, we can derive that for \(u < k\), given \(\alpha_{ik} = 1\), if \(\alpha_{iu} = 0\) or \(\alpha_{iu} = \alpha_{uk} = 1\), then the interfering signal from user u would not be eliminated when user i decoding user k’s signal. Similarly, for \(u > k\), given \(\alpha_{ik} = 1\), if \(\alpha_{iu} = 0\) or \(\alpha_{ku} = 0\), then the interfering signal from user u would not be eliminated when user i decoding user k’s signal. Therefore, the interference Intf\textsubscript{i→k} (\(\alpha, W\)) for decoding the signal of user k at user i can be mathematically formulated by

\[
\text{Intf}_{i \rightarrow k} (\alpha, W) = \sum_{u < k} (1 - \alpha_{iu} + \alpha_{i\alpha_{uk}}) |h_i^H w_u|^2 + \sum_{u > k} (1 - \alpha_{iu} \alpha_{ku}) |h_i^H w_u|^2 + \sigma^2, \quad \forall i, k \in K, \ i \neq k. \tag{6}
\]

The corresponding SINR for user i to decode user k’s signal, defined as SINR\textsubscript{i→k} (\(\alpha, W\)), can be computed by

\[
\text{SINR}_{i \rightarrow k} (\alpha, W) = \frac{|h_i^H w_k|^2}{\text{Intf}_{i \rightarrow k} (\alpha, W)}, \quad \forall i, k \in K, \ i \neq k, \tag{7}
\]

and the achievable data rate can be given by \(R_{i \rightarrow k} (\alpha, W) = (1 + \text{SINR}_{i \rightarrow k} (\alpha, W)), \forall i, k \in K, \ i \neq k\).

The sum rate of the proposed generalized cluster-free NOMA framework can be given by

\[
R (\alpha, W) = \sum_{k \in K} R_{i \rightarrow k} (\alpha, W). \tag{8}
\]

D. Generalization Analysis

Essentially, by introducing the cluster-free SIC, the proposed framework provides a generalized and unified modelling, where the beamformer-based NOMA, cluster-based NOMA, and SDMA can be all regarded as the special cases of the proposed framework, as analysed as follows.

1) Special Case 1 - Beamformer-Based NOMA: When there is only one SIC decoding sequence that involves all the connected users, i.e., \(\alpha_{ik} = 1, \forall i > k\), and \(\alpha_{ik} = 0\) otherwise, the proposed generalized framework is equivalent to beamformer-based NOMA and the achievable sum rate can be given by \(R_{\text{BBNOMA}} (\alpha, W) = \sum_{k \in K} \log_2 \left( 1 + \frac{|h_i^H w_k|^2}{\sum_{u \in K \setminus \{k\}} |h_i^H w_u|^2 + \sigma^2} \right) \). \]

2) Special Case 2 - Cluster-Based NOMA: When user i and user k are served by two aligned beamforming vectors, i.e., \(\exists c_{ik} \in \mathbb{R}, w_k = c_{ik} w_{ij}, i, k \in K\), they are considered to share the same spatial beam, which is similar to the traditional cluster-based NOMA systems. If SIC decoding is sequentially carried out only between users served by aligned beamforming vectors, then the proposed generalized framework reduces to cluster-based NOMA, where \(\alpha_{ik} = 1\) if \(i > k\) and \(\exists c_{ik} \in \mathbb{R}\) such that \(w_k = c_{ik} w_{ij}\), and \(\alpha_{ik} = 0\) otherwise. Suppose there exists \(G\) clusters, indexed by \(g = \{1, 2, \ldots, G\}\). Denote the user set of cluster \(g\) by \(K_g\). Then, the achievable sum rate of the cluster-based NOMA system can be written as \(R_{\text{CBNOMA}} (\alpha, W) = \sum_{k \in K_g} \log_2 \left( 1 + \frac{|h_i^H w_k|^2}{\sum_{u \in K \setminus K_g} |h_i^H w_u|^2 + \sum_{u \in K_g} |h_i^H w_u|^2 + \sigma^2} \right) \). \]

3) Special Case 3 - SDMA: If there is no SIC operation between any users, i.e., \(\alpha_{ik} = 0, \forall i, k \in K, i \neq k\), then...
the proposed generalised NOMA framework is equivalent to SDMA. The sum rate of the SDMA system can be expressed as \( R_{\text{SDMA}}(\alpha, W) = \sum_{k \in K} \log_2 \left( 1 + \frac{|h_{ik} w_k|^2}{\sum_{u \in K} |h_{ik} w_u|^2 + \sigma^2} \right) \).

In addition to unifying the traditional methods, the proposed framework also enables more flexible SIC operations. A specific example is shown in Fig. 1, where users cannot be ideally divided into a single or multiple user clusters, and the cluster-specific SIC schemes are not flexible enough. To empower efficient interference elimination, the proposed cluster-free scheme breaks the clustering limitations, which enables flexible SIC operations between highly channel-correlated users (e.g., user 2 and user 4, user 1 and user 4), while adaptively preventing ineffective SIC operations between less channel-correlated users (e.g., user 1 and user 2).

Remark 1: The proposed framework provides a generalized model to unify traditional methods, and enables more flexible transmission configurations with cluster-free SIC to achieve adaptive inter-user interference mitigation. Therefore, it can overcome the defects of traditional methods and reap their gains to deal with diverse scenarios facing next-generation wireless communications. Owing to these merits, we can straightforwardly derive that the achievable sum rate of the proposed framework can outperform or is at least not worse than that of conventional approaches, i.e.,

\[ R^*(\alpha^*, W^*) \geq \max \{ R_{\text{BBNOMA}}^*, R_{\text{CBNOMA}}^*, R_{\text{SDMA}}^* \}, \]

where \( R^* \) denotes the sum rate achieved at the optimal \( \alpha^* \) and \( W^* \).

E. Problem Formulation

Our goal is to maximize the sum rate while guaranteeing SIC decoding conditions and ensuring users' data rate requirements by jointly optimizing the transmit beamforming and the cluster-free SIC operations between users. Mathematically, the optimization problem can be formulated as

\[ \mathcal{P}_0 : \max_{\alpha, W} \sum_{k \in K} \log_2 \left( 1 + \text{SINR}_{k \rightarrow k}(\alpha, W) \right) \]

s.t. \( R_{i \rightarrow k}(\alpha, W) \geq \alpha_{ik} R_{k \rightarrow k}(\alpha, W), \forall i, k \in K, \]

\[ \log_2 \left( 1 + \text{SINR}_{k \rightarrow k}(\alpha, W) \right) \geq R^\text{min}, \forall k \in K, \]

\[ \sum_{k \in K} \| w_k \|^2 \leq P^\text{max}, \]

\[ \alpha_{ik} + \alpha_{ki} \leq 1, \forall i, k \in K, i \neq k, \]

\[ \alpha_{ik} \in \{0, 1\}, \forall i, k \in K, \]

where constraint (10b) represents the SIC decoding conditions rearranged from (3), (10c) guarantees the minimum data rate of each user \( i \), and (10d) ensures the maximum transmit power of BS does not exceed \( P^\text{max} \). Furthermore, (10e) indicates that user \( i \) and user \( k, i \neq k \), cannot mutually implement the SIC decoding, and (10f) indicates the binary variable constraint.

However, it’s challenging to solve \( \mathcal{P}_0 \) owing to the following reasons. Firstly, the design of SIC operations introduces the binary constraint (10f). Additionally, the optimization variables are highly coupled with each others in both SINR terms and the objective function. Furthermore, the objective and constraints in (10a)–(10c) are neither convex nor concave with respect to optimization variables. Therefore, \( \mathcal{P}_0 \) is a non-convex and highly coupled MINLP problem that is non-deterministic polynomial (NP)-hard. Owing to the coupling mixed-integer variables and the intractable non-convex SIC decoding constraint, it is generally difficult and impractical to search for its globally optimal solution. In the following sections, we aim to propose efficient algorithms that can obtain high-quality locally optimal solutions. Here, a solution \( x \) is locally optimal, implying that at point \( x \) there is no feasible ascent direction for problem maximization.

III. ADMM-SCA Based Solution

In this section, an ADMM-SCA algorithm is developed to solve \( \mathcal{P}_0 \). The highly coupled MINLP is first equivalently reformulated into a tractable AL problem with continuous variables. By invoking SCA to handle the non-convex terms, the AL problem is approximately transformed into a series of biconvex optimization problem. Based on the strongly convergence-guaranteed ADMM method, we further decompose the biconvex problem into two convex subproblems that can be iteratively solved to achieve the stationary solution.

A. Problem Transformation

To tackle the NP-hard MINLP problem (10), we first introduce auxiliary variables \( \beta \in \mathbb{R}^K \times K \), which satisfy

\[ \beta_{ik} = 1 - \alpha_{ik}, \forall i, k \in K, \]

\[ \alpha_{ik} \beta_{ik} = 0, \forall i, k \in K, \]

\[ 0 \leq \alpha_{ik}, \beta_{ik} \leq 1, \forall i, k \in K. \]

Since constraints (11)-(13) enforces \( \alpha_{ik} (1 - \alpha_{ik}) = 0 \) and \( \beta_{ik} (1 - \beta_{ik}) = 0 \), they can stringently guarantee that \( \alpha_{ik}, \beta_{ik} \in \{0, 1\} \). Therefore, the discrete binary constraint (10f) can be equivalently replaced by (11)-(13).

Moreover, since the interference term \( \text{Int}_{i \rightarrow k}(\alpha, W) \) in (6) suffers from the highly coupling variables \( \alpha_{ik} \), \( \alpha_{iu} \), and \( \alpha_{uk} \), we equivalently transform the interference term as follows to make it tractable. Since both \( \{ \alpha_{ik} \} \) and \( \{ \beta_{ik} \} \) are binary variables, the coupling terms \( (1 - \alpha_{iu} + \alpha_{iu} \alpha_{uk}) \) and \( (1 - \alpha_{iu} \alpha_{ku}) \), \forall i, u, k \in K \), in (6) can be directly recast as

\[ 1 - \alpha_{iu} + \alpha_{iu} \alpha_{uk} = \max \{ 1 - \alpha_{iu}, \alpha_{uk} \}, \]

\[ 1 - \alpha_{iu} \alpha_{ku} = \max \{ 1 - \alpha_{iu}, 1 - \alpha_{ku} \}. \]

Hence, the interference terms \( \text{Int}_{i \rightarrow k}(\alpha, W), \forall k \in K, \) in (4) and (6) can be equivalently rewritten into (14), as shown at the bottom of the next page.
Lemma 1: The function \( \text{Intf}_{i-k}(\beta, W) \) defined by (14) is convex with respect to \( \beta \).

Proof: For \( i = k \), \( \text{Intf}_{i-k}(\beta, W) \) is a linear function of \( \beta \). Therefore, we only need to verify the convexity for \( i \neq k \). According to the derivation in [22], the pointwise maximum function \( g(\beta) = \max\{g_1(\beta), g_2(\beta)\} \) is convex if \( g_1(\beta) \) and \( g_2(\beta) \) are both convex functions. Since (14) and (14) are both pointwise maximums of affine functions of \( \beta \), it can be concluded that \( \text{Intf}_{i-k}(\beta, W) \) is convex with respect to \( \beta \), which completes the proof.

To deal with the non-convex data rate expression, we further introduce a series of auxiliary variables \( S = \{S_{ik}\}_{i,k \in K}, I = \{I_{ik}\}_{i,k \in K}, r = \{r_{ik}\}_{i,k \in K} \). Specifically, \( I_{ik} \) indicates the upper bound of the interference \( \text{Intf}_{i-k}(\beta, W) \), \( \forall i, k \in K \).

Moreover, \( S_{ik} \) and \( r_{ik} \) signify the lower bounds of the effective gain and the achievable rate for decoding user \( k \)'s signal at user \( i, \forall i, k \in K \), respectively. Therefore, the intractable MINLP problem (10) can be written as the following continuous problem:

\[
P_1: \max_{\alpha, \beta, W, S, I, r} \sum_{k \in K} r_{ik}
\]  
\[
s.t. \quad r_{ik} \leq \log_2 \left( 1 + \frac{S_{ik}}{I_{ik}} \right), \quad \forall i, k \in K,
\]  
\[
S_{ik} \leq |h_i^k w_k|^2, \quad \forall i, k \in K,
\]  
\[
\text{Intf}_{i-k}(\beta, W) \leq I_{ik}, \quad \forall i, k \in K,
\]  
\[
r_{ik} \geq \alpha_{ik} r_{kk}, \quad \forall i, k \in K, i \neq k,
\]  
\[
r_{kk} \geq R_{\text{min}}, \quad \forall k \in K,
\]  
\[
\alpha_{ik} + \alpha_{ki} \leq 1, \quad \forall i, k \in K, i \neq k,
\]  
\[
\sum_{k \in K} \|w_k\|^2 \leq R_{\text{max}},
\]  
\[
(11) - (13).
\]  

Proposition 1: Problems \( P_1 \) and \( P_0 \) are equivalent in the sense that they have equivalent optimal solutions.

Proof: Owing to the monotonicity of the \( \log_2(\cdot) \) function, the constraints (15b)-(15d) in problem \( P_1 \) always hold with equality at the optimum point. Therefore, the solutions obtained by solving problem \( P_1 \) can satisfy \( r_{ik}^* = \log_2 \left( 1 + \frac{|h_i^k w_k|^2}{\text{Intf}_{i-k}(\beta^*, W^*)} \right) = \log_2 \left( 1 + \frac{\text{SINR}_{i-k}(\alpha^*, W^*)}{\text{Intf}_{i-k}(\beta^*, W^*)} \right), \forall i, k \in K \), which demonstrates the equivalence between the optimal solutions of \( P_1 \) and \( P_0 \).

To tackle the resulting problem \( P_1 \), we invoke the strongly convergence-guaranteed ADMM framework, which is capable of efficiently handling the coupled equality constraints (11) and (12) based on both dual decomposition and method of multipliers. By dualizing and penalizing the coupling equality constraints (11) and (12) into the objective function, the AL problem of \( P_1 \) can be formulated as [24]

\[
P_{\text{AL}}: \max_{\alpha, \beta, W, S, I, r} f_0(r) - \mathcal{L}^{(1)}(\alpha, \beta, \lambda) - \mathcal{L}^{(2)}(\alpha, \beta, \tilde{\lambda})
\]  
\[
s.t. \quad (15b) - (15h), (13),
\]  

where \( f_0(r) = \sum_{k \in K} \log_2(1 + r_{kk}) \) is the original objective function. \( \mathcal{L}^{(1)}(\alpha, \beta, \lambda) \) and \( \mathcal{L}^{(2)}(\alpha, \beta, \tilde{\lambda}) \) respectively denote the AL terms corresponding to equality constraints (11) and (12), given by

\[
\mathcal{L}^{(1)}(\alpha, \beta, \lambda) = \frac{1}{2\rho} \|\beta + \alpha - 1_{K \times K} + \rho \lambda\|^2,
\]  
\[
\mathcal{L}^{(2)}(\alpha, \beta, \tilde{\lambda}) = \frac{1}{2\rho} \sum_{k \in K} \sum_{i \in K} (\alpha_{ik} \beta_{ik} + \rho \tilde{\lambda}_{ik})^2,
\]  

where \( \lambda = \{\lambda_{ik}\} \) and \( \tilde{\lambda} = \{\tilde{\lambda}_{ik}\} \) are the dual variables and \( \rho \) is the non-negative penalty parameter. As proven in [24], by alternatively optimizing the primal variables \( \{\alpha, \beta, W, S, I, r\} \) and dual variables \( \{\lambda, \tilde{\lambda}\} \) of the AL problem, the residuals of equality constraints (11) and (12), i.e., \( (\beta_{ik} + \alpha_{ik} - 1) \) and \( \alpha_{ik} \beta_{ik}, \forall i, k \), will converge to zeros and the binary constraint can be satisfied.

B. ADMM-SCA Algorithm

According to Lemma 1, the AL problem \( P_{\text{AL}} \) is convex over \( \beta \). However, constraint (15b) is still non-convex since \( \log_2 \left( 1 + \frac{S_{ik}}{I_{ik}} \right) = \log_2(S_{ik} + I_{ik}) - \log_2(I_{ik}) \) is a difference-of-concave (d.c.) function over \( I_{ik} \) and constraint (15c) is non-convex over \( W \). Furthermore, since \( \alpha_{ik} \) and \( r_{kk} \) are coupled in constraint (15e), they usually require alternative optimization and cannot be updated simultaneously. This can degrade the achievable performance in practical experiments. To handle the above non-convex and undesirable coupling constraints, we integrate the SCA method [25] into the ADMM framework [23]. By taking the first-order Taylor approximation for a given local point, SCA can approximately and sequentially linearize the non-convex components into a series of convex expressions. In this way, the AL problem can be decomposed into tractable convex subproblems, which can be further optimized based on ADMM iteratively.

The details on SCA transformations can be stated as follows. Specifically, we define \( w_k, I_{ik}, r_{ik} \), and \( \tau_{kk} \) as the values of the optimization variables \( w_k, I_{ik}, \alpha_{ik} \), and \( r_{kk} \) obtained from the previous SCA iteration, respectively. First, to handle constraint (15b), we take the first-order Taylor expansion of

\[
\text{Intf}_{i-k}(\beta, W) = \begin{cases} 
\sum_{u \neq k} \beta_{iu} |h_i^u w_u|^2 + \sigma^2, & \text{if } i = k, \\
\sum_{u < k} \max \{\beta_{iu} - 1, \beta_{ku} - 1\} |h_i^u w_u|^2 + \sum_{u < k} \max \{\beta_{iu}, \beta_{ku}\} |h_i^u w_u|^2 + \sigma^2, & \text{if } i \neq k.
\end{cases}
\]
function $q_2(I_{ik}) = \log_2(I_{ik})$ at point $\bar{I}_{ik}$ and obtain
\[
q_2(I_{ik}) \leq \bar{q}_2(I_{ik}, \bar{I}_{ik}) = \log_2(\bar{I}_{ik}) + \frac{1}{\ln 2} \frac{1}{\bar{I}_{ik}} (I_{ik} - \bar{I}_{ik}).
\] (19)

After rearrangement, constraint (15b) can be transferred into
\[
r_{ik} + \log_2(\bar{I}_{ik}) + \frac{1}{\ln 2} \frac{1}{\bar{I}_{ik}} (I_{ik} - \bar{I}_{ik}) \leq \log_2(I_{ik} + S_{ik}).
\] (20)

Next, we define function $q_1(W_k) = \|h_k^H w_k\|^2$. Based on the first-order Taylor approximation around $\bar{w}_k$, i.e., $q_1(w_k) \geq \tilde{q}_1(W_k, \bar{w}_k) = q_1(\bar{w}_k) + q'_1(\bar{w}_k)(w_k - \bar{w}_k)$, we can reformulate the constraint (15e) as
\[
S_{ik} + \|h_k^H w_k\|^2 \leq 2\rho \|h_k^H h_k^H w_k\|, \quad \forall i, k \in \mathcal{K}.
\] (21)

Furthermore, to decouple $\alpha_{ik}$ and $r_{ik}$ in constraint (15e), the term $\alpha_{ik}r_{kk}$ can be rearranged as $\alpha_{ik}r_{kk} = \frac{1}{4} (\alpha_{ik} + r_{kk})^2 - \frac{1}{4} (\alpha_{ik} - r_{kk})^2$. To deal with this difference of convex expression, we linearize the non-convex term $\frac{1}{4} (\alpha_{ik} - r_{kk})^2$ using the first-order Taylor expansion, i.e.,
\[
-\frac{1}{4} (\alpha_{ik} - r_{kk})^2 \leq -\frac{1}{4} (\bar{\alpha}_{ik} - r_{kk})^2 - \frac{1}{2} (\bar{\alpha}_{ik} - r_{kk}) \times (\alpha_{ik} - \bar{\alpha}_{ik} + r_{kk} - r_{kk}).
\] (22)

Considering (22), constraint (15e) can be transformed into
\[
r_{ik} + \frac{1}{4} (\alpha_{ik} - r_{kk})^2 + \frac{1}{2} (\bar{\alpha}_{ik} - r_{kk}) (\alpha_{ik} - \bar{\alpha}_{ik} + r_{kk} - r_{kk}) \geq \frac{1}{4} (\alpha_{ik} + r_{kk})^2, \quad \forall i, k \in \mathcal{K}, \ i \neq k.
\] (23)

Based on the above analyses, the AL problem $\mathcal{P}_{AL}$ can be approximately transformed into the following problem during each SCA update:
\[
\mathcal{P}_2 : \max_{\alpha, \beta, W, I, S, r} f_0(r) - \mathcal{L}^{(1)}(\alpha, \beta, \lambda) - \mathcal{L}^{(2)}(\alpha, \beta, \bar{\lambda})
\]
\[\text{s.t.} \ (13), (15d), (15f) - (15h), (21), (20), (23).\] (24a)

(24b)

The resulting problem $\mathcal{P}_2$ is a biconvex optimization problem, i.e., the optimization variables of $\mathcal{P}_2$ can be partitioned into two blocks $\{\alpha_{ik}\}_{i \neq k}$, $W$ and $\{\beta_{ik}\}_{i \neq k}$, such that when fixing one block, $\mathcal{P}_2$ is convex over the other block. This biconvex problem can be decomposed into two nested convex subproblems over the variable blocks $\{\alpha_{ik}\}_{i \neq k}$ and $\{\beta_{ik}\}_{i \neq k}$, respectively. Based on ADMM framework, these convex subproblems can be solved alternatively at each iteration, followed by which the dual variables $\lambda$ and $\bar{\lambda}$ are updated. In light of this, we propose an ADMM-SCA algorithm, which has three steps during each iteration.

Firstly, given $\{\beta, \bar{\sigma}, W, I, r, \lambda, \bar{\lambda}\}$, ADMM-SCA algorithm jointly optimizes SIC operations $\{\alpha_{ik}\}_{i \neq k}$ and transmit beamforming $W$ by solving the following convex program
\[
\max_{\alpha, W, I, S, r} f_0(r) - \mathcal{L}^{(1)}(\alpha, \beta, \lambda) - \mathcal{L}^{(2)}(\alpha, \beta, \bar{\lambda})
\]
\[\text{s.t.} \ (13), (15d), (15f) - (15h), (21), (20), (23).\] (25a)

Based on the first-order optimality, the Karush-Kuhn-Tucker (KKT) solution of $\{\alpha_{ik}\}_{i \neq k}$ is given by
\[
\alpha_{ik}^* = \frac{1}{2} (1 + \beta_{ik}) + \rho \omega_{ik}^{(4)} \times \left[ \frac{\rho \omega_{ik}^{(4)} (\bar{\pi}_{ik} - r_{kk} - r_{ik})}{2 (1 + \beta_{ik}) + \rho \omega_{ik}^{(4)}} + 2 \rho (\omega_{ik}^{(2)} - \omega_{ik}^{(1)}) \right],
\] (26)

where $\omega_{ik}^{(1)}$ and $\omega_{ik}^{(2)}$ denotes the Lagrangian multipliers for constraints (13), and $\omega_{ik}^{(3)}$ and $\omega_{ik}^{(4)}$ are the Lagrangian multipliers corresponding to constraints (15g) and (23), respectively.

Thereafter, we update $\beta$ by solving the following problem with fixed $\{\alpha, W, S, \bar{\sigma}, I, r, \lambda, \bar{\lambda}\}$
\[
\max_{\beta, r} f_0(r) - \mathcal{L}^{(1)}(\alpha, \beta, \lambda) - \mathcal{L}^{(2)}(\alpha, \beta, \bar{\lambda})
\]
\[\text{s.t.} \ (13), (15d), (15f), (20), (23).\] (27a)

Since (27) is a convex optimization problem, it can be easily solved by the interior point method using the standard convex optimization tool, such as CVX [26].

Furthermore, at each iteration $t$ the dual variables $\lambda$ and $\bar{\lambda}$ can be updated by
\[
\lambda^{(t+1)} = \lambda^{(t)} + \frac{1}{\rho} \left( \beta^{(t)} + \alpha^{(t)} - 1 \right),
\] (28)
\[
\bar{\lambda}^{(t+1)} = \bar{\lambda}^{(t)} + \frac{1}{\rho} \beta^{(t)} \alpha^{(t)}, \quad \forall i, k \in \mathcal{K}.
\] (29)

The overall ADMM-SCA algorithm can be summarized as Algorithm 1. The computational complexity of solving convex subproblems [(25) and (27)] via the interior point method can be respectively given by $O\left( (4K^2 + M)3.5 \right)$ and $O\left( (3K^2)3.5 \right)$ [22]. Therefore, the computational complexity of Algorithm 1 can be given by $O\left( T_{ADMM} \left( (4K^2 + M)3.5 + (3K^2)3.5 \right) \right)$, where $T_{ADMM}$ denotes the number of iterations for reaching convergence. Although the ADMM-SCA algorithm can converge to a feasible and stationary solution of problem $\mathcal{P}_1$ with polynomial time complexity according to the analyses in [27] and [28], it generally suffers from slow convergence and requires a high computational complexity. Moreover, the obtained discrete variables $\alpha$ and $\beta$ are usually highly sensitive to the initialized parameters, which thus significantly impact the resulting performances. Hence, we randomly initialize $W$ and test $N^{\text{ini}}$ groups of initialized parameters for $\{\beta, \lambda, \bar{\lambda}\}$ to empirically choose appropriate initialization points.

IV. LOW-COMPLEXITY MATCHING-SCA SOLUTION

Although the ADMM-SCA algorithm guarantees the convergence to a desirable suboptimal solution, it may need a large number of iterations for convergence and require high computational complexity when user number increases. Moreover, the achieved performance is typically highly sensitive to the initialized parameters due to the discrete optimization. Motivated by this, in this section we further propose a novel...
The SIC operations disjoint entries, where dynamic two-sided matching game among the multiplexed control for the MINLP optimization. [30], which can achieve a stable and self-organizing discrete promising tool to tackle the combinatorial optimization [29], and inexact SCA. In particular, matching game provides a convex NP-hard MINLP based on the matching game theory and the optimal value.

Algorithm 1 ADMM-SCA Algorithm for Solving $P_1$

1: Initialize the accuracy tolerance $\epsilon_{ADMM} > 0$ and the maximum iteration number $T_{ADMM}^\text{max}$.
2: Initialize $\{\alpha, \beta, W, I, r, \lambda, \bar{\lambda}\}$ with a feasible point, and initialize $\rho > 0$.
3: Set the iteration number as $t = 0$.
4: repeat
5:   By fixing $\{\beta, \bar{r}, W, I, r, \lambda, \bar{\lambda}\}$, update the SIC operations $\alpha$ and the transmit beamforming $W$ by solving problem (25).
6:   By fixing $\{\alpha, W, S, \bar{r}, I, r, \lambda, \bar{\lambda}\}$, update the variables $\beta$ by solving problem (27).
7:   Update dual variables $\lambda$ and $\bar{\lambda}$ using (28) and (29), respectively.
8: until $t = T_{ADMM}^\text{max}$ or the difference of successive objective values satisfies $|f_0(r^t) - f_0(r^{t-1})|^2 \leq \epsilon_{ADMM}$.

Ensure: The SIC operations $\alpha^*$, transmit beamforming $W^*$, and the optimal value.

A. Many-To-Many SIC Matching Problem

Firstly, we model the cluster-free SIC optimization as a dynamic two-sided matching game among the multiplexed users. We define two virtual user sets $U$ and $V$ with logically disjoint entries, where $U$ consists of the users that SIC can cancel the interference imposed by users from $V$. Without loss of generality, we define $U = V = K$. If user $u \in U$ is scheduled to carry out SIC to eliminate interference from user $v \in V$, we say user $u$ and user $v$ are matched to each other, which is denoted by $(u, v)$. The SIC operations can be formulated as a matching problem, which yields the following definitions and remarks.

Definition 1 (Many-to-Many Matching): A man-to-many matching $\mu$ is a function from set $U \cup V$ to the set of all subsets of $U \cup V$, such that

(i) $\mu(u) \subseteq V$ and $|\mu(u)| \leq N_u$, $\forall u \in U$;
(ii) $\mu(v) \subseteq U$ and $|\mu(v)| \leq N_v$, $\forall v \in V$;
(iii) $\mu(u) \subseteq V$ if and only if $\mu(v) \subseteq U$;
(iv) $u \in \mu(v)$ if and only if $v \in \mu(u)$.

In the above definition, condition (i) means that each user $u \in U$ can carry out SIC to a subset $\mu(u)$ of users in $V$, and the cardinality of $\mu(u)$ cannot exceed $N_u$. Condition (ii) indicates that the interference from each user $v \in V$ can be eliminated with SIC by at most $N_v$ users from $U$. Condition (iii) represents that the mapping $\mu(u)$ of user $u \in U$ is the subset of $V$, and vice versa. Condition (iv) implies that when $u \in U$ matches with $v \in V$, $v$ matches with $u$ as well. Without loss of generality, we set $N_u = N_v = K$ here. Thus, both the low-complexity and efficient strategy, which solves the non-convex NP-hard MINLP based on the matching game theory and inexact SCA. In particular, matching game provides a promising tool to tackle the combinatorial optimization [29], [30], which can achieve a stable and self-organizing discrete control for the MINLP optimization.

Lemma 2: The formulated two-sided many-to-many matching has the properties of externality and non-substitutability.

Proof: The properties can be demonstrated as follows.

i) Externality: Owing to the feature of the multiple-antenna NOMA system, the interference suffered by each user $k \in U \cup V$ varies with matching states of the other users. Therefore, the achievable data rate and preference of each user $k$ also depend on other users, and each user should take into account the internal relationship of the other users when determining its matching state. This renders the externality of the SIC matching problem.

ii) Non-substitutability: Given two virtual user sets $U$ and $V$, each user $u \in U$ prefers to match with a subset $V_u$ of $V$, which is defined as the choice of $u$ in $V$, denoted by $C_u(V) = V_u$. Here, user $u$ prefers $V_u$ to any subset of $V$, which can be denoted as $V_u \supseteq V'$, $\forall V' \subseteq V$, $V' \neq V_u$. The preference of $u$ over sets of $V$ possesses substitutability property if and only if $u \in C_u(V)$ and $V_u \in C_u(V \setminus \{v\}')$, $\forall v, v' \in V$, which means that the choice of $u$ in $V$ will not be affected even if one matched user in $V$ is excluded. However, since the optimal beamforming and inter-user interference varies with the SIC operations, the achievable data rate may change under different user matching. Thus, the formulated matching based SIC operation problem lacks the substitutability property.

To address the externality and ensure exchange stability, we first introduce the following matching swap operation as defined by conventional matching theory [29], [30]

$$
\mu_{uv}' = \mu \setminus \{(u, v), (u', v')\} \cup \{(u', v), (v, u)'\},
$$

(31)

which means that users $u$ and $u'$ exchange their matched users $v$ and $v'$ while keeping all other users’ matching states unchanged. Here, we also consider the swap operation over “holes”, i.e., the empty set $\emptyset$ that does not contain any users. Hence, the matching state of user pair $(u, v)$ can be transferred from matched into unmatched after the swap operation $\mu_{uv}'$. Moreover, the state of the user pair $(u, v)$ can shift from unmatched to matched based on $\mu_{uv}'$. With the matching swap operation, the swap-blocking pair can be defined as follows.
Definition 2 (Swap-Blocking Pair): For two users $u$ and $u'$, $(u, u')$ is a swap-blocking pair in matching $\mu$ if and only if
(i) $\forall k \in \{u, u', v, v'\}, U_k \left( \tilde{\mu}_{uv}^{u'v'} \right) \geq U_k(\mu)$;
(ii) $\exists k \in \{u, u', v, v'\}$, such that $U_k \left( \mu_{uv}^{u'v'} \right) > U_k(\mu)$.

B. Extended Many-To-Many Matching for SIC Operations

1) Enhanced Swap Operation: The conventional swap operation in (31) only optimizes the user matching, but cannot dynamically swap the SIC decoding order for the matched user pairs, i.e., swap between $\alpha_{uv} = 1$ and $\alpha_{vu} = 1$ for $(u, v)$. However, in the proposed framework, the SIC operations $\alpha$ determines both the user matching and the decoding order, as discussed in Section II-B. Hence, to efficiently optimize $\alpha$, we extended the conventional many-to-many matching model by considering both matching swap and decoding order swap.

Unlike the conventional matching model that regards the matched pairs $(u, v)$ and $(v, u)$ as the same pair, in our formulated matching model, they will respectively indicate $\alpha_{uv} = \alpha_{vu} = 1$, and thus have different physical meanings. Hence, the swap from $(u, v)$ to $(v, u)$ signifies the swap of SIC decoding order. To enable the exchange of decoding order between two matched users $u$ and $v$, we propose the following decoding order swap operation $\tilde{\mu}_{uv}^{u'v'}$:

$$\tilde{\mu}_{uv}^{u'v'} = \{u \setminus \{(u, v)\} \cup \{(u', v')\} \}, \quad u' = v, \quad v' = u. \quad (32)$$

2) Enhanced Swap Rule: Combining the conventional matching swap operation (31) and the decoding order swap operation (32), we further present the following concept of enhanced swap-blocking pair to determine the swap rule. Different from conventional swap rule that ensures the utility increments of individual users (see Definition 2), we aim at improving the sum utility $U(\mu)$ of all users in matching $\mu$ to maximize the sum rate. Using this enhanced swap rule, the proposed algorithm can converge to a local optimum, as proven in Theorem 1 in Section IV-D.

Definition 3 (Enhanced Swap-Blocking Pair): Given a matching $\mu$, for user $u \in U$ and user $u' \in U$, we define $(u, u')$ as an enhanced swap-blocking pair, if one of the following conditions can be satisfied
(i) $\exists v \in \mu(u), v' \in \mu(u')$ such that $U(\mu_{uv}^{u'v'}) > U(\mu)$ and $\alpha_{v'u'} = \alpha_{uv} = 0$;
(ii) for $u' = v$, $v' = u$, we can obtain that $v \in \mu(u)$ and $U(\tilde{\mu}_{uv}^{u'v'}) > U(\mu)$.

The above condition (i) indicates that after a conventional matching swap (31), the overall utility should be increased. Moreover, to ensure that the matching swap from $\{u, v\} \cup \{(u', v')\}$ to $\{u, v\} \cup \{(u, v')\}$ is feasible, we should always guarantee $\alpha_{uv} + \alpha_{v'u} \leq 1$ and $\alpha_{vu} + \alpha_{u'v} \leq 1$, which leads to the constraint $\alpha_{v'u'} = \alpha_{uv} = 0$ in condition (i). Condition (ii) implies that after a decoding order swap operation (32), the overall utility should be improved.

Note that Definition 3 determines the swap rule of the formulated matching. To be more specific, if there exists an enhanced swap-blocking pair satisfying any of the above conditions (i) and (ii), then the matching is not stable nor convergent, and the corresponding matching $\mu_{uv}^{u'v'}$ or $\tilde{\mu}_{uv}^{u'v'}$ would be “approved”. Therefore, by extending the concept of exchange stability in conventional matching [31], we can define the enhanced exchange stability as follows.

Definition 4 (Enhanced Exchange-Stable Matching): The two-sided matching $\mu$ is an enhanced exchange-stable matching if and only if there is no enhanced swap-blocking pair.

C. Matching-SCA Based Joint Optimization

Based on the proposed extended matching, we further develop a dual-loop iterative algorithm to jointly optimize the SIC operations and the transmit beamforming. In the outer loop, the matching state $\mu$ of SIC operations are updated by the extended many-to-many matching while fixing the transmit beamforming. In the inner loop, given the SIC operations, the transmit beamforming $\mathbf{W}$ is sequentially optimized via an SCA process.

The inner-loop transmit beamforming optimization can be illustrated as follows. Given the current matching state $\mu$ and the corresponding SIC operation variables $\alpha^\mu$, the beamforming $\mathbf{W}$ can be optimized by invoking the SCA method, as analyzed in Section III-B. By fixing $\alpha^\mu$, $\mathbf{W}$ can be optimized by sequentially solving the following convex problem

$$\max_{\mathbf{W} \in S(\mathbf{I})} J_0(\mathbf{r}) \quad (33a)$$

s.t. $r_{ik} + \log_2 (I_{ik}) + \frac{1}{\ln 2} \left( \frac{I_{ik}}{I_{ik} - 1} - 1 \right) \leq \log_2 (I_{ik} + S_{ik}), \quad \forall i, k \in \mathcal{K}, \quad (33b)$

$$r_{ik} \geq \alpha_{ik}^\mu r_{kk}, \quad \forall i, k \in \mathcal{K}, \quad i \neq k, \quad (33c)$$

$$S_{ik} + \left| \mathbf{h}_i^H \tilde{\mathbf{w}}_k \right|^2 \leq 2 \mathbf{R} \left( \tilde{\mathbf{w}}_k^H \mathbf{H}_i \tilde{\mathbf{w}}_k \right), \quad \forall i, k \in \mathcal{K}, \quad (33d)$$

$$\tilde{f}_{i \rightarrow k}(1 - \alpha^\mu, \mathbf{W}) \leq I_{ik}, \quad \forall i, k \in \mathcal{K}, \quad (33e)$$

$$f_{i \rightarrow k}(\alpha^\mu, \mathbf{W}) \leq 1, \quad \forall i, k \in \mathcal{K}. \quad (33f)$$

The developed joint optimization algorithm, namely Matching-SCA, can be summarized as Algorithm 2. Firstly, the initialized matching states of all users is set as unmatched, i.e., $\alpha = \mathbf{I}_{\mathcal{K} \times \mathcal{K}}$. Moreover, the transmit beamforming $\mathbf{W}$ is randomly initialized with a feasible point. The iterative procedure exploits a dual-loop structure. Specifically, given the current matching states $\mu$, the transmit beamforming $\mathbf{W}$ is sequentially optimized via SCA in the inner loop to an inexact solution that is not required to be locally converged (step 7-step 12). Thereafter, by fixing the transmit beamforming $\mathbf{W}$, we further perform an enhanced swap-matching process in the outer loop (step 13-step 20). By searching enhanced swap-blocking pairs, preferable matching swaps and decoding order swaps can be obtained to improve the sum rate. Based on the resulting matching $\alpha^\mu$ and $\mathbf{W}$, we further update the auxiliary variables $\mathbf{I}$ and $\mathbf{S}$ as

$$S_{ik}^\mu = |\mathbf{h}_i^H \tilde{\mathbf{w}}_k|^2, \quad \forall i, k \in \mathcal{K}, \quad (34)$$

$$I_{ik}^\mu = \tilde{f}_{i \rightarrow k}(\alpha^\mu, \mathbf{W}), \quad \forall i, k \in \mathcal{K}. \quad (35)$$

The above process is repeated until the termination criterion is reached. The computational complexity of
Algorithm 2 Matching-SCA Algorithm for Solving $P_0$

1: Initialize the algorithm accuracy $\epsilon_{\text{Matching}} > 0$. Set the outer loop iteration number as $t = 0$. Set the maximal outer and inner loop iterations as $T_{\text{max}}$ and $L_{\text{max}} = 3$.
2: Initialize the matching states as $\alpha = \mathbf{I}_{K \times K}$ with all users unmatched, and initialize $\mathbf{W}$ with a feasible point.
3: repeat
4: // (Inner loop) SCA-based beamforming optimization
5: Set the inner loop iteration number $l = 0$.
6: repeat
7: Update $\mathbf{W}, \mathbf{I}, \mathbf{S}, \mathbf{r}$ by solving (33) and set $l \leftarrow l + 1$.
8: until $l \geq L_{\text{max}}$.
9: // Matching-based SIC operation optimization
10: For every user $u \in \mathcal{U}$, search for another $u' \in \mathcal{U} \cup \{\emptyset\}$.
11: if $(u, u')$ forms an enhanced swap-blocking pair satisfying condition (i) in Definition 3 then
12: Update $\mu \leftarrow \mu_{u'u'}'$. 
13: end if
14: if $(u, u')$ forms an enhanced swap-blocking pair satisfying condition (ii) in Definition 3 then
15: Update $\mu \leftarrow \mu_{u'u'}'$. 
16: end if
17: Update $\mathbf{S}$ and $\mathbf{I}$ based on current $\mathbf{W}$ and $\alpha^\mu$ using (34) and (35) and set $t \leftarrow t + 1$.
18: until the difference of successive objective values satisfy

$$
|f_0(r^t) - f_0(r^{t-1})|^2 \leq \epsilon_{\text{Matching}} \text{ or } t \geq T_{\text{max}}.
$$

Ensure: Matching $\mu$, SIC operation $\alpha^\mu$, transmit beamforming $\mathbf{W}$, and the optimal value.

Algorithm 2 can be analysed as follows. Specifically, by solving problem (33) via the interior point method, the time complexity of the inner-loop SCA optimization can be given by $O\left((3K^2 + MK)^{3.5}\right)$ [22]. Moreover, the computational complexity of the matching-based SIC operation optimization (i.e., step 11 - step 19) is given by $O(K(K+1))$. Therefore, the computational complexity of Algorithm 2 can be expressed as $O\left(T_{\text{matching}}\left((3K^2 + MK)^{3.5} + K(K+1)\right)\right)$, with $T_{\text{matching}}$ being the iteration number of Matching-SCA.

D. Theoretical Analysis

The properties of the proposed Matching-SCA algorithm with regard to the stability, convergence and optimality can be theoretically analysed as follows.

Proposition 2 (Stability): The proposed Algorithm 2 eventually reaches an enhanced two-sided exchange-stable matching.

Proof: We prove this proposition by contradiction. Assume that there exists an enhanced blocking pair $(u, u')$ in the resulting $\mu^*$, which satisfies condition (i) or condition (ii) in Definition 3. According to step 14-step 20 in Algorithm 2, the proposed algorithm will continue swap until no enhanced blocking pair satisfying the swap conditions exists in the current matching. That is to say, $\mu^*$ should not be the resulting matching, which contradicts the initial assumption. Therefore, it can be concluded that an enhanced exchange stability can be achieved by the proposed Matching-SCA Algorithm 2 eventually. This completes the proof.

Proposition 3 (Convergence): Algorithm 2 converges to an enhanced two-sided exchange-stable matching within limited swap operations.

Proof: Given a matching $\mu$ for the SIC operation, assume that $(u, u')$ is an enhanced swap-blocking pair. Based on Definition 3, there are two cases for $(u, u')$: i) for users $v = \mu(u)$ and $v' = \mu(u')$ satisfying $\alpha_{v'u} = \alpha_{v'u'} = 0$, a conventional swap matching is “approved”, i.e., $U(\mu_{u'u'}) > U(\mu)$; ii) for $v = u'$ and $v' = u$ satisfying $\mu(u) = v$, a decoding order swap matching is “approved”, i.e., $U(\mu_{u'u'}) > U(\mu)$. Since the utilities of all users are non-decreasing for both cases, the sum rate is non-decreasing after each swap operations. Furthermore, owing to the limited transmit power and the inter-user interference, it can be observed that the achievable rate is upper bounded in practice. Therefore, the number of swap operations is limited for every swap-matching process in Algorithm 2. Based on Proposition 2, the proposed algorithm eventually converges to an enhanced two-sided exchange-stable matching when neither matching swaps nor decoding order swaps can further improve the total utility. This completes the proof.

Theorem 1 (Local Optimality): Algorithm 2 converges to a locally optimal matching and beamforming solution.

Proof: See Appendix.

V. Numerical Results

In this section, we present simulation results to verify the effectiveness of the proposed generalized cluster-free multi-antenna NOMA framework and algorithms. We consider one BS that is equipped with $M = 4$ antennas to serve $K = \{3, 4, \ldots, 11\}$ users in downlink transmissions. The noise power is $N_0 = -120$ dBm, and the maximum transmission power of the BS is 27 dBm. To characterize channel correlations, we model $\mathbf{H}$ as the commonly adopted exponential correlation Rayleigh fading channel [32], which can be given by

$$
\mathbf{H} = \tilde{\mathbf{H}}^{-1/2}\mathbf{R}_H^{1/2},
$$

where $\tilde{\mathbf{H}} = [\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_K]$ stacks the small-scale fading vectors. $\mathbf{A}$ is a diagonal matrix with the diagonal elements being diag($\mathbf{A}$) = $[L_1(d_k), L_2(d_k), \ldots, L_K(d_k)]$, where the large-scale fading $L_k(d_k)$ is given by the pathloss model $32.6 + 36.7 \log_{10} d_k$ [33]. Moreover, $\mathbf{R}_H^{1/2}$ denotes the correlation matrix at receivers, where the $(i, j)$-th element signifies the channel spatial correlation of user $i$ and user $j$. For each channel realization, $\mathbf{R}_H$ can be mathematically formulated as

$$
\mathbf{R}_H = \begin{bmatrix}
1 & c & c^2 & \ldots & c^K \\
(c^2)^H & 1 & c & \ldots & c^{K-1} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
(c^K)^H & (c^{K-1})^H & (c^{K-2})^H & \ldots & 1
\end{bmatrix},
$$

(37)

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where \( c = \text{corr} \times e^{j\phi} \) with \( \phi \) being the randomly generated phase within \([0, 2\pi]\) and \( \text{corr} \) controlling the mean channel correlation. To enable fair comparisons among different schemes and verify their performance, we set the rate requirement to \( R_{\text{min}} = 0.8 \) bps/Hz, which ensures the feasibility of all the baseline schemes. The beamforming vectors are randomly initialized. Moreover, for ADMM-SCA, we test \( \Lambda_{\text{ini}} = 20 \) groups of initialization parameters \( \{\beta^0, \alpha^0, \lambda^0, \tilde{\lambda}^0\} \), and empirically choose 3 initialization points for different communication regimes. For each channel realization, we will test the chosen parameters until a feasible solution is obtained. Eventually, if the optimization problem remains infeasible, this channel realization would be ignored in the statistical results.

### A. Convergence Behavior

Fig. 3 presents the convergence behaviours of different algorithms, where we set \( \text{corr} = 0.9 \). In Fig. 3(a), we first compare the convergence behaviours under different user numbers. Here, we introduce the exhaustive search-SCA algorithm as a benchmark, which solves beamforming coefficients \( W \) via SCA while obtaining the globally optimal \( \alpha \) by exhaustively searching all possible combinations. As shown in Fig. 3(a), by iteratively solving the approximated problem based on ADMM-SCA, the objective function of the original problem \( P_0 \), namely the system sum rate, can increase until converge to a locally optimum point. Moreover, with the well-tuned initialized parameters, ADMM-SCA algorithm can achieve the near-optimal SIC operations indicated by exhaustive search-SCA. However, ADMM-SCA incurs high computational complexity, which takes more than 60 iterations for convergence when the number of users increases. In comparison, Matching-SCA yields close performance to ADMM-SCA while achieving much faster convergence, which can converge within 30 inner-loop iterations even in the severely overloaded system and thus significantly reduce the computational complexity. It is worth pointing out that Matching-SCA may outperform ADMM-SCA especially in the overloaded systems. This is because the binary SIC operation variables \( \alpha \) obtained by ADMM-SCA is highly sensitive to initialized parameters, but Matching-SCA can alleviate the dependence on the initialized parameters.

Fig. 3(b) further presents the convergence of the equality constraint violation (namely the residual) of ADMM-SCA. Here, we measure the residual of constraints (11) and (12) by \( \text{vio} = ||\beta + \alpha - 1_{K \times K}||^2 + \sum_{k \in K \setminus K} \sum_{k \in K} (\alpha_{ik} \beta_{ik})^2 \). We can observe that given different numbers of users \( K \), the residual \( \text{vio} \) can decrease rapidly and eventually approach the zero value. This demonstrates that both equality constraints (11) and (12) can be satisfied at the local optimum obtained by ADMM-SCA, thus forcing \( \alpha \) to be binary variables. Combining both the objective convergence in Fig. 3(a) and the residual convergence in Fig. 3(b), we can experimentally verify that the integer relaxation and the approximation in ADMM-SCA affect neither the feasibility nor the local optimality of the obtained solutions with respect to the original problem \( P_0 \).

In Fig. 3(c), we further compare the proposed Matching-SCA algorithm with conventional matching-based algorithms. Here, we consider three baseline matching-based algorithms:

- **Conventional Matching-SCA**: where the conventional matching swap operations and the conventional swap rule are utilized, as presented in Definition 2.
- **Pairwise increment based Matching-SCA**: where the conventional matching swap operation (without the swaps of SIC decoding order) is utilized, while the swap rule is modified into ensuring the sum utility increment of the involved user pairs during swaps.

- **Enhanced swap rule based Matching-SCA**: where the conventional matching swap operation (without the swaps of SIC decoding order) is utilized, while the enhanced swap rule as described in condition (i) in **Definition 3** is adopted, which ensures the sum utility increment of all multiplexed users.

From Fig. 3(c), the conventional Matching-SCA algorithm leads to the worst sum rate since it cannot optimize the SIC decoding order and only focuses on users’ individual utilities, which typically lacks the local optimality guarantee at the convergence. Furthermore, the enhanced swap rule based Matching-SCA algorithm outperforms the pairwise increment based Matching-SCA algorithm, since it takes into account the total utility of all users. The proposed Matching-SCA algorithm achieves the highest performance, which validates the effectiveness of the enhanced swap operations and enhanced swap rule as proposed in **Definition 3**.

### B. Performance Comparisons

To demonstrate the performance of the proposed generalized cluster-free framework as analysed in **Remark 1**, we consider four baseline approaches, namely cluster-based NOMA (CB-NOMA), enhanced cluster-based NOMA (enhanced CB-NOMA), beamformer-based NOMA (BB-NOMA), and SDMA. Here, enhanced CB-NOMA is an improved variant of conventional CB-NOMA. Unlike CB-NOMA which shares a single beamforming vector in each cluster, enhanced CB-NOMA exploits a dedicated beamforming vector for each user to achieve potentially better spatial multiplexing. Both CB-NOMA and enhanced CB-NOMA configure SIC operations by performing user clustering according to users’ channel correlations [34]. Without loss of generality, the beamforming vectors of enhanced CB-NOMA and BB-NOMA are solved by sequentially optimizing problem (33) via SCA. Similarly, the beamforming vectors of SDMA can be solved by ignoring constraint (33c) and other constraints/variables related to SIC decoding in (33). Furthermore, for CB-NOMA, a single beamforming vector is constructed for each user cluster. Thereafter, the cluster-specific beamforming vectors and the transmit power allocation are jointly optimized by invoking SCA and alternating optimization. We obtain the following simulation results by averaging over 100 channel realizations.

Fig. 4 demonstrates the performance of different methods under different channel correlations in the underloaded regime, where $M = 4$ and $K = 3$. From Fig. 4(a), it can be observed that the proposed generalized cluster-free NOMA framework achieves the highest sum rate regardless of the variations in the channel correlations. Furthermore, CB-NOMA and enhanced CB-NOMA reduce to SDMA in the underloaded regime since there is only one user in each cluster, so they achieve the same performance. Since BB-NOMA suffers from the *SIC overuse* problem when users experience low channel correlations, in this case it yields worse performance than CB-NOMA/SDMA schemes. However, when user number increases, the BB-NOMA outperforms the CB-NOMA/SDMA schemes since it can eliminate the interference among the highly channel-correlated users more adequately. On the other hand, from the system viewpoint, the overheads for implementing cluster-free NOMA and conventional approaches mainly differ in the SIC decoding complexity, which can be evaluated by

$$
\sum_{i \in K} \sum_{k \in K \setminus \{i\}} (\alpha_{ik} + \alpha_{ki}).
$$

In Fig. 4(b), we compare the SIC decoding complexity caused by different approaches. Owing to the cluster-free feature, the SIC decoding complexity of the proposed framework increases with users’ channel correlations, and is higher than CB-NOMA/SDMA but lower than BB-NOMA. This demonstrates that it can achieve *scenario-adaptive* SIC operations and efficient interference suppression.

Fig. 5 presents performance comparisons under different spatial correlations in the overloaded regime, where the number of users is $K = 6$. From Fig. 5(a), when the channel correlation increases, the sum rate increases in the BB-NOMA approach, but decreases in CB-NOMA, SDMA, and...
TABLE I
COMPARISONS OF SUM RATE FOR DIFFERENT MULTI-ANTENNA NOMA METHODS

| Communication regime | SDMA | CB-NOMA | BB-NOMA | The prop. |
|----------------------|------|---------|---------|-----------|
| Low correlation      | Underloaded | High | High | Low | Best |
|                      | Overloaded | Medium | High | Low | Best |
|                      | Severely overloaded | Low | Medium | High | Best |
| High correlation     | Underloaded | High | High | Low | Best |
|                      | Overloaded | Low | Medium | High | Best |
|                      | Severely overloaded | Low | Medium | High | Best |

Fig. 5. Performance comparisons under various channel correlations $\text{corr}$ in the overloaded regime. $M = 4, K = 6$.

Fig. 6. Sum rate under different numbers of connected users in the overloaded regime.

Based on the above numerical results, the performance comparisons of the proposed framework, SDMA, BB-NOMA, and CB-NOMA under different communication regimes can be summarized as Table I. It can be observed that the proposed generalized cluster-free framework leads to the highest performance regardless of the system loadings and channel correlations.

baseline methods, while the BB-NOMA yields the worst sum rate. Similar to the underloaded scenarios, the performance of BB-NOMA exceeds other baseline schemes when channel correlation increases. From Fig. 5(b), the proposed framework achieves the highest sum rate while maintaining a moderate SIC decoding complexity in the overloaded regime, which verifies its applicability and effectiveness in different scenarios.

Fig. 6 further shows the performance comparisons under different numbers of connected users in the overloaded regime. To reduce the computational complexity, we consider Matching-SCA algorithm here. We set $\text{corr} = 0.6$ and $\text{corr} = 0.9$ for low channel correlation and high channel correlation, respectively. In both scenarios, SDMA yields the lowest sum rate than other methods. When there are fewer users, CB-NOMA outperforms BB-NOMA and the performance gap decreases with channel correlation. While the user number increases, BB-NOMA yields a higher sum rate than CB-NOMA and the performance gap increases with channel correlation. Owing to the efficient multiplexing, the proposed framework outperforms both CB-NOMA and BB-NOMA despite the varying channel correlations.
VI. CONCLUSION

A novel generalized multi-antenna NOMA framework has been proposed based on the cluster-free SIC, which can reap the gains and overcome the shortcomings of conventional approaches, thus enabling a scenario-adaptive multi-antenna NOMA paradigm towards NGMA. The transmit beamforming and the SIC operations were jointly optimized to maximize the sum rate subject to the SIC decoding conditions and data rate constraints of users. To tackle the resulting highly-coupled NP-hard MINLP problem, an ADMM-SCA algorithm was developed to obtain the stationary solution. Furthermore, to accelerate the convergence and overcome the over-dependence on parameter initialization of ADMM-SCA, a Matching-SCA algorithm was proposed. Based on an extended many-to-many matching procedure, the proposed Matching-SCA can converge to an enhanced exchange-stable matching which guarantees the local optimality. Our numerical results showed that the proposed Matching-SCA algorithm has comparable performance to ADMM-SCA, and achieves fast convergence despite the increment of connected users. Numerical results also verified that the proposed framework can outperform conventional multi-antenna NOMA approaches under varying channel correlations and both underloaded and overloaded regimes, which confirmed the effectiveness of the proposed framework and motivated the future research on the generalized cluster-free NOMA for empowering NGMA.

APPENDIX

PROOF OF THEOREM 1

Let \( \{ \alpha^t, W^t, I^t, S^t, r^t \} \) and \( \{ W^{t,l}, I^{t,l}, S^{t,l}, r^{t,l} \} \) denote the optimal values obtained from outer-loop iteration \( t \) and the corresponding inner-loop iteration \( l \) of Algorithm 2, respectively. In each inner-loop SCA iteration \( l \), functions \( q_1 (w_k) \) and \( q_2 (I_{ik}) \) defined in Section III-B are linearized based on the previous iterate points \( w^{t-1} \) and \( I^{t-1} \). From (21), by solving problem (33), the obtained \( S_{ik}^{t,l} \) and \( W_{ik}^{t,l} \) always satisfy \( S_{ik}^{t,l} \leq q_1 \left( w_{ik}^{t,l} \right) \leq q_2 \left( I_{ik}^{t-1} \right) \) from (19), the inequality \( r_{ik}^{t,l} + q_2 \left( I_{ik}^{t-1} \right) \leq \log_2 \left( I_{ik}^{t,l} + S_{ik}^{t,l} \right) \) holds, \( \forall i,k \in K \). Hence, given the fixed SIC operation variables \( \alpha^t \), any feasible solution \( \{ W^{t,l}, I^{t,l}, S^{t,l}, r^{t,l} \} \) of problem (33) is also feasible to \( P_1 \). Therefore, the objective \( f (\alpha, W, I, S, r) \) is non-decreasing at each inner-loop SCA iteration \( 1 < l \leq L_{\text{max}} \) as [27], [35]

\[
\begin{align*}
& f \left( \alpha^{t-1}, W^{t-1}, I^{t-1}, S^{t-1}, r^{t-1} \right) \\
& \leq f \left( \alpha^{t-1}, W^{t-1}, I^{t-1}, S^{t-1}, r^{t-1} \right) \\
& \leq f \left( \alpha^{t-1}, W^{t,l}, I^{t,l}, S^{t,l}, r^{t,l} \right) \\
& \leq f \left( \alpha^{t-1}, W^{t,\text{MAX}}, I^{t,\text{MAX}}, S^{t,\text{MAX}}, r^{t,\text{MAX}} \right).
\end{align*}
\]

Moreover, according to Proposition 3, the sum rate is non-decreasing after each enhanced swap operation. Thus, the sequence \( \{ f \left( \alpha^{t,l}, W^{t,l}, I^{t,l}, S^{t,l}, r^{t,l} \right) \} \) \( t \in \{1,2,...,T_{\text{MAX}}^{\text{IMPAK}} \} \) has monotonic convergence. Considering the limited transmit power, Algorithm 2 will terminate until there are no SCA updates nor swap operations can further increase the sum rate. Hence, the local optimality of the solutions can be guaranteed, which completes the proof.

REFERENCES

[1] Cisco Annual Internet Report (2018–2023) White Paper, Cisco, San Jos, CA, USA, 2020.
[2] W. Saad, M. Bennis, and M. Chen, “A vision of 6G wireless systems: Applications, trends, technologies, and open research problems,” IEEE Netw., vol. 34, no. 3, pp. 134–142, May 2019.
[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, “The roadmap to 6G: AI empowered wireless networks,” IEEE Commun. Mag., vol. 57, no. 8, pp. 86–88, Aug. 2019.
[4] Y. Liu et al., “Evolution of NOMA toward next generation multiple access (NGMA) for 6G,” IEEE J. Sel. Areas Commun., vol. 40, no. 4, pp. 1037–1071, Apr. 2022.
[5] 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 Commun. Surveys Tuts., vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.
[6] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, “Nonorthogonal multiple access for 5G and beyond,” Proc. IEEE, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.
[7] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, “Resource allocation for downlink NOMA systems: Key techniques and open issues,” IEEE Wireless Commun., vol. 23, no. 2, pp. 40–47, Apr. 2018.
[8] Z. Ding, C. Zhang, J. Wang, Y. Liang, L. Yang, and X. You, “Signal processing for MIMO-NOMA: Present and future challenges,” IEEE Wireless Commun., vol. 25, no. 2, pp. 32–38, Apr. 2018.
[9] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, “Massive access for 5G and beyond,” IEEE J. Sel. Areas Commun., vol. 39, no. 3, pp. 613–637, Mar. 2021.
[10] Z. Ding, F. Adachi, and H. V. Poor, “The application of MIMO to nonorthogonal multiple access,” IEEE Trans. Wireless Commun., vol. 15, no. 1, pp. 537–552, Jan. 2016.
[11] Y. Liu, H. Xing, C. Pan, A. Nallanathan, M. Elkashlan, and L. Hanzo, “Multiple-antenna-assisted non-orthogonal multiple access,” IEEE Wirel. Commun., vol. 25, no. 2, pp. 17–23, Apr. 2018.
[12] M. F. Hanif, Z. Ding, T. Ratarajah, and G. K. Karagiannidis, “A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems,” IEEE Trans. Signal Process., vol. 64, no. 1, pp. 36–48, Jan. 2016.
[13] Q. Sun, S. Han, Z. Xu, S. Wang, I. Chih-Lin, and Z. Pan, “Sum rate optimization for MIMO-nomog multiple access systems,” in Proc. IEEE Wireless Commun. Netw. Conf., New Orleans, LA, USA, May 2015, pp. 747–752.
[14] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, “On the application of quasi-degradation to MISO-NOMA downlink,” IEEE Trans. Signal Process., vol. 64, no. 23, pp. 6174–6189, Dec. 2016.
[15] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, “Low complexity beamforming and user selection schemes for 5G MIMO-NOMA systems,” IEEE J. Sel. Areas Commun., vol. 35, no. 12, pp. 2708–2722, Dec. 2017.
[16] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, “Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster,” IEEE J. Sel. Areas Commun., vol. 35, no. 10, pp. 2413–2424, Oct. 2017.
[17] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, “Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA,” IEEE Trans. Veh. Technol., vol. 68, no. 3, pp. 3000–3004, Mar. 2019.
[18] X. Hu, C. Zhong, X. Chen, W. Xu, and Z. Zhang, “Cluster grouping and power control for angle-domain mmWave MIMO NOMA systems,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 5, pp. 1167–1180, Sep. 2019.
[19] Y. Fu, M. Zhang, L. Salaun, C. W. Sung, and C. S. Chen, “Zero-forcing oriented power minimization for multi-cell MISO-NOMA systems: A joint user grouping, beamforming, and power control perspective,” IEEE J. Sel. Areas Commun., vol. 38, no. 8, pp. 1925–1940, Aug. 2020.
[20] Z. Ding, “NOMA beamforming in SDMA networks: Riding on existing beams or forming new ones?” IEEE Commun. Lett., vol. 26, no. 4, pp. 868–871, Apr. 2022.
[21] Z. Xiao, L. Zhu, Z. Gao, D. O. Wu, and X. Xia, “User fairness non-orthogonal multiple access (NOMA) for millimeter-wave communications with analog beamforming,” IEEE Trans. Wireless Commun., vol. 18, no. 7, pp. 3411–3424, Jul. 2019.

[22] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge, U.K.: Cambridge Univ. Press, 2004.

[23] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximation,” Comput. Math. Appl., vol. 2, no. 1, pp. 17–40, 1976.

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Found. Trends Mach. Learn., vol. 3, no. 1, pp. 1–122, Nov. 2011.

[25] B. R. Marks and G. P. Wright, “A general inner approximation algorithm for nonconvex mathematical programs,” Oper. Res., vol. 26, no. 4, pp. 681–683, 1978.

[26] M. Grant and S. Boyd. (2016). CVX: MATLAB Software for Disciplined Convex Programming. [Online]. Available: http://cvxr.com/cvx

[27] G. R. Lanckriet and B. K. Sriperumbudur, “On the convergence of the concave–convex procedure,” in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 1759–1767.

[28] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximation,” SIAM J. Optim., vol. 25, no. 4, pp. 2434–2460, 2015.

[29] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, “Many-to-many matching with externalities for device-to-device communications,” IEEE Wireless Commun. Lett., vol. 6, no. 1, pp. 138–141, Feb. 2017.

[30] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, “Many-to-many matching with externalities for device-to-device communications,” IEEE Wireless Commun. Lett., vol. 6, no. 1, pp. 138–141, Feb. 2017.

[31] W. Ni, X. Liu, Y. Liu, H. Tian, and Y. Chen, “Resource allocation for multi-cell IRS-aided NOMA networks,” IEEE Trans. Wireless Commun., vol. 20, no. 7, pp. 4253–4268, Jul. 2021.

[32] E. Bodine-Baron et al., “Peer effects and stability in matching markets,” in Proc. Int. Symp. Algorithmic Game Theory, 2011, pp. 117–129.

[33] J.-P. Kermoal, L. Schumacher, K. I. Pedersen, P. E. Mogensen, and F. Frederiksen, “A stochastic MIMO radio channel model with experimental validation,” IEEE J. Sel. Areas Commun., vol. 20, no. 6, pp. 1211–1226, Aug. 2002.

[34] Further Advancements for E-UTRA Physical Layer Aspects (Release 9), 3GPP, document TS 36.814, Mar. 2010.

[35] L. Dai, B. Wang, M. Peng, and S. Chen, “Hybrid precoding-based multimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer,” IEEE J. Sel. Areas Commun., vol. 37, no. 1, pp. 131–141, Jan. 2019.

[36] A. Beck, A. Ben-Tal, and L. Tetruashvili, “A sequential parametric convex approximation method with applications to nonconvex truss topology design problems,” J. Glob. Optim., vol. 47, no. 1, pp. 29–51, May 2010.

Yuanwei Liu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, King’s College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. From 2017 to 2021, he was a Lecturer (Assistant Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, where he has been a Senior Lecturer (Associate Professor) since August 2021. His research interests include non-orthogonal multiple access, reconfigurable intelligent surface, integrated sensing and communications, and machine learning. He was a co-recipient of the Best Student Paper Award in IEEE VTC2022-Fall, the Best Paper Award in ISWCS 2022, and the 2022 IEEE SPCC-TC Best Paper Award. He received the IEEE ComSoc Outstanding Young Researcher Award for EMIA in 2020. He received the 2020 IEEE Signal Processing and Computing for Communications (SPCC) Technical Early Achievement Award and the IEEE Communication Theory Technical Committee (CTTC) 2021 Early Achievement Award. He received the IEEE ComSoc Outstanding Nominee for Best Young Professionals Award in 2021. He has been the Academic Chair of the Next Generation Multiple Access Emerging Technology Initiative. He serves as the Publicity Co-Chair for IEEE VTC 2019-Fall and the Symposium Co-Chair for Cognitive Radio and AI-Enabled Networks for IEEE GLOBECOM 2022 and Communication Theory for IEEE GLOBECOM 2023. He serves as the Chair for the Special Interest Group (SIG) in SPCC Technical Committee on Signal Processing Techniques for Next Generation Multiple Access and the Vice-Chair for the SIG WTC on Reconfigurable Intelligent Surfaces for Smart Radio Environments. He serves as a Senior Editor for IEEE COMMUNICATIONS LETTERS and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON COMMUNICATIONS. He serves as the Guest Editor for the Special Issue on Next Generation Multiple Access of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the Special Issue on Signal Processing Advances for Non-Orthogonal Multiple Access of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and the Special Issue on Next Generation Multiple Access for 6G of IEEE NETWORK. He has been a Web of Science Highly Cited Researcher since 2021. He has been a Distinguished Lecturer of the IEEE Vehicular Technology Society and the IEEE Communication Society. More information can be found at: http://www.eecs.qmul.ac.uk/yuanwei

Xiaohua Xu received the B.Eng. degree in geographical condition monitoring from Wuhan University, China, in 2017, where she is currently pursuing the Ph.D. degree with the Electronic Information School. From 2021 to 2022, she was a Visiting Student with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. Her current research interests include millimeter-wave communications, non-orthogonal multiple access, reconfigurable intelligent surface, and machine learning.

Xidong Mu (Member, IEEE) received the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2022. He is currently a Post-Doctoral Researcher with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. His research interests include non-orthogonal multiple access (NOMA), IRSs/RISs aided communications, integrated sensing and communications, semantic communications, and optimization theory.

He was a recipient of the Best Paper Award in ISWCS 2022, the 2022 IEEE SPCC-TC Best Paper Award, and the Best Student Paper Award in IEEE VTC2022-Fall. He received the Exemplary Reviewer Certificate of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2020 and the IEEE COMMUNICATIONS LETTERS in 2021 and 2022. He serves as the Secretary for the IEEE ComSoc Next Generation Multiple Access (NGMA) Emerging Technology Initiative and the Special Interest Group (SIG) in SPCC Technical Committee on Signal Processing Techniques for NGMA. He also serves as a Guest Editor for the Special Issue on Signal Processing for Sensing, Communication and Computation Integration in Next-Generation Wireless Networks of IET Signal Processing. More information can be found at: https://xidongmu.github.io/
Qimei Chen (Member, IEEE) received the Ph.D. degree from the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, in 2017. She was a Visiting Student with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA, USA, from 2015 to 2016. From 2017 to 2022, she was an Associate Researcher with the School of Electronic Information, Wuhan University, Wuhan, China, where she has been an Associate Professor since 2023. Her research interests include intelligent edge communication, unlicensed spectrum, massive MIMO-NOMA, and machine learning in wireless communications. She has served as a TPC Member for IEEE conferences, such as ICC, GLOBECOM, and WCNC. She received the Exemplary Reviewer Certificate of the IEEE WIRELESS COMMUNICATIONS LETTERS in 2020. She also served as a Guest Editor for the Special Issue on Heterogeneous Networks of Sensors (MDPI).

Zhiguo Ding (Fellow, IEEE) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree in electrical engineering from Imperial College London in 2005. He is currently a Professor of communications with The University of Manchester and Khalifa University. Previously, he worked with Queen’s University Belfast, Imperial College London, Newcastle University, and Lancaster University. From October 2012 to September 2023, he has also been an Academic Visitor with the Prof. Vincent Poor’s Group, Princeton University. His research interests include 5G networks, machine learning, cooperative and energy harvesting networks, and statistical signal processing. He is a member of the Global Research Advisory Board of Yonsei University. He recently received the EU Marie Curie Fellowship from 2012 to 2014, the Top IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY Editor in 2017, the IEEE Heinrich Hertz Award in 2018, the IEEE Jack Neubauer Memorial Award in 2018, the IEEE Best Signal Processing Letter Award in 2018, the Friedrich Wilhelm Bessel Research Award in 2020, and the Best Paper Award at IEEE ICCCN-21. He is serving as an Area Editor for the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY and an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He was an Editor of IEEE WIRELESS COMMUNICATION LETTERS, IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE COMMUNICATION LETTERS, from 2013 to 2016. He is a Web of Science Highly Cited Researcher in two disciplines (2019–2023). He is an IEEE ComSoc Distinguished Lecturer.