



**QUEEN'S  
UNIVERSITY  
BELFAST**

## Two-stage reinforcement learning for MIMO-NOMA with hard-latency constraints

Zhang, L., Liu, A., Xu, X., Mu, X., & Liu, Y. (2025). Two-stage reinforcement learning for MIMO-NOMA with hard-latency constraints. *IEEE Transactions on Communications*. Advance online publication. <https://doi.org/10.1109/TCOMM.2025.3576919>

**Published in:**  
IEEE Transactions on Communications

**Document Version:**  
Peer reviewed version

**Queen's University Belfast - Research Portal:**  
[Link to publication record in Queen's University Belfast Research Portal](#)

### **Publisher rights**

Copyright 2025 the authors.

This is an accepted manuscript distributed under a Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

### **General rights**

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### **Take down policy**

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact [openaccess@qub.ac.uk](mailto:openaccess@qub.ac.uk).

### **Open Access**

This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: <http://go.qub.ac.uk/oa-feedback>

# Two-Stage Reinforcement Learning for MIMO-NOMA with Hard-Latency Constraints

Luyuan Zhang, An Liu, *Senior Member, IEEE*, Xiaoxia Xu,  
Xidong Mu, *Member, IEEE*, and Yuanwei Liu, *Fellow, IEEE*

**Abstract**—A novel hard-latency guaranteed cluster-free multiple-input multiple-output non-orthogonal multiple access (MIMO-NOMA) framework is proposed to deal with burst traffics that commonly occur in real-world scenarios. The hard-latency constrained effective throughput (HLC-ET) maximization problem is formulated, which jointly optimizes the beamforming and cluster-free success interference cancellation (SIC) operations. To address the resultant problem, a two-stage reinforcement learning (RL)-based algorithm is developed to capture system uncertainty, where the large-dimension optimization is decoupled into two stages to reduce the action space and fasten convergence of RL. In the long-term stage, we aim to maximize the HLC-ET, and a hybrid RL algorithm with policy reuse is adopted to control the priority weights to construct the weighted sum rate (WSR) function of users. In the short-term stage, a branch-and-bound (BB) based algorithm is further developed to obtain the optimal solution of the WSR maximization problem. The BB-based algorithm is proved to guarantee the convergence to an  $\epsilon$ -optimal solution of the WSR maximization problem within a finite number of steps. To accelerate computation in the short-term stage, a channel correlation based two-loop greedy (CC-TLG) algorithm is proposed to significantly reduce the complexity with almost no performance loss compared to the BB-based algorithm. Finally, simulations demonstrate the advantages of the proposed two-stage RL based joint beamforming and SIC optimization (TSRL-JBSO) algorithm over conventional RL-based and non-RL based algorithms.

**Index Terms**— Beamforming, hard latency, non-orthogonal multiple access (NOMA), reinforcement learning.

## I. INTRODUCTION

One of critical performance targets of sixth-generation (6G) wireless systems is that the spectral efficiency (SE) and energy efficiency (EE) have to be 5-10 and 10-100 times higher than for 5G, respectively. With the increasing demand for large capacity in wireless networks, the conventional multiple access schemes cannot fully meet the SE target in 6G. Non-orthogonal multiple access (NOMA) [1] is a promising technique, which adopts the non-orthogonal principle to enable multiple users to share time domain, frequency domain or code domain resources. Superposition coding (SC) and successive interference cancellation (SIC) are two key technologies in

NOMA for achieving the non-orthogonal use of radio resources and interference management. Therefore, NOMA can achieve high SE and better user fairness [2], [3] compared to orthogonal multiple access (OMA). To employ NOMA with multiple antenna technologies, two categories of beamforming strategies have been proposed, namely, beamformer-based MIMO-NOMA (BB-NOMA) and cluster-based NOMA (CB-NOMA), but their effectiveness relies on specific scenarios [4]. Therefore, a unified cluster-free NOMA framework was proposed in [5], which enables SIC to be flexibly implemented, thus breaking the shortcoming of the existing approaches.

Another important target of 6G is strict latency constraint. Ultra-reliable and low latency communications (URLLC), has always been a key requirement for many applications such as public safety, telemedicine and etc [6], [7], [8] in fifth-generation (5G). However, previous research on conventional 5G URLLC use cases has primarily focused on short packet transmission, which fails to meet the comprehensive requirements of future wireless communication systems. Extend reality (XR), which is an umbrella term for different types of realities such as virtual reality (VR), augmented reality (AR), and mixed reality (MR), has been regarded as an emerging application for URLLC in 6G with new traffic characteristics and more stringent requirements. Different from short-packet transmissions in conventional URLLC, XR frame has a much larger size and requires multiple timeslots to complete the transmission, which makes resource scheduling more difficult to meet the hard delay constraints [9].

In addition to the quasi-periodical traffic in XR, burst traffic with both large frame size and random arrivals in some real world low latency communication scenarios has become the leading cause of network congestion or even collapse [10]. There have been many works apply NOMA to URLLC [11] and XR/VR networks [12] to improve SE as well as to reduce latency. However, to the best of our knowledge, existing works in the literature have not considered NOMA for hard-latency transmission under burst traffic, which still face various technical challenges to be addressed.

## A. Related Works

1) *Studies on NOMA* : In the past few years, extensive efforts have been devoted to the development of MIMO-NOMA. Existing MIMO-NOMA systems can be broadly classified into beamformer-based NOMA and cluster-based NOMA, which exploit different beamforming and SIC operation designs. Beamformer-based NOMA directly serves users through different beamforming vectors, and meanwhile reduce the spatial

An Liu and Luyuan Zhang are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China. (email: {22131107, anliu}@zju.edu.cn).

Xiaoxia Xu is with the School of Electronic Engineering and Computer Science, Queen Mary University of London. (email: x.xiaoxia@qmul.ac.uk).

Xidong Mu is with the Centre for Wireless Innovation (CWI), Queen's University Belfast, U.K. (email: x.mu@qub.ac.uk).

Yuanwei Liu is with the Department of Electrical and Electronic Engineering, The University of Hong Kong. (email: yuanwei@hku.hk).

interference by carrying out SIC between the multiplexed users [13], [14], [15]. The authors of [14] investigated the optimal power allocation in a two-user downlink MIMO-NOMA, which can achieve the capacity region of the MIMO broadcast channel under the derived channel state information (CSI) condition. Different from beamformer-based NOMA, cluster-based NOMA [16], [17], [18] partitions the highly channel correlated users into the same cluster, and allows each cluster to share the same beamforming vector. The authors of [19] proposed a distributed user grouping, beamforming and power control algorithm for power consumption minimization. It can be observed that both beamformer-based NOMA and cluster-based NOMA adopt scenario-centric SIC operation designs [4], since the former assigns all users to a single cluster for SIC decoding, while the latter assumes that the users in the same cluster have high channel correlations. To overcome non-ideal clustering, the authors of [5] proposed a novel generalized downlink NOMA transmission framework with the concept of cluster-free SIC, and it provides a generalized modeling, which unifies the existing approaches.

2) *Studies on NOMA-URLLC Systems:* URLLC has always been a key requirement for a number of applications such as public safety, telemedicine and etc, and several works have applied NOMA to URLLC since NOMA is able to reduce the transmission delay for each user by providing additional access in the power domain [20]. As URLLC uses finite blocklength (FBL) transmissions, the well-known Shannon capacity is no longer the accurate approximation of data rates [21]. The authors of [22] proposed a downlink MIMO-NOMA framework for the URLLC networks, which focus on a two-user case and can be extended to multi-user scenarios to enhance the connectivity. In [23], the authors proposed a static multi-user NOMA-URLLC framework based on hybrid automatic repeat request re-transmissions. The authors of [24] proposed three multi-agent deep reinforcement learning (MADRL) based frameworks to maximize energy efficiency while satisfying URLLC requirements in an uplink URLLC-NOMA system. However, all of these works consider short-packet communications (SPC), which are not suitable for communications with large packets, e.g., VR/XR transmissions.

3) *Studies on NOMA-XR/VR Systems:* XR/VR has been regarded as an emerging application for URLLC in 6G with new traffic characteristics and more stringent requirements, whose frame has a much larger size and requires multiple timeslots to complete the transmission. NOMA has been leveraged in XR/VR to improve SE of transmitting XR/VR content as well as to reduce transmission latency. The authors of [12] considered a NOMA assisted VR content transmission network, and formulated an optimization problem to minimize the sum of weighted total energy consumption and VR content distortion with the delay constraint. The authors of [25] constructed a multi-user uplink NOMA system to address the challenges brought by XR devices such as ultra-massive access, real-time synchronization, and applied an exact linear search based algorithm for finding the optimal policy. A cooperative NOMA (Co-NOMA) scheme was introduced in [26], to strike a trade-off between the throughput and fairness between XR devices.

## B. Motivations and Contributions

In this paper, we aim to solve the hard-latency constrained transmission problem in the generalized cluster-free NOMA framework under burst traffic. Different from the short packet in URLLC that can be transmitted in one timeslot and the quasi-periodical traffic in XR, the packets in the considered burst traffic have both large frame size and random arrivals. Packets with large frame size need multiple timeslots transmission and can only be viewed transmitted successfully at the final timeslot. Therefore, in order to capture the transmission state of packets, we define a hard-latency constrained effective throughput (HLC-ET), which only considers packets which have been successfully delivered before the hard delay constraints. To get rid of the unrealistic assumptions on traffic/channel statistics, e.g., simple and known traffic/channel statistics, we adapt reinforcement learning (RL) method to tackle long-term latency constraints. However, the problem is rather challenging due to two main reasons: On the one hand, our objective is to maximize the HLC-ET, which is a non-convex problem, and the sparse reward brought by large packets that need multi-timeslots transmission as well as the large state space make the algorithm harder to converge. On the other hand, different from some works in the literature [11] [27] where the clustering is either neglected or addressed separately from beamforming, this paper jointly optimizes the beamforming and SIC operations for a cluster-free NOMA network, which introduces both integer and continuous variables. Therefore, directly assigning all decision variables, i.e., beamforming, SIC operations, and priority weights as RL agent's action will lead to a large-dimension action space and slow convergence. To overcome this challenge, we propose a novel two-stage reinforcement learning based joint beamforming and SIC optimization (TSRL-JBSO) algorithm. The main contributions of this work are:

- **A two-stage RL based hard-latency constrained NOMA framework:** To handle both the environment uncertainty and the large-dimension optimization variables, we propose a novel latency-guaranteed transmission framework, which consists of two stages in different time scales. In the long-term stage, the priority weights for weighted sum rate (WSR) of users are determined by RL, which ensures hard latency constraints by solving the Markov Decision Process (MDP) problem. Using the assigned priority weights, the short-term stage (i.e., each iteration of the long-term stage) further maximizes the WSR by jointly optimizes the beamforming and SIC operations. The proposed framework significantly fastens the RL convergence speed by reducing the action space and adopting a hybrid RL algorithm with policy reuse.
- **The optimal solution of WSR maximization for joint beamforming and cluster-free SIC:** The joint beamforming and cluster-free SIC optimization problem for WSR maximization in the short-term stage is a challenging mixed-integer nonlinear programming (MINLP) problem. The globally optimal solution for this coupled NP-hard problem is still unexplored currently. In this paper, we develop a global optimization algorithm based

on branch-and-bound (BB) to address this challenge. The developed BB can find the global optimum solution of the MINLP problem by successively branching the feasible solution space and solving the convex relaxation problem to evaluate its bounding.

- **A low-complexity and near-optimal algorithm for WSR maximization:** Although the BB-based algorithm can find the optimal solution of the MINLP problem, it requires many iterations to converge and has high complexity, which is not suitable to be applied in each iteration of the RL algorithm. Therefore, to accelerate computation on each iteration, we propose a channel correlation based two-loop greedy (CC-TLG) algorithm to significantly reduce the complexity of solving the MINLP with almost no performance loss compared to the BB-based algorithm. CC-TLG maximizes the WSR of users based on the user channel correlation coefficients, where the outer loop add users one by one based on the WSR in a greedy manner, and the inner loop generates a near-optimal SIC operation based on both the channel correlation coefficients and WSR in a greedy way.
- **Convergence analysis and performance evaluation:** We prove that the BB-based algorithm in the short-term stage can guarantee the convergence to an  $\epsilon$ -optimal solution of the WSR maximization problem within a finite number of steps. The convergence of the RL algorithm in the long-term stage is also established. Simulations show that the proposed low-complexity algorithm is able to solve the MINLP with almost no performance loss and much lower complexity compared to the BB algorithm, and that the proposed TSRL-JBSO achieves higher HLC-ET as well as lower packet loss probability compared to the baseline.

### C. Organization and Notation

The rest of the paper is organized as follows. In Section II, we illustrate the system model considered in this paper. In Section III, the resource scheduling problem in the cluster-free NOMA with hard-latency constraint is formulated as a two-stage RL problem. In Section IV, a two-stage reinforcement learning based joint beamforming and beamforming optimization algorithm is proposed. In Section V, a low-complexity method to solve the MINLP problem in the short-term stage is proposed. Section VI showcases the simulation results. Finally, we conclude this paper in Section VII.

*Notation:* Vectors and matrices are denoted by bold-face letters.  $\|\mathbf{x}\|$  denotes the Euclidean norm of a vector  $\mathbf{x}$ .  $\mathbf{x}^T$  and  $\mathbf{x}^H$  denote the transpose and Hermitian conjugate of vector  $\mathbf{x}$ .  $\mathbf{I}_{N \times N}$  indicates an identity matrix of size  $N$ .  $\mathbf{1}_{N \times N}$  denotes an  $M \times N$  all ones matrix.

## II. SYSTEM MODEL AND PROBLEM FORMULATION

As shown in Fig. 1, we consider a downlink cluster-free MIMO NOMA system, which is a generalized downlink NOMA transmission framework with the concept of cluster-free SIC. It consists a base station (BS) which is equipped with  $N_T$  antennas, and a set of  $K$  users  $\mathcal{K} = \{1, \dots, K\}$  which is equipped with  $N_R$  antennas ( $N_T > N_R$ ). For

clarify, we consider single-stream transmission mode where each user only transmits a single stream using maximum ratio combining (MRC) at the receiver. In this case, each user can be equivalently viewed as a single-antenna user, and we denote the equivalent channel of user  $i$  after MRC as  $\mathbf{h}_i$ . In contrast to cluster-based NOMA which assigns a single beamforming vector for each cluster, each user  $k$  is assigned a dedicated transmit beamforming vector  $\mathbf{w}_k \in \mathbb{C}^{N_T \times 1}$  in the considered cluster-free system.

### A. Signal Model

For each user, the BS utilizes NOMA superposition and beamforming simultaneously. Let  $\mathbf{W}(t) = [\mathbf{w}_k]_{k \in \mathcal{K}} \in \mathbb{C}^{N_T \times K}$ , where  $[\mathbf{w}_k]$  denotes the transmit beamforming matrix for user  $k$ . Let  $s_k$  denote the transmitted data symbol for user  $k$ , then the received signal at user  $k$  is expressed as

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \sum_{k' \neq k} \mathbf{h}_k^H \mathbf{w}_{k'} s_{k'} + n_k, \forall k \in \mathcal{K}, \quad (1)$$

where  $n_k \sim \mathcal{CN}(0, \delta^2)$  indicates the additive white Gaussian noise (AWGN) at user  $k$  with zero mean and variance  $\delta^2$ . Note that the time slot index  $t$  is omitted for concise. The data symbols are normalized, i.e.,  $\mathbb{E}[|s_k|^2] = 1$ . The first item of (1) denotes the desired signal, and the second item denotes the multi-user interference.

### B. SIC Procedure

To efficiently mitigate the multi-user interference, cluster-free SIC is flexibly implemented between any two channel-correlated users without the pre-defined user clusters. Denote  $\alpha_{m,k} \in \{0, 1\}$ ,  $\forall m, k \in \mathcal{K}$  as the indicator which specifies whether the SIC operation is carried out at user  $i$  to decode the signal of user  $k$ . Specifically,  $\alpha_{m,k} = 1$  indicates that user  $m$  will first employ the SIC to decode the signal of user  $k$  before decoding its own signal for eliminating interference from user  $k$ , and  $\alpha_{m,k} = 0$  otherwise. As it is generally impossible to mutually implement the SIC decoding at both users, we have

$$\alpha_{m,k} + \alpha_{k,m} \leq 1, \quad \forall m, k \in \mathcal{K}, m \neq k. \quad (2)$$

Given  $\alpha$  and the corresponding ascending-channel-gain decoding order (such decoding order is shown to be near-optimal in [28]), user  $m$  will sequentially decode the signals of each user  $k$  that satisfies  $\alpha_{m,k} = 1$ . Once user  $k$ 's signal is decoded, user  $m$  can remove the interference from user  $k$  when decoding the remaining users' signals. In other words, the users in the cluster-free NOMA framework are able to decode and subtract the interference of all weaker users. When  $\alpha_{m,k} = 1$ , to successfully implement SIC for interference elimination, the following SIC decoding constraint should be satisfied

$$R_{m \rightarrow k} \geq \alpha_{m,k} R_{k \rightarrow k}, \quad \forall m, k \in \mathcal{K}, m \neq k \quad (3)$$

where  $R_{m \rightarrow k}$  is the achievable rate for user  $m$  to decode user  $k$ 's signal, and  $R_{k \rightarrow k}$  denotes the achievable rate for user  $k$  to decode its own signal. The SIC decoding constraint indicates that the achievable rate  $R_{m \rightarrow k}$  for decoding user  $k$ 's signal

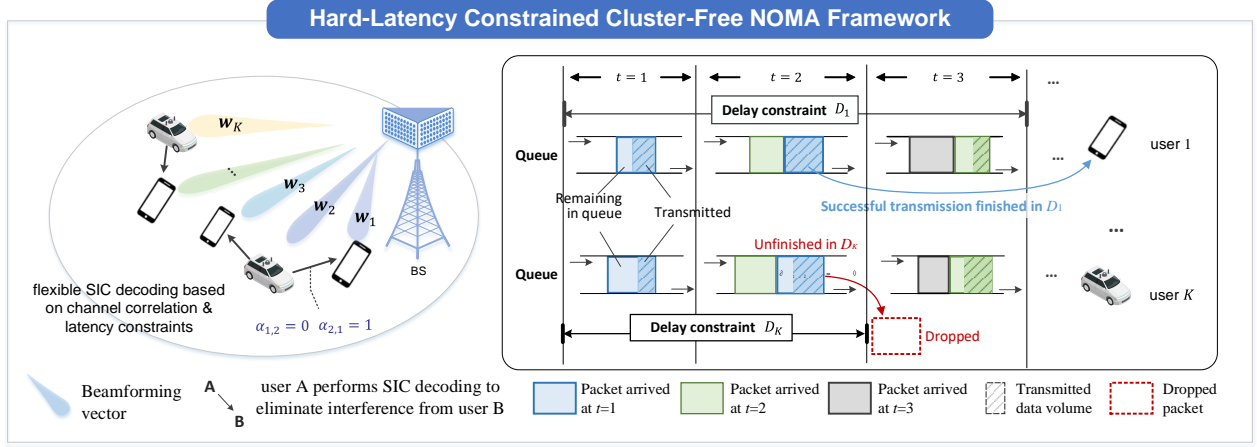


Figure 1. The cluster-free NOMA system and illustration of the queue dynamic model.

at user  $m$  should be maintained at a sufficiently high level to successfully eliminate the interfering signal from user  $k$ .

The signal-to-interference-plus-noise ratio (SINR)  $\text{SINR}_{k \rightarrow k}$  for user  $k$  to decode its own signal can be given by

$$\text{SINR}_{k \rightarrow k} = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{m \neq k} (1 - \alpha_{k,m}) |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma^2}, \quad \forall k \in \mathcal{K}. \quad (4)$$

For ease of expression, we sort users according to the ascending-channel-gain ordering method, which is a commonly adopted ordering method and has been demonstrated to be a rational and effective predefined decoding order [29], i.e.,  $\|\mathbf{h}_m\|^2 \leq \|\mathbf{h}_k\|^2, \forall m < k$ . Then the SINR for user  $m$  to decode user  $k$ 's signal  $\text{SINR}_{m \rightarrow k}$  can be given by

$$\begin{aligned} \text{SINR}_{m \rightarrow k} &= |\mathbf{h}_m^H \mathbf{w}_k|^2 \left( \sum_{u < k} (1 - \alpha_{m,u} + \alpha_{m,u} \alpha_{u,k}) \right. \\ &\quad \left. |\mathbf{h}_m^H \mathbf{w}_u|^2 + \sum_{u > k} (1 - \alpha_{m,u} \alpha_{k,u}) |\mathbf{h}_m^H \mathbf{w}_u|^2 \right. \\ &\quad \left. + \sigma^2 \right)^{-1}, \quad \forall m, k \in \mathcal{K}, m \neq k. \end{aligned} \quad (5)$$

As a result, the achievable data rate  $R_{k \rightarrow k}$  for user  $k$  to decode its own signal can be expressed as  $R_{k \rightarrow k} = \log_2(1 + \text{SINR}_{k \rightarrow k}), \forall k \in \mathcal{K}$ . The achievable rate  $R_{m \rightarrow k}$  for decoding user  $k$ 's signal at user  $m$  can be given by  $R_{m \rightarrow k} = \log_2(1 + \text{SINR}_{m \rightarrow k}), \forall m, k \in \mathcal{K}, m \neq k$ .

### C. Traffic and Queue Dynamic Model

We assume that the data packets only arrive at the start of each timeslot. Specifically, at the  $t$ -th timeslot, a batch of  $\varsigma_{i,t}$  data packets  $\{\omega_{i,t}^b\}$  of length  $Q_{i,t}^b$  arrives at the queue of user  $i$  with a probability  $PA_i$ . The length of arrived data is random with  $\mathbb{E}(Q_{i,t}^b) = \lambda_i$ .

The delay constraint for user  $i$  is  $D_i$ , which means that if a packet arrives at user  $i$ 's queue at the  $t$ -th timeslot, and at the  $(t + D_i)$ -th timeslot it has not been successfully delivered, then it would be dropped out of the queue at this timeslot. Apparently, there are at most  $D_i$  batches of packets in the

queue of user  $i$ . To better capture the state of each packet in the queue, we denote  $Q_{i,t}^b$  as the remaining data size of the packet  $\omega_{i,t}^b$ , and  $B(\omega_{i,t}^b) = \sum_{t'=1}^{D_i-1} \sum_{b'=1}^{\varsigma_{i,t-t'}} Q_{i,t-t'}^{b'} + \sum_{b'=1}^{t-1} Q_{i,t}^{b'}$  as the length of the packet backlog in front of  $\omega_{i,t}^b$ . The arrived data packets are served according to the first-come-first-served (FCFS) protocol. Thus, the packet  $\omega_{i,t}^b$  will not be served until  $B(\omega_{i,t}^b) = 0$ . In hard-latency constrained transmissions, we focus on the following two crucial cases:

- **Packet being dropped:** Packets failed to be delivered before their deadlines would be dropped. Specifically, at the  $t$ -th time slot, the packet  $\omega_{i,t-D_i}^b$  in the queue of user  $i$  would be dropped if  $Q_{i,t-D_i}^b > R_i(t-1)\tau$ .
- **Packet being successfully delivered:** We define a binary functions  $\mathbb{I}_{i,t}(\omega_{i,t'}^b)$ ,  $\forall i, t$ , to indicate whether the packet  $\omega_{i,t'}^b$  is successfully delivered at the  $t$ -th time slot:

$$\mathbb{I}_{i,t}(\omega_{i,t'}^b) = \begin{cases} 1, & \text{if } B(\omega_{i,t'}^b) + Q_{i,t'}^b \leq R_i(t)\tau, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

For ease of understanding, we show a possible state of the data queues in Fig. 1, and we assume  $\varsigma_{i,t} = 1$ . As shown in Fig. 1, packet  $\omega_{1,1}^1$  arrived at user 1's queue at the 1st timeslots ago, and at the 3rd timeslot, packet  $\omega_{1,1}^1$  has been successfully delivered to user 1. At the 2nd timeslot, user  $K$ 's packet  $\omega_{K,1}^1$  has not been delivered before delay constraint so it is dropped.

### D. Problem Formulation

In this paper, we focus on optimizing the hard-latency constrained effective throughput (HLC-ET) of users. Specifically, at the  $t$ -th time slot, the instantaneous HLC-ET of users can be defined as

$$\frac{1}{\tau} \sum_{i=1}^K \sum_{t'=0}^{D_i-1} \sum_{b'=1}^{\varsigma_{i,t-t'}} \mathbb{I}_{i,t}(\omega_{i,t-t'}^{b'}) \bar{Q}_{i,t-t'}^{b'}, \quad (7)$$

where  $\tau$  is the duration of each timeslot,  $\omega_{i,t-t'}^{b'}$  is the  $b'$ -th packet in the batch that arrived at user  $i$ 's queue at  $(t-t')$ -th timeslot, and  $\bar{Q}_{i,t-t'}^{b'}$  is the original length of packet  $\omega_{i,t-t'}^{b'}$ . The delay constraint of user  $i$  is  $D_i$ , so at the  $t$ -th timeslot, the oldest packet in user  $i$ 's queue is  $\omega_{i,t-(D_i-1)}^1$ . The indication

function  $\mathbb{I}_{i,t}(\omega_{i,t-t'}^{b'}) = 1$  means that  $\omega_{i,t-t'}^{b'}$  with the original packet length  $\bar{Q}_{i,t-t'}^{b'}$  is successfully delivered at the  $t$ -th timeslot, which contributes to the HLC-ET with the term  $\frac{1}{\tau} \bar{Q}_{i,t-t'}^{b'}$ . Without loss of generality, we set  $\tau = 1$ . When packet  $\omega_{i,t-t'}^{b'}$  is delivered successfully at the  $t$ -th time slot, the original data size  $\bar{Q}_{i,t-t'}^{b'}$  would be included in the HLC-ET. To simplify the presentation, we define vectors  $\mathbf{Q}(t) = [\{Q_{1,t-D_1+1}^b, \dots, \{Q_{1,t}^b, \dots, \{Q_{K,t-D_K+1}^b, \dots, \{Q_{K,t}^b\}^T \in \mathbb{R}^{\sum_{i=1}^K \varsigma_{i,t} D_i}$  and  $\bar{\mathbf{Q}}(t) = [\{\bar{Q}_{1,t-D_1+1}^b, \dots, \{\bar{Q}_{1,t}^b, \dots, \{\bar{Q}_{K,t-D_K+1}^b, \dots, \{\bar{Q}_{K,t}^b\}^T \in \mathbb{R}^{\sum_{i=1}^K \varsigma_{i,t} D_i}$ .

Our objective is to maximize the HLC-ET by jointly designing the beamforming and SIC operation:

$$\max_{\mathbf{W}, \boldsymbol{\alpha}} \sum_{i=1}^K \sum_{t'=0}^{D_i-1} \sum_{b'=1}^{\varsigma_{i,t-t'}} \mathbb{I}_{i,t}(\omega_{i,t-t'}^{b'}) \cdot \bar{Q}_{i,t-t'}^{b'} \quad (8a)$$

$$\text{s.t. } \alpha_{m,k} \in \{0, 1\}, \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (8b)$$

$$\alpha_{m,k} + \alpha_{k,m} \leq 1, \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (8c)$$

$$\sum_k \|\mathbf{w}_k\|_2^2 \leq P_{\max}, \quad (8d)$$

$$R_{m \rightarrow k}(\boldsymbol{\alpha}, \mathcal{K}) \geq \alpha_{m,k} R_{k \rightarrow k}(\boldsymbol{\alpha}, \mathcal{K}), \quad \forall m, k \in \mathcal{K}, m \neq k. \quad (8e)$$

However, it is hard to converge if we directly use the beamforming and SIC operation as the action and apply a RL algorithm, due to the large action space and the complicated constraints. To address this issue, we proposed a two-stage RL framework, with details to be presented in Section III.

### III. TWO-STAGE RL FRAMEWORK

#### A. Motivation and Outline of Two-Stage RL Framework

Some previous works applying the RL algorithm to solve user clustering and SIC operation problem have chosen discrete actions, e.g., the author in [11] adopted deep Q-network (DQN) to solve the mean error minimization problem to handle URLLC constraints in a NOMA-aided uplink URLLC system with short data blocks. However, since the values of the state, e.g., the channel state information (CSI), are continuous, applying discrete actions may degrade the performance of RL. On the other hand, the action space would be extremely large when all of the variables are considered in the action.

In this paper, we design a two-stage RL algorithm which controls the *priority weights* of users in the long-term stage based on a hybrid RL algorithm, and optimizes the beamforming and the SIC operation in the short-term stage based on maximizing the WSR using an iterative algorithm. The WSR maximization algorithm in the short-term stage is viewed as part of the environment for the hybrid RL algorithm in the long-term stage. As illustrated in Fig. 2, at the  $t$ -th iteration, in the *short-term* stage, based on the action  $\mathbf{a}_t = \{\beta_t\}$  generated by the agent, the environment aims to maximize the WSR; In the *long-term* stage, the agent's policy is updated according to past experiences to maximize the long-term HLC-ET, achieving the hard-latency constraint. Such a two-stage RL formulation can significantly reduce the action space and

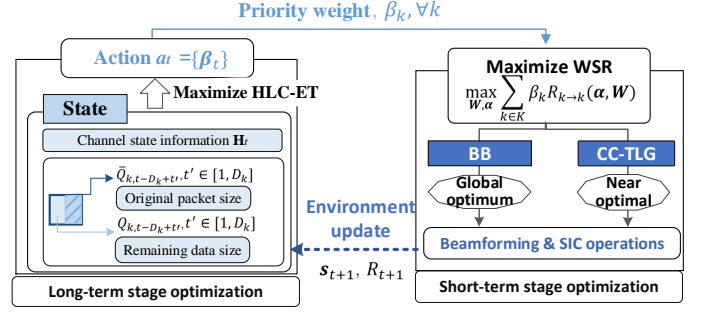


Figure 2. Illustration of the two-stage RL algorithm.

the resulting RL algorithm converges much faster than the conventional single-stage RL formulation that controls all of the variables (i.e., the beamforming and the SIC operation) directly based on RL. Moreover, different from the common one-timescale RL algorithm which can only consider long-term constraints, the WSR optimization algorithm in the short-term stage guarantees the instantaneous constraints in each time slot. In the following, we first formulate the WSR maximization problem (optimization problem for the beamforming and SIC operation) in the short-term stage as a MINLP problem. Then we formulate the problem of controlling the priority weight as a MDP. Finally, we discuss the optimality of two-stage RL problem formulation compared to the conventional single-stage RL formulation.

#### B. Weighted Sum-Rate Maximization Problem in the Short-term Stage

In each iteration, the beamforming and SIC operation are indirectly determined by maximizing the WSR based on a given priority weight  $\beta_t$  and we denote variable  $\boldsymbol{\alpha} = \{\alpha_{mk}\}_{\forall m,k \in \mathcal{K}}$ . Suppose that at the  $t$ -th time slot, the priority weight vector is obtained by the policy based on the current state, then the joint design of the beamforming and SIC operation is obtained by maximizing the WSR of users under practical constraints, which can be formulated as a MINLP problem:

$$\mathcal{P}_0 : \max_{\mathbf{W}, \boldsymbol{\alpha}} \sum_{k \in \mathcal{K}} \beta_k R_{k \rightarrow k}(\boldsymbol{\alpha}, \mathbf{W}) \quad (9a)$$

$$\text{s.t. } \alpha_{m,k} \in \{0, 1\}, \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (9b)$$

$$\alpha_{m,k} + \alpha_{k,m} \leq 1, \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (9c)$$

$$\sum_k \|\mathbf{w}_k\|_2^2 \leq P_{\max} \quad (9d)$$

$$R_{m \rightarrow k}(\boldsymbol{\alpha}, \mathcal{K}) \geq \alpha_{m,k} R_{k \rightarrow k}(\boldsymbol{\alpha}, \mathcal{K}), \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (9e)$$

where Constraint (9b) indicates the binary variable constraint. (9c) indicates that user  $m$  and user  $k$ ,  $m \neq k$ , cannot mutually implement the SIC decoding<sup>1</sup> and Constraint (9d) is the total transmission power constraint. Constraint (9e) represents the SIC decoding conditions. Note that  $\{\beta_k\}$  satisfies that  $\sum_{k \in \mathcal{K}} \beta_k = 1$ .

<sup>1</sup>Considering the fact that each user would always decode its own signal, we directly set  $\text{diag}(\boldsymbol{\alpha}) = \{\alpha_{i,i}\} = \mathbf{1}_{K \times 1}$ .

However, it is challenging to solve  $\mathcal{P}_0$  because of several reasons. Firstly, the design of SIC operation introduces the binary constraint. Secondly, the variables are highly coupled with each other in the objective function as well as rate terms. Therefore,  $\mathcal{P}_0$  is a non-convex and highly coupled MINLP problem that is NP-hard.

### C. MDP Problem Formulation in the Long-term Stage

A MDP is denoted as a tuple  $(\mathcal{S}, \mathcal{A}, R, P)$ , where  $\mathcal{S}$  is the state space,  $\mathcal{A}$  is the action space,  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function.  $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  is the transition probability function, and  $P(s' | s, a)$  denotes the transition probability from state  $s$  to state  $s'$  under action  $a$ . A policy  $\pi : \mathcal{S} \rightarrow \mathbf{P}(\mathcal{A})$  is a map from states to probability distributions over actions, and  $\pi(a | s)$  denotes the probability of choosing action  $a$  in state  $s$ . Due to the curse of dimensionality, modern RL algorithms, e.g., deep reinforcement learning (DRL)-based algorithms, usually parameterize the policy by function approximations with high representation capability, e.g., DNN. In this paper, we denote  $\pi_\theta$  as the policy parameterized by  $\theta$ .

- **State space  $\mathcal{S}$ :**  $\mathcal{S}$  is a composite space consisting of the queue state space and the channel state space, i.e., the current state information at the  $t$ -th time slot is denoted as  $\mathbf{s}_t = \{\mathbf{Q}(t), \bar{\mathbf{Q}}(t), \mathbf{H}(t)\}$ , where  $\mathbf{H}(t) \in \mathbb{C}^{K \times N_T}$  is channel matrix formed by merging the channels of all users.
- **Action space  $\mathcal{A}$ :** the priority weight vector space constitute the action space  $\mathcal{A}$ , i.e., the action at the  $t$ -th time slot is  $\mathbf{a}_t = \{\beta_t\}$ , where  $\beta_t$  is the priority weight vector. Specifically, the action  $\mathbf{a}_t$  is sampled according to a policy  $\pi_\theta : \mathcal{S} \rightarrow \mathbf{P}(\mathcal{A})$ .
- **Transition probability function  $P$ :** the function  $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  is an unknown transition probability function related to the statistics of the unknown statistics of environment model, where  $P(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$  denotes the probability of transition to state  $\mathbf{s}_{t+1}$  from state  $\mathbf{s}_t \in \mathcal{S}$  with an action  $\mathbf{a}_t$ . The transition probability  $P$  and policy  $\pi_\theta$  together determine the probability distribution of the trajectory  $\{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \dots\}$ .
- **Reward function  $R$ :** at each timeslot  $t$ , the instantaneous HLC-ET of users is set to be the reward, i.e.,  $R(\mathbf{s}_t, \mathbf{a}_t) = \sum_{i=1}^K \sum_{t'=0}^{D_i-1} \mathbb{I}_{i,t}(\omega_{i,t-t'}) \cdot \bar{Q}_{i,t-t'}$ .

The objective of long-term optimization stage is to maximize the long-term hard-latency constrained effective throughput, by optimizing the parameter of the DNN policy:

$$\min_{\theta \in \Theta} J(\theta) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{p_s \sim \pi_\theta} \left[ - \sum_{t=0}^{T-1} \sum_{i=1}^K \sum_{t'=0}^{D_i-1} \sum_{b'=1}^{\varsigma_{i,t-t'}} \mathbb{I}_{i,t}(\omega_{i,t-t'}^{b'}) \right] \quad (10)$$

$$\bar{Q}_{i,t-t'}^{b'}], \quad (11)$$

where  $p_s \sim \pi_\theta$  denote the probability distribution of the trajectory under policy  $\pi_\theta$ . Note that there is no need to add an explicit constraint for the probability of violating the hard delay constraint due to the following reason. When all of the packets have the same size and delay constraint  $D_{max}$ , the average HLC-ET is equal to the product of the packet

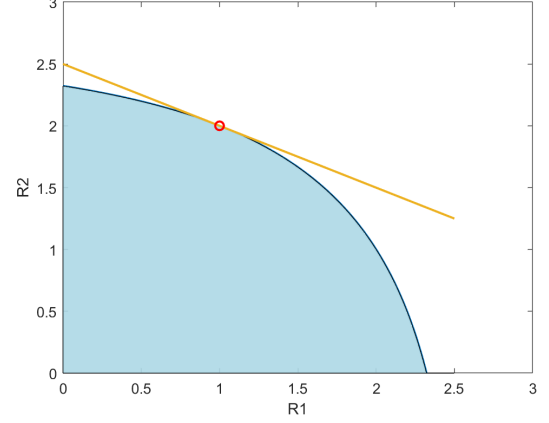


Figure 3. Illustration of a strongly convex rate region.

arrival rate and the successful transmission probability, i.e.  $A * (1 - \Pr(D > D_{max}))$ , where  $A$  is the packet arrival rate. Therefore, maximizing the HLC-ET is equivalent to minimizing the probability of violating the hard delay constraint  $\Pr(D > D_{max})$ .

By such a two-timescale design, the RL algorithm only needs to consider the simple continuous action and optimize a MDP problem with a small action space. Otherwise, the common one-timescale RL algorithms has a extremely large action space (considering all variables), and cannot obtain any reward until the action satisfies the strict constraints in each iteration, making it almost impossible to converge.

### D. Optimality of the Two-stage RL Problem Formulation

When the data rate region is strongly convex, using the priority weights as the control action and maximizing the WSR will not lose any optimality compared to directly controlling all of the variables. The reason is that because if the rate region is strongly convex, with a given weight, the unique WSR maximization rate point is the tangent point of the plane determined by the weight and the rate region, as shown in Fig. 3. This means that any Pareto rate point on the boundary of the rate region can be achieved by maximizing the WSR with a proper weight vector. Since the the optimal solution for maximizing the average effective throughput must achieve a certain Pareto rate point of the rate region (otherwise, we can find a better solution that achieves a strictly higher rate vector to further improve the effective throughput), and we can always achieve the same Pareto rate point by controlling the weight vector, directly controlling the weight vector will not lose any optimality.

It is well-known that the capacity region of Gaussian MIMO broadcast channel (BC) is strongly convex under a total power constraint. Thus directly controlling the priority weights will not lose any optimality for capacity achieving physical layer schemes (such as dirty paper coding [30]). Simulations show that directly controlling the priority weights is still very efficient under other sub-optimal but more practical physical layer scheme such as NOMA beamforming, even when the rate region is not strongly convex in this case. As such, we adopt

this method in our framework since it significantly reduce the action space and can achieve a near optimal solution.

#### IV. THE PROPOSED TWO-STAGE REINFORCEMENT LEARNING BASED JOINT BEAMFORMING AND SIC OPTIMIZATION ALGORITHM

In this section, we first outline the two-stage RL-based algorithm. Then, the BB-based algorithm in the short-term stage is proposed. Finally, we illustrate the hybrid RL algorithm in the long-term stage.

##### A. Outline of the Two-Stage RL based Algorithm

The proposed TSRL-JBSO algorithm chooses the priority weights as the action to maximize the HLC-ET of users, breaking the entire optimization problem into two timescales. In the long-term stage, the priority weights are controlled by the RL algorithm to achieve the hard latency constraints by solving the MDP problem (11), which aims to maximize the long term HLC-ET of users. The short-term stage (i.e., each iteration of the long-term stage) aims to solve the MINLP problem (9a), where the objective is to maximize the WSR of users based on the given priority weight. For the RL algorithm in the long-term stage, the WSR maximization in the short-term stage can be viewed as part of the environment. The details of the proposed algorithm in the short-term stage and the long-term stage would be illustrated in the following.

##### B. The BB-based Algorithm in the Short-term Stage

During each iteration, with a given priority weight, the optimization of the beamforming and SIC operation under practical constraints is a non-convex MINLP which is NP-hard. In particular, the application of the BB method to the MINLPs has shown promising results [31]. BB is a systematic method to solve non-convex optimization problems, and can be applied to Problem (9a) by constructing and solving its convex relaxation and branching the feasible space successfully. As such, the key challenge to design the BB-based algorithm is to find a proper convex relaxation for the considered MINLP  $\mathcal{P}_0$ , as elaborated below.

1) *Convex Relaxation of  $\mathcal{P}_0$* : To deal with the non-convex data rate expression, we introduce a series of auxiliary variables  $\mathbf{S} = \{S_{mk}\}_{\forall m,k \in \mathcal{K}}$ ,  $\mathbf{I} = \{I_{mk}\}_{\forall m,k \in \mathcal{K}}$  and  $\mathbf{r} = \{r_{mk}\}_{\forall m,k \in \mathcal{K}}$ . Specifically,  $S_{mk}$  and  $r_{mk}$  indicate the lower bounds of the effective gain and the achievable rate for decoding user  $k$ 's signal at user  $m$ ,  $\forall m, k \in \mathcal{K}$ , respectively.  $I_{mk}$  is the upper bound of the interference for decoding user  $k$ 's signal at user  $m$ ,  $\forall m, k \in \mathcal{K}$ . Therefore,  $\mathcal{P}_0$  can be rewritten as:

$$\mathcal{P}_1 : \max_{\mathbf{W}, \mathbf{\alpha}, \mathbf{S}, \mathbf{I}, \mathbf{r}} \sum_{k \in \mathcal{K}} \beta_k r_{kk} \quad (12a)$$

$$\text{s.t. (9b)-(9d)}$$

$$r_{mk} \leq \log_2 \left( 1 + \frac{S_{mk}}{I_{mk}} \right) \quad \forall m, k \in \mathcal{K}, \quad (12b)$$

$$S_{mk} \leq |\mathbf{h}_m^H \mathbf{w}_k|^2 \quad \forall m, k \in \mathcal{K}, \quad (12c)$$

$$\sum_{i \neq k} (1 - \alpha_{k,i}) |\mathbf{h}_k^H \mathbf{w}_i|^2 + \sigma^2 \leq I_{kk} \quad \forall k \in \mathcal{K}, \quad (12d)$$

$$\sum_{u < k} (1 - \alpha_{m,u} + \alpha_{m,u} \alpha_{u,k}) |\mathbf{h}_m^H \mathbf{w}_u|^2 + \sum_{u > k} (1 - \alpha_{m,u} \alpha_{k,u}) |\mathbf{h}_m^H \mathbf{w}_u|^2 + \sigma^2 \leq I_{mk} \quad (12e)$$

$$\forall m, k \in \mathcal{K}, m \neq k,$$

$$r_{mk} \geq \alpha_{m,k} r_{kk}, \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (12f)$$

It has been proved that Problems  $\mathcal{P}_0$  and  $\mathcal{P}_1$  are equivalent in the sense that they have equivalent optimal solutions [5].

We further introduce auxiliary variables  $\gamma = \{\gamma_{mk}\}_{\forall m,k \in \mathcal{K}}$ ,  $\mathbf{U} = \{u_{mk}\}_{\forall m,k \in \mathcal{K}}$  and  $\mathbf{L} = \{l_{mk}\}_{\forall m,k \in \mathcal{K}}$ , such that

$$\gamma_{mk} = \mathbf{h}_m^H \mathbf{w}_k, \quad (13)$$

$$|\gamma_{mk}|^2 \leq u_{mk}, \quad (14)$$

$$|\gamma_{mk}| \geq l_{mk}. \quad (15)$$

Without loss of optimality, Problem  $\mathcal{P}_1$  can be reformulated into a more tractable form as given below

$$\mathcal{P}_2 : \min_{\mathbf{W}, \mathbf{\alpha}, \mathbf{S}, \mathbf{I}, \mathbf{r}, \gamma, \mathbf{U}, \mathbf{L}} \sum_{k \in \mathcal{K}} \beta_k r_{kk} \quad (16a)$$

$$\text{s.t. (9b)-(9d), (12b), (12f), (13)-(15),}$$

$$S_{mk} \leq l_{mk}^2 \quad \forall m, k \in \mathcal{K}, \quad (16b)$$

$$\sum_{i \neq k} (1 - \alpha_{k,i}) u_{ki} + \sigma^2 \leq I_{kk} \quad \forall k \in \mathcal{K}, \quad (16c)$$

$$\sum_{u < k} (1 - \alpha_{m,u} + \alpha_{m,u} \alpha_{u,k}) u_{mu} + \sum_{u > k} (1 - \alpha_{m,u} \alpha_{k,u}) u_{mu} + \sigma^2 \leq I_{mk} \quad (16d)$$

$$\forall m, k \in \mathcal{K}, m \neq k,$$

Note that in  $\mathcal{P}_2$ , the term  $\mathbf{h}_m^H \mathbf{w}_k$ ,  $\forall m, k$  in constraints (12c), (12d) and (12e) are replaced with the newly introduced variables  $\gamma_{mk}$ . The feasible region of  $\mathcal{P}_2$  is *non-convex* due to the constraints (9b), (12b), (12f), (15), (16b), (16c) and (16d), while the objective function is *convex*. We construct a convex relaxation of  $\mathcal{P}_2$  by out-approximating its feasible space with a convex set. By doing so, we can simply drop the binary constraints and treat the variables as continuous ones in the range  $[0, 1]$ .

First, we construct convex relaxations for the constraint (15) by applying the following proposition.

**Proposition 1.** Let  $\mathcal{D}_{[\varphi_{mk}, \bar{\varphi}_{mk}]}(l_{mk})$  denote the subset of complex numbers  $\gamma_{mk} = \rho_{mk} e^{j\varphi_{mk}}$ , with amplitude and phase respectively satisfying the inequalities  $\rho_{mk} \geq l_{mk}$  and  $\varphi_{mk} \leq \varphi_{mk} \leq \bar{\varphi}_{mk}$ , where  $l_{mk} \geq 0$  and  $0 \leq \varphi_{mk} \leq \bar{\varphi}_{mk} \leq 2\pi$ . Suppose that  $\bar{\varphi}_{mk} - \varphi_{mk} \leq \pi$ , then the convex envelop of  $\mathcal{D}_{[\varphi_{mk}, \bar{\varphi}_{mk}]}(l_{mk})$  is given by:

$$\begin{aligned} & \text{Conv}(\mathcal{D}_{[\varphi_{mk}, \bar{\varphi}_{mk}]}(l_{mk})) \\ &= \{ \gamma_{mk} \in \mathbb{C} \mid \sin(\varphi_{mk}) \text{Re}(\gamma_{mk}) - \cos(\varphi_{mk}) \text{Im}(\gamma_{mk}) \leq 0, \\ & \quad \sin(\bar{\varphi}_{mk}) \text{Re}(\gamma_{mk}) - \cos(\bar{\varphi}_{mk}) \text{Im}(\gamma_{mk}) \geq 0, \\ & \quad f_{mk} \text{Re}(\gamma_{mk}) + g_{mk} \text{Im}(\gamma_{mk}) \geq (f_{mk}^2 + g_{mk}^2) l_{mk} \} \end{aligned} \quad (17)$$

where  $f_{mk} = (\cos(\varphi_{mk}) + \cos(\bar{\varphi}_{mk}))/2$  and  $g_{mk} = (\sin(\varphi_{mk}) + \sin(\bar{\varphi}_{mk}))/2$ .

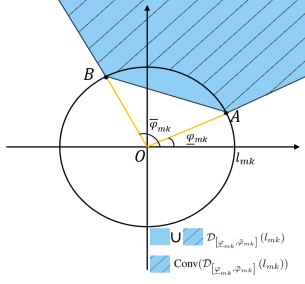


Figure 4. An illustration of  $\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk})$  and  $\text{Conv}(\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk}))$ .

Fig. 4 shows the relationship between sets  $\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk})$  and  $\text{Conv}(\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk}))$ . In Fig. 4, set  $\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk})$  is the light blue region outside the arc AB, which is obviously a non-convex set, and its convex envelope  $\text{Conv}(\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk}))$  is the light blue region, which is convex and is determined by three lines  $\overline{OA}$ ,  $\overline{OB}$  and  $\overline{AB}$ . Proposition 1 has been proved in [32], and it can be verified that as  $\bar{\varphi}_{mk} - \varphi_{mk}$  goes to zero, the convex envelope becomes tight, i.e.,  $\text{Conv}(\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk})) = \mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk})$ . Note that the convex envelope does not take effect when  $\bar{\varphi}_{mk} - \varphi_{mk} > \pi$ . According to Proposition 1, the constraint (15) can be replaced by  $\gamma_{mk} \in \text{Conv}(\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk}))$ .

Next, by defining  $\mu_{mk} = \alpha_{m,k} r_{kk}$ ,  $\psi_{mk} = l_{mk}^2$ ,  $\zeta_{mu} = \alpha_{m,u} u_{mu}$ ,  $\eta_{km}^{u1} = \alpha_{u,k} \zeta_{mu}$  ( $u < k$ ) and  $\eta_{km}^{u2} = \alpha_{k,u} \zeta_{mu}$  ( $u > k$ ), we can reformulate (12f), (16b), (16c) and (16d) as below

$$r_{mk} \geq \mu_{mk}, \quad \forall m, k \in \mathcal{K}, m \neq k, \quad (18)$$

$$S_{mk} \leq \psi_{mk}, \quad \forall m, k \in \mathcal{K}, \quad (19)$$

$$\sum_{i \neq k} (u_{ki} - \zeta_{ki}) + \sigma^2 \leq I_{kk} \quad \forall k \in \mathcal{K}, \quad (20)$$

$$\sum_{u < k} (u_{mu} - \zeta_{mu} + \eta_{km}^{u1}) + \sum_{u > k} (u_{mu} - \zeta_{mu} + \eta_{km}^{u2}) + \sigma^2 \leq I_{mk}, \quad \forall m, k \in \mathcal{K}, m \neq k. \quad (21)$$

We address the relaxation of the bilinear terms defined above using their convex and concave envelopes. According to [31], for three variables  $x$ ,  $y$  and  $z$ , the set  $\mathcal{S}_{(x,y,z)} = \{(x, y, z) \in \mathcal{R}^3 | x \in [x^l, x^u], y \in [y^l, y^u], z = xy\}$  is relaxed using the convex and concave envelopes:

$$\begin{aligned} \text{Conv}(\mathcal{S}_{(x,y,z)}) = \{(x, y, z) \in \mathcal{R}^3 | \\ z \geq x^u y + y^u x - x^u y^u, \\ z \geq x^l y + y^l x - x^l y^l, \\ z \leq x^u y + y^l x - x^u y^l, \\ z \leq x^l y + y^u x - x^l y^u \} \end{aligned} \quad (22)$$

Then, by defining  $\nu_{mk} = x_{mk} I_{mk}$  and introducing a constraint that

$$x_{mk} \leq 2^{r_{mk}} - 1 \quad (23)$$

the constraints (12b) is relaxed as:

$$v_{mk} \leq S_{mk} \quad \forall m, k \in \mathcal{K}, \quad (24)$$

Finally, we can write the convex relaxation of  $\mathcal{P}_2$  as below, where  $\mathbf{Z} = \{\mu, \psi, \zeta, \eta, \mathbf{V}, \mathbf{X}\}$ . Constraint (9b) is eliminated and  $\mathcal{P}_3$  is a convex problem. Note that although Constraint (9b) is eliminated here, and the branch method in Section IV-B2 ensures the binary constraints.

$$\mathcal{P}_3 : \min_{\mathbf{W}, \alpha, \mathbf{S}, \mathbf{I}, \mathbf{r}, \gamma, \mathbf{U}, \mathbf{L}, \mathbf{Z}} - \sum_{k \in \mathcal{K}} \beta_k r_{kk} \quad (25a)$$

$$\text{s.t. (9c)-(9d), (13)-(14), (18)-(21), (23), (24),}$$

$$\gamma_{mk} \in \text{Conv}(\mathcal{D}_{[l_{mk}, \bar{\varphi}_{mk}]}(l_{mk})) \quad (25b)$$

$$(\mu_{mk}, \alpha_{m,k}, r_{kk}) \in \text{Conv}(\mathcal{S}_{(\mu_{mk}, \alpha_{m,k}, r_{kk})}) \quad (25c)$$

$$(\psi_{mk}, l_{m,k}, l_{m,k}) \in \text{Conv}(\mathcal{S}_{(\psi_{mk}, l_{m,k}, l_{m,k})}) \quad (25d)$$

$$(v_{mk}, x_{mk}, I_{mk}) \in \text{Conv}(\mathcal{S}_{(v_{mk}, x_{mk}, I_{mk})}) \quad (25e)$$

$$(\zeta_{mu}, \alpha_{m,u}, u_{mu}) \in \text{Conv}(\mathcal{S}_{(\zeta_{mu}, \alpha_{m,u}, u_{mu})}) \quad (25f)$$

$$(\eta_{km}^{u1}, \alpha_{u,k}, \zeta_{mu}) \in \text{Conv}(\mathcal{S}_{(\eta_{km}^{u1}, \alpha_{u,k}, \zeta_{mu})}) \quad (25g)$$

$$(\eta_{km}^{u2}, \alpha_{k,u}, \zeta_{mu}) \in \text{Conv}(\mathcal{S}_{(\eta_{km}^{u2}, \alpha_{k,u}, \zeta_{mu})}) \quad (25h)$$

2) *BB-based Algorithm*: We adopt the BB-based algorithm in [32] to solve  $\mathcal{P}_2$ , which is equivalent to  $\mathcal{P}_0$ . For ease of notation, we denote the variable matrix of interest as  $\mathbf{Z} = [\alpha, [r_{11}, r_{22}, \dots, r_{KK}]^T]$ , where  $\alpha$  is defined in  $\mathcal{P}_0$ . Initially, the matrix belongs to the box  $\mathcal{B}_0 = [\underline{\Upsilon}, \bar{\Upsilon}]$  where

$$\underline{\Upsilon} = \mathbf{0}_{K \times (K+1)}, \bar{\Upsilon} = [\mathbf{1}_{K \times K}, R_{up} \mathbf{1}_{K \times 1}]. \quad (26)$$

where  $R_{up}$  is an upper bound of the user rate.

The BB-based algorithm involves a sequence of iterations indexed by  $l \in \mathbb{N}$ . In the  $l$ -th iteration,  $\mathcal{L}^l$  denotes the box list, which is the set of boxes with possible feasible regions, and is initialized with  $\mathcal{B}_0$ .  $\Phi_U^l$  and  $\Phi_L^l$  denote an upper bound and a lower bound of the optimal objective function value of  $\mathcal{P}_2$ , respectively. For convenience, we denote  $\Phi_U(\mathcal{B})$  and  $\Phi_L(\mathcal{B})$  as an upper bound and a lower bound of the objective function value of  $\mathcal{P}_2$  over a given box  $\mathcal{B}$ , which are obtained as follows.

- **Lower bound**: Given a box  $\mathcal{B}$ , the lower bound  $\Phi_L(\mathcal{B})$  is obtained by solving the convex relaxation problem  $\mathcal{P}_3$  using any general-purpose solver [33], where the variables  $\alpha$  and  $[r_{11}, r_{22}, \dots, r_{KK}]^T$  are searched over  $\mathcal{B}$ .
- **Upper bound**: Given a box  $\mathcal{B}$ , the upper bound  $\Phi_U(\mathcal{B})$  is a feasible solution of  $\mathcal{P}_2$  that satisfies all constraints, and can be obtained by scaling the optimal solution of  $\mathcal{P}_3$  to meet the constraints in  $\mathcal{P}_2$ . Specifically, we denote  $\{\alpha^*, [r_{11}^*, r_{22}^*, \dots, r_{KK}^*]^T\}$  as the optimal solution of  $\mathcal{P}_3$  over the box  $\mathcal{B}$ . To make the scaling process more tractable, the upper bound is simply set to  $+\infty$  if  $\alpha^* \notin \mathbb{B}^{K \times K}$ . If  $\alpha^* \in \mathbb{B}^{K \times K}$  and constraints (12b), (12f) and (15)-(16d) are satisfied, then the optimal solution of  $\mathcal{P}_3$  is already an upper bound of  $\mathcal{P}_2$ . If these constraints are not satisfied, we can scale  $[r_{11}^*, r_{22}^*, \dots, r_{KK}^*]^T$  to be feasible. In particular, we define the scaling factor for  $\mathbf{L}, \mathbf{S}, \mathbf{I}$  and  $\mathbf{R}$  as  $\kappa^1 \triangleq \{\kappa_{mk}^1, \forall m, k \in \mathcal{K}\}$ ,  $\kappa^2 \triangleq \{\kappa_{mk}^2, \forall m, k \in \mathcal{K}\}$ ,  $\kappa^3 \triangleq \{\kappa_{mk}^3, \forall m, k \in \mathcal{K}\}$  and  $\kappa^4 \triangleq \{\kappa_{kk}^4, \forall k \in \mathcal{K}\}$ , respectively. That is, the scaled solution is given by  $l_{mk} = l_{mk}^* / \kappa_{mk}^1$ ,  $S_{mk} = S_{mk}^* / \kappa_{mk}^2$ ,  $I_{mk} = I_{mk}^* \kappa_{mk}^3$ ,  $r_{kk} = r_{kk}^* / \kappa_{kk}^4$ . The scaling factor  $\kappa^1, \kappa^2, \kappa^3$

and  $\kappa^4$  is chosen such that the scaled solution meets the constraints (12b), (12f) and (15)-(16d):

$$\kappa_{mk}^1 = \max(1, \frac{l_{mk}^*}{|\gamma_{mk}^*|}), \forall m, k \in \mathcal{K}, \quad (27)$$

$$\kappa_{mk}^2 = \max(1, \frac{S_{mk}^*}{l_{mk}^*}), \forall m, k \in \mathcal{K}, \quad (28)$$

$$\kappa_{mk}^3 = \begin{cases} \max((\sum_{u < k} (1 - \alpha_{m,u}^* + \alpha_{m,u}^* \alpha_{u,k}^*) u_{mu}^* + \sum_{u > k} (1 - \alpha_{m,u}^* \alpha_{k,u}^*) u_{mu}^* + \sigma^2) / I_{mk}^*, 1), & \forall m, k \in \mathcal{K}, m \neq k, \\ \max(1, \frac{\sum_{i \neq k} (1 - \alpha_{k,i}^*) u_{ki}^* + \sigma^2}{I_{kk}^*}), & \forall m, k \in \mathcal{K}, m = k, \end{cases} \quad (29)$$

$$\kappa_{kk}^4 = \max(1, \max_{m \neq k} \frac{r_{mk}^* \alpha_{mk}^*}{\log_2(1 + \frac{r_{mk}^*}{I_{mk}^*})}, \frac{r_{kk}^*}{\log_2(1 + \frac{S_{kk}^*}{I_{kk}^*})}), \quad \forall k \in \mathcal{K}. \quad (30)$$

From the scaled feasible solution, we can obtain an upper bound of  $\mathcal{P}_2$  over the box  $\mathcal{B}$  as  $\Phi_U(\mathcal{B}) = -\sum_{k \in \mathcal{K}} \beta_k r_{kk}^*$ .

Now we are ready to present the BB-based algorithm, which consists of two main parts, i.e., Branch and Bound in each iteration, as elaborated below.

- 1) **Branch:** At the  $l$ -th iteration, we select the box with the least lower bound from the box list  $\mathcal{L}^l$ , i.e.,  $\mathcal{B}^* = \arg \min_{\mathcal{B} \in \mathcal{L}^l} \Phi_L(\mathcal{B})$ . Then the selected box  $\mathcal{B}^* = [L, U]$  is split along the longest edge, i.e.,  $i^*, j^* = \arg \max_{i,j} \{u_{i,j} - l_{i,j}\}$  to create two boxes with equal size, which is

$$\mathcal{B}_1^* = \begin{cases} [L, U - \mathbf{J}_{i^*,j^*}], & \text{if } j^* \leq K+1 \\ [L, U - \frac{1}{2}(b_{i^*,j^*} - a_{i^*,j^*})\mathbf{J}_{i^*,j^*}], & \text{else} \end{cases} \quad (31)$$

$$\mathcal{B}_2^* = \begin{cases} [L + \mathbf{J}_{i^*,j^*}, U], & \text{if } j^* \leq K+1 \\ [L + \frac{1}{2}(b_{i^*,j^*} - a_{i^*,j^*})\mathbf{J}_{i^*,j^*}, U], & \text{else} \end{cases} \quad (32)$$

where  $\mathbf{J}_{i^*,j^*}$  is a  $K \times (K+1)$  matrix with  $i^*, j^*$ -th entry equal to 1 and all other entries equal to 0. This branch method makes sure that when the split edge belongs to the variable  $\alpha$ , this edge will be divided into two edges,  $[0,0]$  and  $[1,1]$ , satisfying Constraint (9b). When the iteration number is large enough, some boxes will satisfy the constraint that  $\alpha \in \mathbb{B}^{K \times K}$ .

- 2) **Bound:** The bounding operation consists in computing the upper bound and the lower bound over the newly added box  $\mathcal{B} \in \{\mathcal{B}_1^*, \mathcal{B}_2^*\}$ , and update the lower bound  $\Phi_L^l$  and the upper bound  $\Phi_U^l$ . Specifically, recall that  $\Phi_L(\mathcal{B})$  and  $\Phi_U(\mathcal{B})$  are the lower bound and upper bound over a given box  $\mathcal{B}$ . In the  $l$ -th iteration, we obtain  $\Phi_L(\mathcal{B}_i^*)$ ,  $i = 1, 2$  and  $\Phi_U(\mathcal{B}_i^*)$ ,  $i = 1, 2$ . Note that the lower bound as well as the upper bound are set to  $+\infty$  if the relaxation problem  $\mathcal{P}_3$  is infeasible over the box. After obtaining the lower bounds of two new boxes, we update the box list by removing  $\mathcal{B}^*$  and adding  $\mathcal{B}_1^*$  and  $\mathcal{B}_2^*$  if their lower bounds are not bigger than the current upper bound  $\Phi_U^l$ , i.e.,  $\mathcal{L}^{l+1} = (\mathcal{L}^l - \{\mathcal{B}^*\}) \cup$

---

### Algorithm 1 The BB-based Algorithm

---

**Initialization:** Initialize  $\mathcal{L}^l$  with  $\mathcal{B}_0$ . Find the lower bound  $\Phi_L(\mathcal{B}_0)$  by solving the convex relaxation problem  $\mathcal{P}_3$ , and the upper bound  $\Phi_U(\mathcal{B}_0)$ . Set  $l = 0$ ,  $\Phi_L^0 = \Phi_L(\mathcal{B}_0)$ ,  $\Phi_U^0 = \Phi_U(\mathcal{B}_0)$ , and the tolerance  $\epsilon > 0$ .

**While**  $(\Phi_U^l - \Phi_L^l) / \Phi_L^l > \epsilon$ :

1. **Branch:** Select the box  $\mathcal{B}^*$  with the least lower bound from the box list  $\mathcal{L}^l$ , and split it into two boxes  $\mathcal{B}_1^*$  and  $\mathcal{B}_2^*$ .
2. **Bound:** For each box  $\mathcal{B}_i^*$  ( $i = 1, 2$ ), find its lower bound  $\Phi_L(\mathcal{B}_i^*)$  and its upper bound  $\Phi_U(\mathcal{B}_i^*)$ .
3. Update  $\mathcal{L}^{l+1} = (\mathcal{L}^l - \{\mathcal{B}^*\}) \cup \{\mathcal{B}_i^* | \Phi_L(\mathcal{B}_i^*) \leq \Phi_U^l, i = 1, 2\}$ .
4. Update  $\Phi_L^{l+1} = \min_{\mathcal{B} \in \mathcal{L}^{l+1}} \Phi_L(\mathcal{B})$ .
5. Update  $\Phi_U^{l+1} = \min_{\mathcal{B} \in \mathcal{L}^{l+1}} \Phi_U(\mathcal{B})$ .
6.  $t = t + 1$ .

**End**

---

$\{\mathcal{B}_i^* | \Phi_L(\mathcal{B}_i^*) \leq \Phi_U^l, i = 1, 2\}$ . Then we update the lower bound and upper bound of the optimal objective function value of  $\mathcal{P}_2$ , i.e.,  $\Phi_L^{l+1} = \min_{\mathcal{B} \in \mathcal{L}^{l+1}} \Phi_L(\mathcal{B})$ ,  $\Phi_U^{l+1} = \min_{\mathcal{B} \in \mathcal{L}^{l+1}} \Phi_U(\mathcal{B})$ .

The proposed BB-based algorithm is summarized in **Algorithm 1**.

3) *Convergence and Complexity Analysis:* We denote  $\text{size}(\mathcal{B})$  as the maximum half-length of the sides of box  $\mathcal{B}$ . Following the Theorem 1 in [32], the upper bound and the lower bound of a box region become tight as the box is small enough and shrinks to a point. In other words, as  $\text{size}(\mathcal{B})$  goes to zero, the gap between the lower bound and the upper bound converges to zero.

By adopting the splitting rule in section IV-B2, at least one box in the partition has size not exceeding  $\epsilon$  if  $l$  is sufficiently large. It follows from the Theorem 1 in [32] that when the corresponding box is added to the list at the  $l$ -th iteration, the algorithm should terminate and return  $\epsilon$ -optimal solution.

Following the similar analysis in [32], we can prove that the number of iterations of the BB-based algorithm for obtaining the solution is finite:

**Theorem 1.** For any given  $\epsilon > 0$ , the proposed BB-based algorithm returns an  $\epsilon$ -optimal solution of the given problem within at most  $T_B$  iterations, where

$$T_B = \left( \frac{2R_{\text{up}}}{\epsilon^2} \right)^K + 1. \quad (33)$$

At each iteration, the complexity of the proposed BB-based algorithm is dominated by calculating the lower bounds. Obtaining the lower bound requires solving a convex quadratic problem via a general-purpose solver, e.g., MOSEK in CVX [34] with a complexity of  $\mathcal{O}((K^2)^{3.5})$ . Assuming that the BB-based algorithm converges after  $T_B$  iterations, the worst-case computational complexity can be expressed as  $\mathcal{O}(T_B K^7)$ . Theorem 1 shows that  $T_B$  can be very large if the tolerance  $\epsilon$  is small. Nevertheless, the proposed BB-based algorithm can be used as a performance benchmark.

---

**Algorithm 2** TSRL-JBSO Algorithm
 

---

**Input:** The initial policy parameters  $\theta^0 \in \Theta$ .

**for**  $t = 0, 1, \dots$  **do**

**Long-term Stage Optimization:**

1. Sample an action  $\mathbf{a}_t \sim \pi_{\theta^t}(\cdot | \mathbf{s}_t)$ .

2. **Short-term Stage Optimization:**

(a) Optimize the MINLP problem (9a) based on the action  $\mathbf{a}_t = \{\beta_t\}$ , and obtain the transmitted rate of each user.

(b) Obtain reward and update the environment status.

3. Update the data storage.

4. Update policy parameter  $\theta^{t+1}$  using SCAOPO in [37]. algorithm.

**end for**

---

### C. The Hybrid RL Algorithm in the Long-term Stage

In the long-term stage, we adopt a fast converged hybrid reinforcement learning (HRL) framework to solve the MDP problem (11). In HRL framework, both policy reuse [35] and domain specific knowledge are exploited to accelerate the convergence speed. Specifically, there are  $N \geq 1$  old policies  $\pi_1, \dots, \pi_N$  trained under other similar environments (parameterized by DNNs), a domain knowledge (DK) policy  $\pi_{N+1}$  (e.g., we use the Q-weighted algorithm [36] as the DK policy in the simulations), and the new policy  $\pi_0 \triangleq \pi_{\gamma_0}$  (parameterized by DNN with parameter  $\gamma_0$ ). In each iteration  $t$ , the agent randomly chooses policy  $\pi_m, m \in \{0, 1, \dots, N, N+1\}$  with probability  $p_m$ . Then the agent generates the action  $\mathbf{a}_t$  according to  $\pi_m$  based on the current state  $\mathbf{s}_t$ , interacts with environment and obtains the cost, and updates the data storage. Finally, the data storage is used to update the hybrid policy  $\pi_{\theta}$  with parameters  $\theta = [\mathbf{p}; \gamma_0]$ , where  $\mathbf{p} = [p_0, \dots, p_{N+1}]$ .

The hybrid policy  $\pi_{\theta}$  can be viewed as a stochastic policy whose actions are generated from a mixture distribution  $\pi_{\theta}(\mathbf{a} | \mathbf{s}) = \sum_{m=1}^M p_m \pi_m(\mathbf{a} | \mathbf{s})$ , and the parameters of this stochastic policy are given by  $\theta = [\mathbf{p}; \gamma_0] \in \Theta$ , where  $\gamma_0$  is the parameter of the new policy and  $\mathbf{p}$  is the probability of using each sub-policy. The old policies  $\pi_1, \dots, \pi_N$  and DK policy  $\pi_{N+1}$  help to accelerate the initial convergence speed. Such a stochastic policy can be seen as a generalization of the conventional stochastic policy with only a single sub-distribution, e.g., the Gaussian policy, is a special case when there is only one sub-policy/sub-distribution. As such, all the existing RL algorithms that work for stochastic policy can be directly applied to update the hybrid policy  $\pi_{\theta}$ . In the simulations, we adopt the successive convex approximation based off-policy optimization (SCAOPO) algorithm in [37] to update the hybrid policy  $\pi_{\theta}$ . The SCAOPO enables to reuse old experiences from previous updates, thereby significantly reducing the implementation cost when deployed in the real-world engineering systems that need to online learn the environment. Compared to conventional DRL approaches, the proposed DRL can better adapt to the change of state distribution and maintain a much smaller probability of violating the delay constraints.

The overall two-stage RL based algorithm is summarized in **Algorithm 2**.

### D. Convergence Analysis of the Long-term Stage Algorithm

It has been explained in Section IV-B3 that when the data rate region is strongly convex, using the priority weights as the control action and maximizing the WSR will not lose any optimality compared to directly controlling all of the variables. Therefore, the optimal solution of the proposed two-stage approach is equivalent to the optimal solution of the conventional single-stage approach, when the data rate region is strongly convex. In other words, if the BB-based algorithm finds the optimal solution of the WSR maximization problem in the short-term stage and the hybrid RL algorithm finds the optimal solution of the MDP problem in the long-term stage, the proposed two-stage approach can find the optimal solution of the original problem. Simulations show that the proposed algorithm is still very efficient even when the rate region considered in this paper is not strongly convex.

In Section IV-B3, we have proved that the BB-based algorithm can find an  $\epsilon$ -optimal solution of the short-term problem. Now we focus on the convergence of the hybrid RL algorithm in the long-term stage of the two-stage approach, i.e., the hybrid reinforcement learning framework, which can be viewed as a stochastic policy whose actions are generated from a mixture distribution, and both the probability of using policy  $\mathbf{p}$  and the parameter of new policy  $\gamma_0$  are updated using SCAOPO in [37]. Following the similar convergence analysis as in [37], we can prove that the long-term stage algorithm can converge to a KKT point. If the KKT point found by the hybrid RL algorithm is also the optimal solution of the MDP problem in the long-term stage, then the overall solution found by the proposed two-stage approach is the optimal solution of the original problem, when the data rate region is strongly convex. Of course, since the problem is non-convex, a KKT point is not necessarily optimal. However, this is also true for the conventional single-stage approach, which still cannot guarantee the convergence to the global optimum for non-convex problem. In fact, due to the huge action space and complicated constraints, directly applying the single-stage approach to this problem cannot even converge, as explained above. Simulations show that the two-stage approach has a much better convergence behavior and the KKT point found by it is indeed a good solution compared to the baselines. As such, we adopt this method in our framework since it significantly reduce the action space and can achieve a good solution that satisfies the KKT conditions.

Please refer to [37] for the detailed convergence proof.

## V. PROPOSED LOW COMPLEXITY ALGORITHM FOR THE SHORT-TERM PROBLEM

In this section, we first propose a low complexity two-loop greedy algorithm based on the user channel correlation coefficients, where the outer loop add users one by one based on the WSR in a greedy manner, and the inner loop generates a near-optimal SIC operation based on both the channel correlation coefficients and WSR in a greedy way. Then, the details of the channel correlation based greedy SIC operation algorithm are introduced. Finally, we compare the complexity of the proposed two-loop greedy algorithm with an exhaustive search method.

### A. The Channel Correlation based Two-loop Greedy Algorithm

Although the short-term BB-based algorithm can find the optimal solution of the MINLP problem, it has high complexity, so it is not suitable to be applied in each iteration of the RL algorithm. To accelerate computation on each iteration, we propose a low complexity channel correlation based two-loop greedy (CC-TLG) algorithm to optimize the MINLP with almost no performance loss compared to the BB-based algorithm. CC-TLG maximizes the WSR of users by adding one user at each iteration in a greedy way.

Specially, in each *outer loop* iteration, CC-TLG adds the user which maximizes the WSR of users under the operation constraints to the user set. We define the set of selected users as  $\mathcal{K}_{GI}^q$ , then at the  $q$ -th iteration, the user that would be selected from  $\mathcal{K}/\mathcal{K}_{GI}^q$  can be expressed as

$$k^q = \arg \max_{k \in \mathcal{K}/\mathcal{K}_{GI}^q} \text{WSR}(k \cup \mathcal{K}_{GI}^q, \alpha_{k \cup \mathcal{K}_{GI}^q}^*), \quad (34)$$

where  $\text{WSR}(k \cup \mathcal{K}_{GI}^q, \alpha_{k \cup \mathcal{K}_{GI}^q}^*)$  is the maximum WSR that scheduling the user set  $k \cup \mathcal{K}_{GI}^q$  can achieve while meeting the operation constraints,  $\alpha_{k \cup \mathcal{K}_{GI}^q}^* \in \mathbb{B}^{K \times K}$  is a near-optimal SIC operation generated based on the user channel correlation coefficients in a greedy way, which would be further explained in Section V-B. Theoretically, it needs to go through all  $3^{(|\mathcal{K}_{GI}^q|+1)|\mathcal{K}_{GI}^q|/2}$  possible SIC operations of the user set  $k \cup \mathcal{K}_{GI}^q$  to obtain the maximum WSR, but the search space is very large. Therefore, we adopt a channel correlation (CC) based greedy SIC operation method to reduce the search space among possible SIC operations. Moreover, to reduce the complexity of beamforming optimization, the rate of each user is calculated by assuming the simple RZF precoder with equal power allocation, which is widely used in practical systems and is asymptotically optimal for large  $N_t$  and/or high SNR [38].

CC-TLG adds user iteratively until the WSR does not increase or  $\mathcal{K}/\mathcal{K}_{GI}^q = \emptyset$ . Finally, after obtaining the SIC operation and user selection, the beamforming is optimized using the successive convex approximation (SCA) algorithm in [39] for a better performance.

### B. Channel Correlation based Greedy SIC Operation Method (Inner Loop)

For a given user set  $\mathcal{U}$ , we have to determine the SIC operation between all pairs of users in it, and there are three kinds of SIC operations between user  $i$  and user  $j$ , i.e.,  $\{(\alpha_{i,j}, \alpha_{j,i}) | (\alpha_{i,j}, \alpha_{j,i}) \in \{(0,0), (0,1), (1,0)\}\}$ . Therefore, the total possible number of SIC operations for a user set  $\mathcal{U}$  is  $3^{\frac{|\mathcal{U}|^2 - |\mathcal{U}|}{2}}$ , which is extremely large. In this subsection, a channel correlation based greedy method is proposed to reduce the search space by generating a near-optimal SIC operation in a heuristic way.

In the *inner loop*, for a given user set  $\mathcal{U}$ , a near-optimal SIC operation  $\alpha_{\mathcal{U}}^*$  is generated based on the channel correlation coefficients in a greedy way.

**Step 1:** Calculate the channel correlation coefficients between all pairs of users in the user set, and sort the user

pairs by channel correlation in descending order. Define the sorted user pairs and the corresponding coefficients as  $\Gamma_{\mathcal{U}} = \{(\rho_{i_1, j_1}, i_1, j_1), \dots, (\rho_{i_{|\mathcal{U}|^2 - |\mathcal{U}|}, j_{|\mathcal{U}|^2 - |\mathcal{U}|}}, i_{|\mathcal{U}|^2 - |\mathcal{U}|}, j_{|\mathcal{U}|^2 - |\mathcal{U}|})\}$  satisfying  $\rho_{i_1, j_1} \geq \dots \geq \rho_{i_{|\mathcal{U}|^2 - |\mathcal{U}|}, j_{|\mathcal{U}|^2 - |\mathcal{U}|}}$  and  $\{(\rho_{i,j}, i, j) | 1 \leq i < j \leq |\mathcal{U}|\}$ , where  $\rho_{i,j}$  denotes the channel correlation coefficient between user  $i$  and  $j$ , which can be expressed as

$$\rho_{i,j} = \frac{|\mathbf{h}_i^H \mathbf{h}_j|}{\|\mathbf{h}_i\| \cdot \|\mathbf{h}_j\|}. \quad (35)$$

**Step 2:** Determine the SIC operation sequentially according to the channel correlation order. Specifically, the SIC operation between the user pair  $i_1$  and  $j_1$  with the highest channel correlation coefficient is the first to be determined, and the  $(i_1, j_1)$ -th and the  $(j_1, i_1)$ -th elements of  $\alpha_{\mathcal{U}}^*$  are chosen from  $\{(0,1), (1,0), (0,0)\}$  that could maximize the WSR, then  $\alpha_{\mathcal{U}}^*$  is updated; The SIC operation between the user pair  $i_2$  and  $j_2$  with the second highest channel correlation coefficient is the second to be determined; and so on.

Define  $\alpha_{\mathcal{U}}^\rho$  as the SIC operation in the  $\rho$ -th inner-loop iteration, with  $\alpha_{\mathcal{U}}^0 = \mathbf{I}_{|\mathcal{U}|}$ . Then the update of  $\alpha_{\mathcal{U}}^\rho$  can be expressed as

$$\alpha_{\mathcal{U}}^\rho = \arg \max_{(\alpha = \alpha_{\mathcal{U}}^{\rho-1}, (\alpha_{i_\rho, j_\rho}, \alpha_{j_\rho, i_\rho}) \in \{(0,1), (1,0), (0,0)\})} \sum_{m \in \mathcal{U}} \beta_m \cdot R_{m \rightarrow m}(\alpha, \mathcal{U}) \cdot \mathbb{I}^1(\alpha), \quad (36)$$

where  $R_{m \rightarrow m}$  is the user rate defined in Section II, and  $\mathbb{I}^1(\alpha)$  is the indication function that indicates whether the SIC decoding conditions (9e) are satisfied, i.e.,  $\mathbb{I}^1(\alpha) = 1$  indicates that all SIC decoding conditions are satisfied, and  $\mathbb{I}^1(\alpha) = 0$  otherwise.

The channel correlation based greedy SIC operation method determines the SIC operation iteratively until the WSR does not increase or  $\rho = |\mathcal{U}|$ , and finally obtains  $\alpha_{\mathcal{U}}^*$ . The motivation of such greedy design is that a user pair with higher channel correlation coefficient is more likely to require SIC with  $((\alpha_{i,j}, \alpha_{j,i}) \in \{(0,1), (1,0)\})$  to eliminate the interference, intuitively it is desired to first determine the SIC operation between the user pair with the highest channel correlation coefficient.

The overall two-loop algorithm is summarized in **Algorithm 3**.

### C. Complexity Analysis

For the proposed CC-TLG algorithm, the main computational complexity comes from the calculation of the WSR, which can be expressed as  $\mathcal{O}(K^2 N_t + K^3)$ . The proposed CC-TLG requires at most  $\sum_{i=1}^K (K - i + 1) 3i = \frac{K(K+1)(K+2)}{2}$  calculations of WSR. We choose exhaustive greedy method as the baseline to compare with. Exhaustive greedy algorithm also adds user iteratively in a greedy manner, but for a given user set, it searches all possible SIC operations exhaustively. Exhaustive greedy requires  $\sum_{i=1}^K (K - i + 1) 3^{\frac{i(i-1)}{2}}$  calculations of WSR. As illustrated before, the worst-case computational complexity of the BB-based algorithm can be expressed as  $\mathcal{O}(T_B K^7)$ . Note that  $T_B$  can be very large if the

---

**Algorithm 3** CC-TLG Algorithm
 

---

**Initialization:** Initialize the selected user set  $\mathcal{K}_{GI}^0 = \emptyset$ ,  $R_{total}^0 = 0$ , and the SIC operation matrix  $\alpha^* = \mathbf{I}_K$ .

**for**  $q = 0, 1, \dots$  **do**

**Outer loop**

1. **for** each user  $k \in \mathcal{K} / \mathcal{K}_{GI}^q$  **do**

(a) Obtain the sorted channel coefficients  $\Gamma_{k \cup \mathcal{K}_{GI}^q}$  according to **step 1**.

(b) **Inner loop**

Obtain  $\alpha_{\mathcal{U}}^*$  based on the channel correlation based greedy method according to **step 2**.

2. Select the user  $k^q$  with the maximum WSR.

3. **If**  $\text{WSR}(k^q \cup \mathcal{K}_{GI}^q, \alpha_{k^q \cup \mathcal{K}_{GI}^q}^*) \geq R_{total}^q$ :

Update  $\mathcal{K}_{GI}^{q+1} = \mathcal{K}_{GI}^q \cup k^q$ ,  $\alpha^* = \alpha_{k^q \cup \mathcal{K}_{GI}^q}^*$ ,

and  $R_{total}^{q+1} = \text{WSR}(k^q \cup \mathcal{K}_{GI}^q, \alpha_{k^q \cup \mathcal{K}_{GI}^q}^*)$ .

**until**  $\mathcal{K} / \mathcal{K}_{GI}^q = \emptyset$

**Optimize**  $\mathbf{W}$  using SCA.

---

tolerance  $\epsilon$  is small ( $T_B = \left(\frac{2R_{up}}{\epsilon^2}\right)^K + 1$ ). The complexity for BB-based algorithm is too high for it to converge in the simulation, so we delete the unfeasible region of the first several iterations and choose a small range for the rate, e.g.  $R_{up} = 0.1$ . In this case, the computational complexity for the BB-based algorithm is reduced to  $(200^K + 1)K^7$  when  $\epsilon = 0.01$ , making the simulation for  $K = 3, 4$  able to converge in a limited time. In Table I, we compare the complexity of the proposed CC-TLG algorithm with exhaustive greedy and the BB-based algorithm for different values of  $K$ . It can be observed in the simulations that the CC-TLG can significantly reduce the complexity with almost no performance loss.

## VI. NUMERICAL RESULTS

### A. Simulation Setup

In the simulations, we adopt the commonly used exponential correlation Rayleigh fading channel model [40] to generate the channel of users, which can be given by

$$\mathbf{H} = \tilde{\mathbf{H}}\mathbf{\Lambda}^{-1/2}\mathbf{R}_H^{1/2}, \quad (37)$$

where  $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \dots, \tilde{\mathbf{h}}_K]$  stacks the small-scale fading vectors.  $\mathbf{\Lambda}$  is a diagonal matrix with the diagonal elements being  $\text{diag}(\mathbf{\Lambda}) = [L_1(d_1), \dots, L_K(d_K)]$ , where the large-scale fading  $L_k(d_k)$  is given by the pathloss model  $32.6 + 36.7 \lg d_k$ . 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 a channel realization,  $\mathbf{R}_H$  can be mathematically formulated as

$$\mathbf{R}_H = \begin{bmatrix} 1 & c & \dots & c^{K-1} \\ c^H & 1 & \dots & c^{K-2} \\ \vdots & \vdots & \ddots & \vdots \\ (c^{K-1})^H & (c^{K-2})^H & \dots & 1 \end{bmatrix}, \quad (38)$$

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. For each user  $i$ , data packets whose lengths follow a Poisson arrival distribution with mean  $\lambda_i$

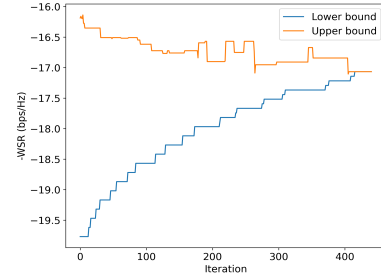


Figure 5. Convergence behavior of the BB-based algorithm for  $K = 3$  under configuration (1).

arrive at the start of each timeslot with probability  $PA$ . We set  $K = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ ,  $N_t = 2$  and the batch of packets  $\varsigma = 1$ . In simulations, we have chosen various values of  $\{D_i\}$ ,  $\{\lambda_i\}$  and  $PA$  according to the parameters reported in [41], which are typical values in burst traffic application scenarios, and the proposed algorithm works under all of those traffic conditions. We report in this section, the simulation results under two configurations: (1)  $D_i \in [4, \dots, 7]$  timeslots,  $\lambda_i \in [15, 25]$  Kbit,  $PA = 0.3$ ; (2)  $D_i \in [4, \dots, 7]$  timeslots,  $\{\lambda_i\} \in [20, 40]$  Kbit,  $PA = 0.3$ .

### B. Simulation Results and Discussions

#### 1) Convergence of the Short-term BB-based Algorithm:

Fig 5 presents the convergence behavior of the proposed BB-based algorithm for  $K = 3$  under configuration (1). It can be observed that the lower bound are non-decreasing and the upper bound are always higher than the lower bound. As the iteration index increases, the gap between the upper bound and the lower bound becomes smaller and infeasible sub-regions are removed. The iteration ends when the gap is smaller than the tolerance  $\epsilon$ .

#### 2) Performance of the Short-term CC-TLG Algorithm:

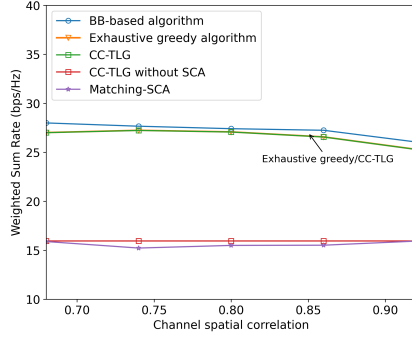
We compare the performance of the proposed CC-TLG algorithm with the BB-based algorithm and exhaustive greedy algorithm, which is introduced in Section V-C. To further demonstrate the performance of the proposed algorithms, we also consider a low complexity matching-SCA algorithm proposed in [5] as a baseline approach, which is shown to guarantee the local optimality in cluster-free NOMA framework. CC-TLG without SCA is the solution (scheduled user set, SIC operation and beamforming) obtained by the proposed low complexity algorithm, but without the final step of optimizing the beamforming by SCA.

Fig 6a shows the achieved WSR versus channel correlations,  $\text{corr}$ , for the case with  $K = 3$  users. It can be observed that the proposed CC-TLG and exhaustive achieve almost the same performance, and obtain a near-optimal solution compare to the benchmark, i.e., the BB-based algorithm. The matching-SCA leads to the worst performance compared to other algorithms.

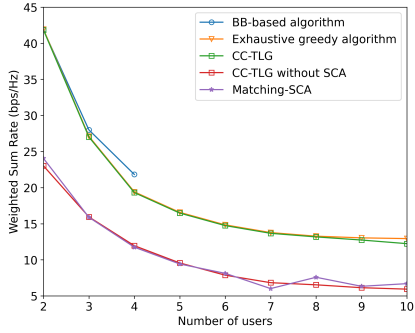
Fig 6b shows the achieved WSR versus number of users,  $K$ , for the case with  $\text{corr} = 0.68$ . Note that we only present the performance of the BB-based algorithm under the case of  $K = \{2, 3, 4\}$ , due to its high computational complexity. It

Table I  
COMPLEXITY COMPARISON

	Computational Complexity	$K = 3$	$K = 4$	$K = 5$
CC-TLG algorithm	$\frac{K(K+1)(K+2)}{2}(K^2 N_t + K^3)$	$1.35e^3$	$5.76e^3$	$1.83e^4$
Exhaustive greedy	$\sum_{i=1}^K (K-i+1)3^{\frac{i(i-1)}{2}}(K^2 N_t + K^3)$	$1.62e^3$	$7.64e^4$	$1.06e^7$
BB-based algorithm	$(200^K + 1)K^7$	$1.74e^{10}$	$2.6e^{13}$	$2.5e^{16}$



(a) Weighted sum rate versus channel correlations  $corr$ .  $K = 3$ .



(b) Weighted sum rate versus number of users  $K$ .  $corr = 0.68$ .

Figure 6. Performance comparisons for different cases.

can be observed that the proposed CC-TLG achieves almost the same performance with exhaustive greedy method, but with dramatically reduced complexity, as presented in Table I. When  $K$  is small, i.e.,  $K = 2, 3$ , the gap between CC-TLG and the BB-based algorithm can be almost ignored. As the number of users increases, the performance loss of the proposed CC-TLG grows, reaches 9.5% when  $K = 4$ . In all cases presented in this section, the proposed CC-TLG outperforms than the matching-SCA and the CC-TLG without SCA, and achieves a near-optimal solution of the problem, with significantly reduced complexity compared to the BB-based algorithm

### 3) Convergence of the Two-Stage RL based Algorithm:

To demonstrate the performance of the proposed TSRL-JBSO algorithm, we choose Q-weighted algorithm [36] as the baseline, which is shown to perform well for light to moderate traffic loading and can provide a stable scheduling performance. Q-weighted JBSO decides the priority weight based on the queue length, then optimizes the MINLP problem using the same method as TSRL-JBSO, while TSRL-JBSO

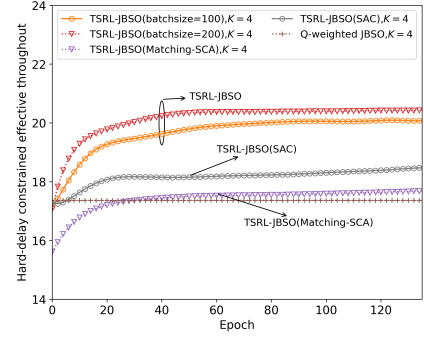
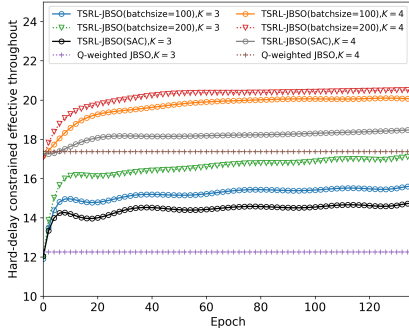


Figure 7. Comparison of different short-term stage methods.

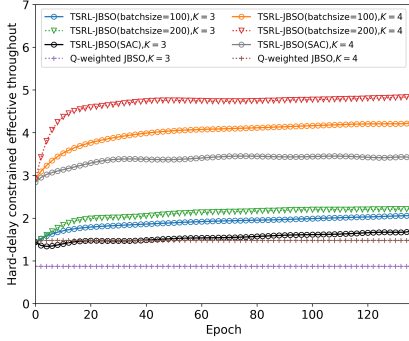
optimizes the policy to generate the weight, whose objective is to maximize the long-term HLC-ET. We also compare the proposed TSRL-JBSO which adopts the HRL algorithm in the long-term stage, with TSRL-JBSO that adopts soft actor-critic (SAC) [42], to illustrate the fast convergence of HRL.

In the simulations, we choose the low-complexity CC-TLG as the short-term stage method since exhaustive greedy and BB-based method have much higher complexity and the simulation time will be unacceptable, while the Matching-SCA proposed in [5] has a worse performance. We compare the proposed algorithm with a two-stage algorithm which chooses the Matching-SCA as the short-term stage method and also adopts the hybrid RL framework to update the policy. As shown in Fig. 7, the baseline with the Matching-SCA as the short-term stage method converges to a stationary point that better than the Q-weighted JBSO, but it still performs much worse than the proposed TSRL-JBSO with the CC-TLG as the short-term stage method.

Fig 8a shows different algorithms' learning curves of HLC-ET under configuration (1), where the traffic loading is moderate. In this paper, batchsize means the number of experience that used to update the policy. It can be observed that when the traffic loading is moderate, Q-weighted JBSO can achieve a fair performance, but much worse than both TSRL-JBSO with 100 batchsize and TSRL-JBSO with 200 batchsize. When the traffic loading goes heavy, as shown in Fig. 8b, the performance gain of the proposed TSRL-JBSO becomes much higher, with the highest gain of 226% (TSRL-JBSO with 200 batchsize compared to the Q-weighted JBSO). In both scenarios, TSRL-JBSO using SAC performs better than Q-weighted JBSO, but converges slower than TSRL-JBSO using HRL. Simulation results also show that with the same old experiences, the proposed algorithm with larger batchsize (newly added data) converges faster than that with smaller batchsize at the cost of higher complexity per iteration, which



(a) The learning curve of hard-latency constrained effective throughput under configuration (1).



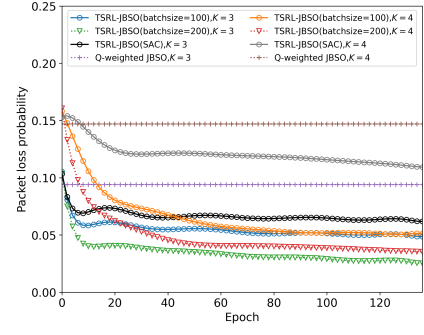
(b) The learning curve of hard-latency constrained effective throughput under configuration (2).

Figure 8. HLC-ET comparison under different traffic scenario.

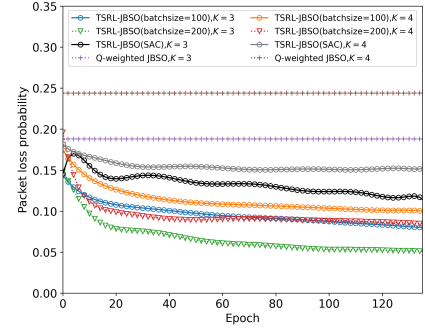
may due to a more stable gradient update of large batchsize.

To better illustrate the effectiveness of the proposed TSRL-JBSO, we also compare the learning curve of packet loss probability under different traffic cases, as shown in Fig. 9a and Fig. 9b. It can be seen that following the same rule of the learning curve of HLC-ET, when the traffic loading is moderate, Q-weighted JBSO can achieve a fair performance, but much worse than the proposed TSRL-JBSO. The performance gain of TSRL-JBSO becomes much higher when the traffic loading is heavy, which illustrate the effectiveness of TSRL-JBSO in burst traffic transmission.

In order to prove the superiority of our two-stage RL algorithm relative to single-stage methods, we simulate a single-stage RL approach to solve the objective problem by trying several commonly used DRL algorithms, such as SAC and SCAPO, to update the parameters of the policy, but found the reward was very small and cannot converge at all. As shown in Fig. 10, we compare the proposed TSRL-JBSO with a single-stage RL approach, whose action is the beamforming and the SIC operations, and the SIC and power constraints are satisfied by projecting the action to the feasible region. It can be found that the reward of the single-stage RL approach is very small and the algorithm converges very slow. This is not only because of the mixed large action space, but also due to the complicated SIC operation constraints. The proposed two-stage algorithm converges significantly faster, since the agent only needs to generate the priority weight, and the problem with complicated constraints can be solved in the short-term



(a) The learning curve of packet loss probability under configuration (1).



(b) The learning curve of packet loss probability under configuration (2).

Figure 9. Packet loss probability comparison under different traffic scenario.

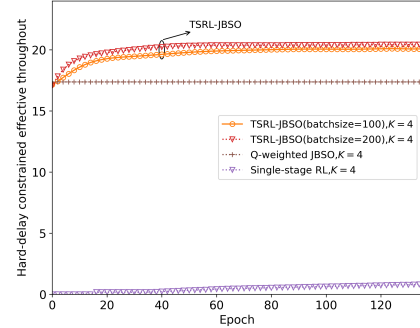


Figure 10. Comparison with the single-stage RL approach.

stage by the designed optimization-based method.

## VII. CONCLUSION

We proposed a novel TSRL-JBSO algorithm, which breaks the entire optimization problem into two stages in different timescales. We developed a BB-based algorithm to obtain the optimal solution of the WSR maximization problem in the short-term stage. We proved that the BB-based algorithm can guarantee the convergence to an  $\epsilon$ -optimal solution of the WSR maximization problem, which belongs to the challenging MINLP, within a finite number of steps. To accelerate computation in the short-term stage, we proposed a low-complexity CC-TLG algorithm based on greed user selection in the outer loop and channel correlation based greedy SIC operation in the inner loop to significantly reduce the complexity with almost no performance loss compared to the BB-based algorithm.

Simulations shows the effectiveness of the proposed CC-TLG in the short-term stage, and that the proposed overall TSRL-JBSO algorithm achieves much better performance than the baseline. In the non-stationary case when the channel/traffic statistics change at a timescale comparable to the timeslot duration, the performance of the proposed algorithm may degrade, which however, is also true for most resource allocation algorithms. Future research may adopt the efficient context-aware meta-learning to address this issue as well as consider imperfect CSI.

## REFERENCES

- [1] 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, Nov. 2017.
- [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 J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Jul. 2017.
- [3] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179–189, Jan. 2020.
- [4] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of NOMA toward next generation multiple access (NGMA) for 6G," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1037–1071, Jan. 2022.
- [5] X. Xu, Y. Liu, X. Mu, Q. Chen, and Z. Ding, "Cluster-free NOMA communications toward next generation multiple access," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2184–2200, Feb. 2023.
- [6] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, pp. 1–74, Nov. 2021.
- [7] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, Oct. 2019.
- [8] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.
- [9] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Network*, vol. 32, no. 2, pp. 78–84, Apr. 2018.
- [10] A. Yu, H. Yang, K. K. Nguyen, J. Zhang, and M. Cheriet, "Burst traffic scheduling for hybrid E/O switching DCN: An error feedback spiking neural network approach," *IEEE Trans. Netw. Serv. Manage.*, vol. 18, no. 1, pp. 882–893, Nov. 2020.
- [11] W. Ahsan, W. Yi, Y. Liu, and A. Nallanathan, "A reliable reinforcement learning for resource allocation in uplink NOMA-URLLC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5989–6002, Jan. 2022.
- [12] Y. Li, C. Dou, Y. Wu, W. Jia, and R. Lu, "Noma assisted two-tier vr content transmission: A tile-based approach for qoe optimization," *IEEE Trans. Mob. Comput.*, vol. 23, no. 5, pp. 3769–3784, May. 2023.
- [13] 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 Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Sep. 2015.
- [14] 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 *2015 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, Jun. 2015, pp. 747–752.
- [15] 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, Aug. 2016.
- [16] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Sep. 2015.
- [17] M. Zeng, A. Yadav, O. A. Dobie, 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, Jul. 2017.
- [18] X. Xu, Q. Chen, X. Mu, Y. Liu, and H. Jiang, "Graph-embedded multi-agent learning for smart reconfigurable THz MIMO-NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 259–275, Nov. 2021.
- [19] Y. Fu, M. Zhang, L. Salaün, 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, Jun. 2020.
- [20] W. Yi, Y. Liu, A. Nallanathan, and M. ElKashlan, "Clustered millimeter-wave networks with non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4350–4364, Feb. 2019.
- [21] C. She, C. Yang, and T. Q. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.
- [22] C. Xiao, J. Zeng, W. Ni, X. Su, R. P. Liu, T. Lv, and J. Wang, "Downlink MIMO-NOMA for ultra-reliable low-latency communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, Feb. 2019.
- [23] R. Kotaba, C. N. Manchon, N. M. K. Pratas, T. Balercia, and P. Popovski, "Improving spectral efficiency in URLLC via NOMA-based retransmissions," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–7.
- [24] D.-D. Tran, S. K. Sharma, V. N. Ha, S. Chatzinotas, and I. Woungang, "Multi-agent DRL approach for energy-efficient resource allocation in URLLC-enabled grant-free NOMA systems," *IEEE Open J. Commun. Soc.*, Jul. 2023.
- [25] J. Chen, J. Wang, C. Jiang, and J. Wang, "Age of incorrect information in semantic communications for NOMA aided XR applications," *IEEE J. Sel. Top. Signal Process.*, vol. 17, no. 5, pp. 1093–1105, Jun. 2023.
- [26] R.-J. Reifert, H. Dahrouj, and A. Sezgin, "Extended reality via cooperative NOMA in hybrid cloud/mobile-edge computing networks," *IEEE Internet Things J.*, Nov. 2023.
- [27] Q. Gao, Y. Liu, X. Mu, M. Jia, D. Li, and L. Hanzo, "Joint location and beamforming design for STAR-RIS assisted NOMA systems," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2532–2546, Feb. 2023.
- [28] Y. Xiu, J. Zhao, W. Sun, M. Di Renzo, G. Gui, Z. Zhang, and N. Wei, "Reconfigurable intelligent surfaces aided mmwave noma: Joint power allocation, phase shifts, and hybrid beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8393–8409, 2021.
- [29] Z. Xiao, L. Zhu, Z. Gao, D. O. Wu, and X.-G. Xia, "User fairness non-orthogonal multiple access (noma) for millimeter-wave communications with analog beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3411–3423, 2019.
- [30] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Aug. 2006.
- [31] M. Tawarmalani and N. V. Sahinidis, *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*. Springer Science & Business Media, 2013, vol. 65.
- [32] S. Norouzi, B. Champagne, and Y. Cai, "Joint optimization framework for user clustering, downlink beamforming, and power allocation in MIMO NOMA systems," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 214–228, Nov. 2022.
- [33] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [34] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.
- [35] F. Fernández and M. Veloso, "Probabilistic policy reuse in a reinforcement learning agent," in *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, May. 2006, pp. 720–727.
- [36] G. Venkatraman, A. Tölli, J. Janhunen, and M. Juntti, "Low complexity multiuser MIMO scheduling for weighted sum rate maximization," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, Nov. 2014, pp. 820–824.
- [37] C. Tian, A. Liu, G. Huang, and W. Luo, "Successive Convex Approximation Based Off-Policy Optimization for Constrained Reinforcement Learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1609–1624, Mar. 2022.
- [38] B. Saleeb, M. Shehata, H. Mostafa, and Y. Fahmy, "Performance evaluation of RZF precoding in multi-user MIMO systems," in *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, Oct. 2019, pp. 1207–1210.
- [39] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations research*, vol. 26, no. 4, pp. 681–683, Jul. 1978.

- [40] 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, Nov. 2002.
- [41] C. Li, B. Liu, X. Su, and X. Xu, "eMBB-URLLC Multiplexing: A Greedy Scheduling Strategy for URLLC Traffic with Multiple Delay Requirements," in *TENCON 2023-2023 IEEE Region 10 Conference (TENCON)*. IEEE, Oct. 2023, pp. 1216–1221.
- [42] H. Tang, A. Wang, F. Xue, J. Yang, and Y. Cao, "A novel hierarchical soft actor-critic algorithm for multi-logistics robots task allocation," *IEEE Access*, vol. 9, pp. 42 568–42 582, Feb. 2021.