Statistical Delay Guarantee for the URLLC in IRS-Assisted NOMA Networks with Finite Blocklength Coding

Thi My Tuyen Nguyen, The Vi Nguyen, Wonjong Noh, and Sungrae Cho

***Abstract*—One of the essential factors for enabling sixth-generation systems is efficiently ensuring diverse quality-of-service (QoS) performance metrics to support the upcoming massive ultra-reliable low-latency communication (URLLC). This work proposes efficient transmission control in intelligent reflecting surface (IRS)-assisted nonorthogonal multiple access (NOMA) networks in the finite blocklength (FBL) regime that statistically guarantee stringent URLLC QoS requirements. Thus, we formulate a nonconvex problem that maximizes the sum effective capacity (SEC) while ensuring statistical delay QoS constraints. To make the problem more tractable, we propose a tight upper bound for the objective function based on Jensen's inequality and employ the concept of opportunistically minimizing an expectation. Then, we decompose the problem into two subproblems: active beamforming at the base station and phase-shift optimization at the IRS. Each subproblem is convexified by employing slack variables, penalty functions, and linear approximation, and solved using successive convex approximations. The subproblems are iteratively solved until convergence using alternating optimization. The convergence to a suboptimal stationary solution and the computing complexity of the proposed algorithm are rigorously analyzed. Finally, extensive numerical evaluations confirm that the proposed control in the FBL regime significantly improves the SEC under various QoS parameters compared to existing benchmark schemes. In particular, as the number of antennas and IRS elements increases, the proposed method becomes more efficient than the semi-definite relaxation-based approach in terms of complexity and performance.**

***Index Terms*—Finite blocklength coding, statistical quality-of-service guarantee, ultra-reliable low-latency communication**

I. Introduction

ULTRA-RELIABILITY low latency communication (URLLC) is becoming increasingly critical in upcoming sixth-generation communication systems due to the exponential increase Internet of Things (IoT) devices with strict latency and reliability requirements [1]. Some representative URLLC services envisioned to support mission-critical applications include autonomous driving, telesurgery systems, industrial automation, augmented reality, and virtual reality. In URLLC systems, performance can be improved in three ways: by improving throughput through resource reuse, directly enhancing dependability, and directly reducing latency [2]–[4]. Thus, paradigm-shifting technology, such as nonorthogonal multiple access (NOMA), intelligent reflecting surface (IRS)-enabled systems, and short-packet transmission, can be considered.

Recently, NOMA technology was introduced to boost the system capacity over orthogonal multiple access (OMA) systems with restricted resources [5]. In contrast to OMA methods, NOMA can reduce the transmission latency of URLLC services by enabling multiple users to access the same resource in the time, frequency, and spatial domains. Specifically, power-domain NOMA employs the successive interference cancellation (SIC) technique to decode the desired signals at the receivers. In multi-antenna communications, NOMA can be more efficient in exploiting limited spatial degrees of freedom (DoFs) than space division multiple access (SDMA), which is only applicable in the underloaded and loaded scenarios when sufficient spatial DoFs are exploited to mitigate the inter user-interference [6].

However, the random nature of the propagation environment caused by multipath fading poses a significant challenge to achieving high-reliability URLLC. The IRS is a potential technique to address this challenge for URLLC [7]. Specifically, the IRS is a meta-surface with a massive number of reflecting components. By properly adjusting the phase shift of all elements, the reflected signals can be added constructively to the direct signal from the base station (BS) to improve the received signal power for the intended users. Hence, the signal-to-ratio is significantly enhanced. Even if the direct link between transceivers is blocked or hindered, IRSs can exploit the smart reflection to build a virtual line-of-sight link between transceivers. Therefore, the IRS can reorganize the wireless environment and convert random wireless channels into partially predictable ones. Thus, the integration of IRSs into a communication system improves reliability and reduces packet retransmission and delay. In addition, most earlier efforts optimized the Shannon capacity with assumptions of an infinite blocklength and zero-error probability. If infinite blocklength coding (IFBC) with vanishing error probability is applied for URLLC applications, there are two problems. First, the delay is *underestimated* because an extremely large code-word incurs a large processing and transmission delay, which is prohibitive in URRLC applications. Second, the reliability is

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2022-00156353) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was supported (in part) by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2023R1A2C1003003) *(Corresponding authors: Wonjong Noh and Sungrae Cho.)*

T. M. T. Nguyen, T. V. Nguyen, and S. Cho are with the School of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, South Korea (email: tuyen@uclab.re.kr, tvnguyen@uclab.re.kr, srcho@cau.ac.kr).

W. Noh is with the School of Software, Hallym University, Chuncheon 24252, South Korea (email: wonjong.noh@hallym.ac.kr).

overestimated because error-free transmission cannot be guaranteed for URLLC systems [8] Therefore, practical URLLC systems must use finite blocklength (FBL) transmission [9]. Recent technology or communication protocols that operate under the FBL regime must be studied to ensure URLLC services [10]. Therefore, integrating the IRS and NOMA in the FBL regime has merits for URLLC services.

A. Related Work and Motivation

Owing to the capability to improve spectral efficiency considerably, a NOMA system can be used in conjunction with an FBL code (FBC) to ensure the quality-of-service (QoS) criteria for latency and reliability. In [11], a two-user downlink NOMA system in the FBL regime was studied to optimize the transmission rates and power distribution and investigate the trade-offs between the transmission rate, decoding error probability, and blocklength-based transmission delay. In [12], a deep state-action reward-state-action learning strategy was presented to optimize the uplink resource allocation in a NOMA-aided URLLC system in the FBL regime. In a time-varying network, the authors provided a reliable learning technique to reduce the mean decoding error probability.

The topic of IRS-assisted URLLC systems for short-packet communication is relatively new. Despite the interesting results on phase-shift control in IRS-aided communication, few studies have investigated the performance of the URLLC system in the FBL regime. The performance of the IRS-aided URLLC system in a factory automation setting was analyzed in terms of the average data rate and decoding error probability [13], [14]. In [13], the authors analyzed the performance with various cases, including Rayleigh, Rician, Nakagami-m, and correlated fading channels under a limited channel blocklength. In [14], the performance was assessed in the presence of phase errors due to limited quantization levels and hardware impairments in IRS components. Moreover, optimization with various performance metrics for the IRS-assisted URLLC system has also been considered. For instance, an energy-efficiency maximization problem in a downlink IRS-assisted URLLC system was considered in [15]. In [16], the blocklength allocation and reflecting beamforming were jointly optimized to minimize the total latency for all users while guaranteeing their reliability.

The IRS can assist NOMA in improving communication performance by facilitating the implementation of the NOMA scheme via effectively aligning the direction of the user channel vectors [26]. Due to the potential of the IRS and NOMA, some research has focused on the integration of the IRS and NOMA in URLLC networks [17]–[23]. In [17], the authors proposed a resource allocation strategy for URLLC services in IRS-aided NOMA networks. They aimed to maximize a weighted sum rate by jointly optimizing the reflection coefficients of the IRS and transmission power allocation of the BS. Moreover, the authors of [18] maximized the sum throughput of all users in IRS-assisted NOMA-URLLC networks using the same set of optimization variables. In [19], the authors considered a STAR-IRS-aided uplink NOMA IoT networks with FBL transmission. The authors investigated the resource allocation design aiming to achieve high rate and low error for rate-target and error-target IoT devices, respectively. In [20], the authors proposed a joint optimization design, where power allocation, transmission blocklength, receiving beamforming, IRS reflection, and user pairing optimization are jointly optimized to minimize the maximum decoding error probability. In [21], the authors investigated an IRS-assisted NOMA-based mobile edge computing (MEC) network in the FBL regime. Accordingly, they formulated the energy efficiency maximization problem under the constraints on the codelength and maximum decoding error probability. In [22], the authors proposed a spectral-efficient resource allocation design for a simultaneous transmitting and reflecting (STAR)-IRS-assisted multi-user multiple-input multiple-output (MU-MIMO) downlink (DL) system subject to given QoS requirements in terms of rate, latency, and reliability. In [23], the authors investigated a resource allocation strategy for an IRS-aided full-duplex (FD) NOMA URLLC system. The sum rate maximization problem is formulated under given latency and reliability requirements. However, most current work on optimizing for IRS-aided NOMA-URLLC networks (e.g., [15]–[21]) is based on alternating optimization (AO) and semidefinite relaxation (SDR) methods. Nevertheless, Gaussian randomization after using SDR is not guaranteed to obtain a rank-one matrix and fails to generate a feasible solution in some instances, which results in an impediment to the local optimality of the overall algorithm.

In addition, existing studies (e.g., [11]–[18]) have only considered non-delay-sensitive-based performance metrics, such as spectral efficiency and energy efficiency. In addition, the current performance analysis based on only physical-layer channel models can barely meet the delay-sensitive requirements of URLLC services [27]. For specific delay-sensitive applications, arrival data must be stored in the queue buffer until transmitted. In such cases, the system performance is significantly influenced by the queuing behavior. Moreover, due to the random variations in the wireless channel conditions, statistically guaranteeing delay requirements should be considered to provide a practical model for QoS. Therefore, a new link-layer channel model, the *effective capacity*, was introduced, which models the channel under the statistical delay QoS constraints in terms of the queuing delay violation probability [28]. Specifically, effective capacity determines the maximum constant arrival rate that a service process can support under statistical QoS constraints. The effective capacity with statistical delay QoS guarantee in NOMA systems was studied in [24], [25]. In [24], the authors proposed a low-latency scheme in the FBL regime that combines the advantages of NOMA and time-division multiple access. Two scenarios were considered based on queuing behaviors that address the error-probability minimization and effective capacity maximization problems. In addition, the authors of [25] proposed two dynamic power allocation schemes under the statistical delay QoS guarantee in an uplink NOMA system with paired users. Specifically, they considered the sum effective capacity (SEC) and effective energy efficiency under statistical delay QoS constraints. However, these studies only focused on the effective capacity under the statistical delay

TABLE I
COMPARISON OF THE PROPOSED AND EXISTING STUDIES

Ref.	**URLLC**	**NOMA**	**IRS**	**Multi-antenna BS**	**Imperfect SIC**	**Statistical delay QoS guarantee**	**Performance metric**	**Optimization method**
[11]	✓	✓	✗	✗	✓	✗	Throughput	Closed-form solution
[12]	✓	✓	✗	✗	✗	✗	Average decoding error probability	Reinforcement learning
[13], [14]	✓	✗	✓	✗	✗	✗	Average data rate/ error probability	Performance analysis
[15]	✓	✗	✓	✓	✗	✗	Energy efficiency	AO, SCA, SDR
[16]	✓	✗	✓	✗	✗	✗	Latency	AO, SCA, SDR
[17], [18]	✓	✓	✓	✗	✗	✗	Sum rate	AO, SCA, SDR
[19]	✓	✓	✓	✗	✓	✗	Sum rate	AO, SCA, SDR, penalty-based method
[20]	✓	✓	✓	✓	✗	✗	Maximal decoding error probability	SCA, SDR, Hungarian matching-based method
[21]	✓	✓	✓	✓	✗	✗	Energy efficiency	AO, SDR, user clustering algorithm
[22]	✓	✓	✓	✓	✓	✗	Sum rate	AO,SCA
[23]	✓	✓	✓	✓	✗	✗	Sum rate	AO, SCA, penalty-based method
[24]	✓	✓	✗	✗	✗	✓	Error probability/Effective capacity	Lagrangian dual method
[25]	✗	✓	✗	✗	✗	✓	Effective capacity	Lagrangian dual method
This work	✓	✓	✓	✓	✓	✓	**Effective capacity**	**AO, SCA, penalty-based method**

AO: alternating optimization algorithm, SCA: successive convex approximation algorithm, and SDR: semidefinite relaxation algorithm

QoS constraints for NOMA-URLLC systems without IRS employment. In addition, the proposed solution is problem-specific, which is challenging to extend to general scenarios, such as multi-antenna and IRS-aided systems, which are addressed in our study. Therefore, an efficient algorithm must be developed for an optimal design in IRS-aided NOMA-URLLC systems under statistical delay QoS constraints.

Moreover, most of the aforementioned works have assumed that SIC can be performed perfectly, which is idealistic and challenging to achieve in practical scenarios. In real-world implementations, signal reception can be affected by various impairments, e.g., fast varying channels, strong channel correlation, and hardware issues, resulting errors in detecting the transmitted symbols. Moreover, because recovery of each symbol using SIC relies on prior decoding, errors will unavoidably propagate and affect the overall system performance. Therefore, considering SIC residual error propagation is important in designing practical NOMA systems [29].

B. Contributions and Organization

To the best of our knowledge, the IRS-aided NOMA system with an FBL for URLLC services is still in its early stages and requires further investigation. The main contributions of this study are summarized as follows. Table I summarizes the differences between this work and existing studies.

- This paper develops a new efficient transmission control that stochastically guarantees URLLC service requirements in IRS-aided NOMA networks with FBC.
- Considering the impact of the residual error by imperfect SIC, we formulate a nonconvex problem that jointly optimizes active beamforming and the IRS phase shift to maximize the SEC for the given statistical delay QoS constraints of the users, the blocklength, and the decoding error probability. To make the problem more tractable, we propose a tight upper bound for the objective function based on Jensen's inequality and employ the concept of opportunistically minimizing an expectation. We decompose the problem into subproblems: active beamforming at the BS and phase-shift optimization at the IRS. The subproblems are convexified by employing linear approximation and solved using SCA and penalty-based methods. These subproblems are iteratively solved using the AO technique until convergence. Furthermore, the convergence to a suboptimal stationary solution and the computing complexity of the proposed algorithm are rigorously analyzed.
- Finally, through extensive numerical experiments, we evaluate the proposed control in the FBL regime and confirm that it outperforms benchmark schemes in terms of the SEC. Especially, when the numbers of antennas and IRS elements are considerable, the proposed scheme achieves more improvement compared to the SDR-based scheme, which has been widely used to optimize active and passive beamforming.

The remainder of this paper is organized as follows. Section II establishes the system model, and Section III formulates the optimization problem. Next, Section IV discusses the proposed transmission controls and details the convergence and complexity analysis. Finally, Sections V and VI present the simulation results and conclusions, respectively.

II. SYSTEM MODEL

A. System Model

We consider an IRS-aided downlink MISO-NOMA system where a BS equipped with M antennas serves two single-antenna users $U_i, i \in \{1, 2\}$ [1]. In this system, the direct

[1] Our proposed system can be extended to a multi-user downlink NOMA system. For the multi-user system, the users can be grouped into multiple clusters. The BS employs orthogonal multiple access (OMA) schemes, such as time-division multiple access (TDMA) or frequency-division multiple access (FDMA) schemes, to send packages to different clusters [29], [30]. According to [29], clusters of two users are preferable in practical downlink NOMA systems with reduced SIC complexity, less processing overload, and feedback overhead. In addition, this setting was implemented in LTE-A [30]. In this regard, the extension for multi-user transmission based on hybrid OMA-NOMA is interesting and will be left for future work.

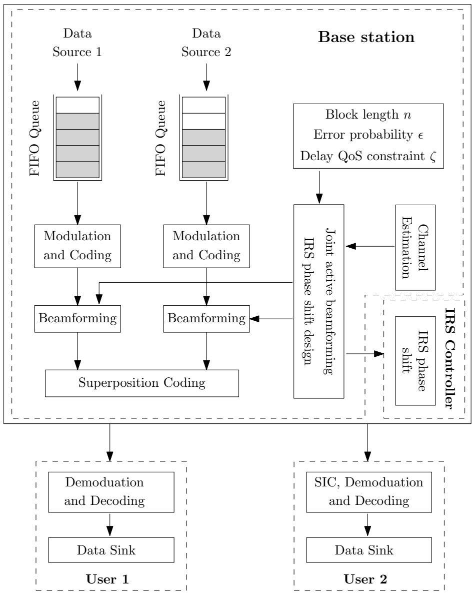

Fig. 1. System block diagram model of the joint active beamforming and IRS phase-shift design in an IRS-aided NOMA URLLC system for the given statistical delay QoS constraints.

link between the BS and users can be weak due to blockage or distance. Therefore, the communication can be assisted by an IRS with K phase shift elements. The IRS phase shift configuration is controlled by a controller connected to the BS through a strong link that provides sufficiently low control overhead for the IRS configuration [2]. The controller has transmission/reception and signal processing capabilities to receive and decode the configuration signal from BS. In addition, we assumed that the perfect channel state information (CSI) can be attained at the BS by conducting appropriate channel estimation.[3] The assumption of perfect CSI for various IRS-aided systems has been considered in [13]–[18]. The perfect CSI acquisition is a challenging problem; however, the results in this study can serve as theoretical performance bounds for systems with imperfect CSI. We consider block fading (i.e., where fading remains constant during a block and varies independently from one block to another). We also assume that the size of a fading block is equal to the blocklength, taken as n symbols [36]. The channels from the BS to the IRS, the IRS to U_i, and the BS to U_i are denoted by $\boldsymbol{A} \in \mathbb{C}^{K \times M}$, $\boldsymbol{h}_{r,i}^H \in \mathbb{C}^{1 \times K}$, and $\boldsymbol{h}_{d,i}^H \in \mathbb{C}^{1 \times M}$, respectively. In addition, we let $\boldsymbol{\Theta} \triangleq \text{diag}(e^{j\theta_1}, \ldots, e^{j\theta_K}) \in \mathbb{C}^{K \times K}$ denote the IRS phase-shift matrix, where $\theta_k \in [0, 2\pi)$ is the phase shift.

As illustrated in Fig. 1, at the BS, packets from each user in the upper layer are assembled into frames and buffered at the first-in-first-out (FIFO) queue, and then split into bit streams for transmission over the wireless channel. In addition, the BS exploits superposition coding and beamforming; therefore, the transmit signal vector can be expressed as $\boldsymbol{x} = \boldsymbol{w}_1 s_1 + \boldsymbol{w}_2 s_2$, where s_i is the transmit data symbol for U_i with $\mathbb{E}\left\{|s_i|^2\right\} = 1$, and $\boldsymbol{w}_i \in \mathbb{C}^{M \times 1}$ denotes the beamforming vector for U_i. Then, the received signal for user U_i can be expressed as follows:

$$y_i = \boldsymbol{h}_i^H (\boldsymbol{w}_1 s_1 + \boldsymbol{w}_2 s_2) + n_i, \tag{1}$$

where $n_i \sim \mathcal{CN}(0, \sigma^2)$ represents the additive white Gaussian noise with zero mean and variance of σ^2, and $\boldsymbol{h}_i^H \triangleq \boldsymbol{h}_{d,i}^H + \boldsymbol{h}_{r,i}^H \boldsymbol{\Theta} \boldsymbol{A} \in \mathbb{C}^{1 \times M}$ denote the combined channel including direct and reflected channels for U_i.

B. Channel Coding Rate in Finite Blocklength Coding

The BS employs FBC for short-packet (message) transmission to support the low-latency requirements for URLLC. The blocklength of an information message is denoted by n. A channel coding process is performed by encoding the message into codewords of length n at the BS. Afterward, at U_i, the received message is decoded into an estimated message, possibly with an error denoted as ϵ_i. In the FBC, the nonzero decoding error probability is nonnegligible; therefore, Shannon's capacity theory is inapplicable. Polyanskiy *et al.* [37] proposed a new channel coding rate in the FBC, which is approximated as follows:

$$R(n, \epsilon_i) \approx \underbrace{\log_2(1 + \gamma_i)}_{\text{Shannon capacity}} - \underbrace{\sqrt{V(\gamma_i)/n} Q^{-1}(\epsilon_i)}_{\text{channel dispersion}}, \tag{2}$$

where γ_i denotes the SINR for U_i, $V(\gamma_i) \triangleq 1 - \frac{1}{(1+\gamma_i)^2}$, and $Q^{-1}(\cdot)$ represents the inverse function of the Q-function $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$. The notation $R(n, \epsilon_i)$ is used to emphasize that this is an achievable rate with a blocklength of n and a decoding error probability of ϵ_i. From (2), for a given blocklength n and packet error probability ϵ_i, the communication rate $R(n, \epsilon_i)$ corresponds to the Shannon capacity minus a channel dispersion term (penalty), which is proportional to $1/\sqrt{n}$. According to [38], when SINR is higher than 5 dB, the approximation $V = 1 - 1/(1 + \gamma_i)^2 \approx 1$. This

[2] According to [31], the IRS reflection matrix can be reconfigured multiple times within the channel coherence time. Specifically, the switching frequency of reflection elements made by positive-intrinsic-negative (PIN) diodes can reach 5 megahertz (MHz), corresponding to the switching time of 0.2 μs, much smaller than the typical coherence time on the order of ms. Therefore, it is practical for the dynamical configuration of the IRS elements among different time slots within the channel coherence time. In addition, it is important to reduce configuration overhead between the BS and the IRS controller, when the IRS has a large number of IRS elements. The authors in [32] aimed to reduce IRS phase shifts feedback overhead by transmitting only the factors (obtained by the tensor-based low-rank factorization approach) to the IRS controller instead of transmitting full IRS phase shift vectors. Once received the smaller factors, the IRS controller can reconstruct the full IRS phase shift vector by adopting known multi-linear structure of the selected low-rank tensor model. The proposed control allows significant feedback overhead reduction, enabling frequent IRS phase shift feedback in fast varying channels, where IRS configuration should catch up with the change of environment.

[3] Because the IRS is not equipped with radio-frequency chains to transmit or receive the pilot symbols, conventional channel estimation methods cannot be applied to obtain the IRS-associated CSI. Many recent studies have been devoted to IRS-channel estimation to address this issue, including compress sensing [33], matrix factorization [34], and deep learning [35]. Specifically, *compress sensing* exploits the channel sparsity to reduce training overhead. *Matrix factorization* decomposes a cascaded channel with high dimensions into sub-channels with low dimensions that are easier to estimate with lower training overhead. Finally, *deep learning model* learns a non-linear function that maps from a training input data to the output cascaded CSI.

approximation can be easily achieved in supporting URLLC. Therefore, (2) becomes

$$R(n,\epsilon_i) \approx \log_2(1+\gamma_i) - Q^{-1}(\epsilon_i)/\sqrt{n}. \tag{3}$$

C. NOMA Transmission Scheme

In NOMA systems, users employ SIC to decode their signals. We assumed that the decoding order is (U_1, U_2) [4]. Specifically, at the user U_1, it decodes its signal while treating the signal of U_2 as interference, therefore, the SINR of user U_1 for decoding itself is written as follows:

$$\gamma_1 = \frac{|\boldsymbol{h}_1^H \boldsymbol{w}_1|^2}{|\boldsymbol{h}_1^H \boldsymbol{w}_2|^2 + \sigma^2}. \tag{4}$$

Accordingly, decoding error probability of U_1's signal at U_1 is denoted as ϵ_1. At user U_2, it applies SIC to decode the signal of U_1 and removes it from the received signal. Accordingly, the SINR of U_2 for decoding U_1's signal can be expressed as

$$\gamma_{1\to 2} = \frac{|\boldsymbol{h}_2^H \boldsymbol{w}_1|^2}{|\boldsymbol{h}_2^H \boldsymbol{w}_2|^2 + \sigma^2}. \tag{5}$$

In the FBC regime, perfect SIC cannot be guaranteed, therefore, decoding error probability cannot be ignored. Then, the decoding error probability of U_1's signal at U_2 is $\bar{\epsilon}_1$. The decoding error probability of U_2's signal at U_2 is as follows:

$$\begin{aligned}\epsilon_2 = \mathbb{P}(D_{2\to 2}=0) = \mathbb{P}(D_{2\to 2}=0|D_{1\to 2}=1)\mathbb{P}(D_{1\to 2}=1)\\ + \mathbb{P}(D_{2\to 2}=0|D_{1\to 2}=0)\mathbb{P}(D_{1\to 2}=0),\end{aligned} \tag{6}$$

where $(D_{i\to j}=1)$ and $(D_{i\to j}=0)$ respectively denote the event that U_i is successfully and unsuccessfully decoded at U_j, $\forall i,j \in \{1,2\}$, and $\mathbb{P}(\cdot)$ denotes the probability of an occurring event. Here, $\mathbb{P}(D_{2\to 2}=0|D_{1\to 2}=1) \overset{\Delta}{=} \bar{\epsilon}_2$ indicates decoding error probability of U_2's signal at U_2 given successful SIC (i.e., $(D_{1\to 2}=1)$ occurs with probability $1-\bar{\epsilon}_1$). In addition, $\mathbb{P}(D_{2\to 2}=0|D_{1\to 2}=0) \overset{\Delta}{=} \tilde{\epsilon}_2$ indicates decoding error probability of U_2's signal at U_2 given failed SIC (i.e., $(D_{1\to 2}=0)$ occurs with probability $\bar{\epsilon}_1$). Then, Eq. (6) can be rewritten as $\epsilon_2 = \bar{\epsilon}_2(1-\bar{\epsilon}_1) + \tilde{\epsilon}_2\bar{\epsilon}_1$.

In practice, SIC might be imperfect due to hardware limitations, errors in data detection, and decoding [40]. Therefore, at the user U_2, the signal of U_1 cannot be completely removed, which remains as the residual interference. As reported in [40], the residual interference can be modeled as a linear function that effectively represents the linear relationship between the residual interference and power of the received signal. Under this model of imperfect SIC, the signal after SIC processing and SINR at U_2 are respectively given as follows:

$$y_2' = \boldsymbol{h}_2^H \boldsymbol{w}_2 s_2 + \sqrt{\varpi}\boldsymbol{h}_2^H \boldsymbol{w}_1 s_1 + n_2, \tag{7}$$

$$\gamma_2 = \frac{|\boldsymbol{h}_2^H \boldsymbol{w}_2|^2}{\varpi|\boldsymbol{h}_2^H \boldsymbol{w}_1|^2 + \sigma^2}, \tag{8}$$

where $0 \le \varpi \le 1$ represents the level of imperfect SIC [5]. Specifically, $\varpi = 0$ indicates the perfect SIC and $\varpi = 1$ indicates no SIC. In this study, the negative impact of imperfect SIC can be alleviated by jointly optimizing the BS beamforming and IRS phase shift matrix.

D. Effective Capacity

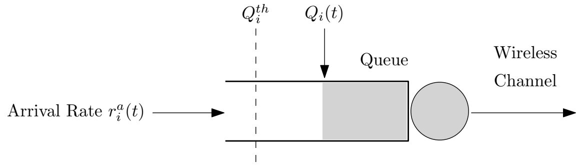

Fig. 2. Queuing model for user i.

Fig. 2 illustrates a discrete-time queuing system for U_i. Every time slot t, new data randomly arrive at the BS for transmission to U_i. Arrival data are stored in queue $Q_i(t)$ and await transmission. We let $(\{r_i^a(t)\}, \{r_i^s(t)\})$ denote the arrival and service processes, respectively. Queue $Q_i(t)$ evolves according to the following equation:

$$Q_i(t+1) = \max\{Q_i(t) - r_i^s(t) + r_i^a(t), 0\}. \tag{9}$$

When stochastic processes $\{r_i^a(t)\}$ and $\{r_i^s(t)\}$ are stationary and ergodic, and $\mathbb{E}\{r_i^a(t)\} < \mathbb{E}\{r_i^s(t)\}$; thus, $Q_i(t)$ converges to a steady state $Q_i(\infty)$ [41]. This paper considers statistical delay QoS constraints that restrict the buffer overflow probability. According to the large deviation theory, the buffer overflow probability is approximated as follows [42]:

$$\Pr\{Q_i(\infty) \ge Q_i^{th}\} \approx e^{-\zeta_i Q_i^{th}}, \tag{10}$$

where Q_i^{th} represents the queue length threshold, ζ_i represents the delay QoS exponent or decay rate of the tail distribution of the queue length. In (10), the statistical delay QoS guarantees are characterized by the delay QoS exponent ζ_i (i.e., the delay QoS exponent ζ_i controls the decay rate of the overflow probability). Specifically, a higher value of ζ_i indicates a stricter limitation on the probability of overflow (i.e., the QoS requirement is more stringent). In contrast, a lower value of ζ_i indicates a looser QoS requirement. According to [43], the following is required to ensure the target buffer overflow probability in (10)

$$\Lambda_a(\zeta_i) + \Lambda_s(-\zeta_i) = 0, \tag{11}$$

where $\Lambda_a(\zeta_i) = \lim_{T\to\infty}\frac{1}{T}\log(\mathbb{E}\{e^{\zeta_i\sum_{t=1}^{T} r_i^a(t)}\})$ and $\Lambda_s(\zeta_i) = \lim_{T\to\infty}\frac{1}{T}\log(\mathbb{E}\{e^{\zeta_i\sum_{t=1}^{T} r_i^s(t)}\})$ are the Gärtner-Ellis limits of the arrival and service processes, respectively.

[4] In NOMA transmission, since the reflected link depends on the unknown IRS phase shift matrix, hence, optimizing decoding order is important for enhancing system performance at cost of increasing complexity. A low-complexity decoding order design in IRS-aided NOMA URLLC system is an interesting topic and worth for further investigation in our future work. For simplicity, according to [39], we replace the phase shift matrix by an identity matrix so that two UEs can be sorted in ascending order.

[5] By sampling over a long training period, the residual interference can be approximated using the Gaussian distribution according to the central limit theorem. The coefficient ϖ can be obtained at the BS by comparing residual interference power and received signal power [40]. In our work, we note that the SIC is only performed once at user U_2, which does not cause excessive accumulated residual interference. Moreover, with the advancements in SIC technology and hardware capabilities, the value of residual interference can be significantly small [19].

When the arrival rate is constant [6] (i.e., $r_i^a(t) = r_i^a$), we have $\Lambda_a(\zeta_i) = \zeta_i r_i^a$. Substituting into (11), we have [46]

$$r_i^a = -\frac{\Lambda_s(-\zeta_i)}{\zeta_i} \triangleq EC(\zeta_i), \tag{12}$$

which is called the *effective capacity*. Intuitively, the effective capacity is a valuable concept that determines the maximum constant arrival rate that a service process can support under a queuing constraint specified by the delay QoS exponent ζ_i.

E. Effective Capacity in the Finite Blocklength Coding Regime

This section introduces the concept of effective capacity with a non-vanishing error probability in FBC. The achievable rate in (3) can be attained with a probability of $1-\epsilon_i$. With the decoding error probability ϵ_i, an error occurs, retransmission is required, and the achievable rate is zero. Accordingly, the service rate (in bits per n channel uses) can be written as follows:

$$r_i^s = \begin{cases} nR(n,\epsilon_i), & \text{with probability } (1-\epsilon_i) \\ 0, & \text{with probability } \epsilon_i. \end{cases} \tag{13}$$

Inspired by [47], we obtain the following result for the effective capacity by inserting the above service rate into (12).

Proposition 1. *For a given SINR γ_i, decoding error probability ϵ_i, blocklength n, and delay QoS exponent ζ_i, the effective capacity in bits per channel use is given by*

$$EC(\zeta_i, n, \epsilon_i) = -\frac{1}{n\zeta_i}\log\left(\mathbb{E}\left\{\epsilon_i + (1-\epsilon_i)e^{-\zeta_i n R(n,\epsilon_i)}\right\}\right), \tag{14}$$

where $\mathbb{E}\{\cdot\}$ is the expectation with respect to the channel state.

Proof. We have:

$$\begin{aligned} EC(\zeta_i, n, \epsilon_i) &\overset{(a)}{=} -\lim_{T\to\infty}\frac{1}{\zeta_i T}\log(\mathbb{E}\{e^{-\zeta_i \sum_{t=1}^{T} r_i^s(t)}\}) \\ &= -\lim_{T\to\infty}\frac{1}{\zeta_i T}\log\left(\mathbb{E}\left\{\prod_{t=1}^{T} e^{-\zeta_i r_i^s(t)}\right\}\right) \\ &\overset{(b)}{=} -\lim_{T\to\infty}\frac{1}{\zeta_i T}\log\left(\prod_{t=1}^{T}\mathbb{E}\left\{e^{-\zeta_i r_i^s(t)}\right\}\right) \\ &\overset{(c)}{=} -\lim_{T\to\infty}\frac{1}{\zeta_i T}\log\left(\mathbb{E}\left\{e^{-\zeta_i r_i^s(t)}\right\}\right)^T \\ &= -\lim_{T\to\infty}\frac{1}{\zeta_i}\log\left(\mathbb{E}\left\{e^{-\zeta_i r_i^s(t)}\right\}\right) \\ &\overset{(d)}{=} -\lim_{T\to\infty}\frac{1}{\zeta_i}\log\left(\mathbb{E}\left\{e^{-\zeta_i n R(n,\epsilon_i)}\right\}\right), \end{aligned} \tag{15}$$

where (a) is obtained from (12). The equalities (b) and (c) are obtained by the fact that the service process changes independently from one block to another and follows the same distribution. The effective capacity in (14) follows by normalizing the expression (d) with n to obtain a unit of bits per channel use, which completes the proof. □

Effective capacity has the following properties.

Proposition 2. ***(i)*** *For a given SINR γ_i, decoding error probability ϵ_i, blocklength n, the effective capacity $EC(\zeta_i, n, \epsilon_i)$ is a monotonically decreasing function of ζ_i, with $\zeta_i \in [0, \infty)$.*
(ii) *When $\zeta_i \to 0$, we obtain $\lim_{\zeta_i\to 0} EC(\zeta_i, n, \epsilon_i) = (1 - \epsilon_i)\mathbb{E}\{R(n, \epsilon_i)\}$, where $R(n, \epsilon_i)$ is given in* (3). *In addition, if $\epsilon_i \to 0$, then $EC = \mathbb{E}\{R\}$, where $R = \log_2(1 + \gamma_i)$ is the ergodic capacity.*
(iii) *When $\zeta_i \to \infty$, we obtain $\lim_{\zeta_i\to\infty} EC(\zeta_i, n, \epsilon_i) = 0$.*
(iv) *For a given SINR γ_i, blocklength n, and delay QoS exponent ζ_i, the effective capacity $EC(\zeta_i, n, \epsilon_i)$ is a quasiconcave function of ϵ_i.*

Proof. (i) The effective capacity is a monotonically decreasing function of $\zeta_i \in [0, \infty)$, which can be checked in [44].
(ii) When $\zeta_i \to 0$, we have

$$\begin{aligned} \lim_{\zeta_i\to 0} EC(\zeta_i, n, \epsilon_i) &= -\lim_{\zeta_i\to 0}\frac{\log\left(\mathbb{E}\left\{\epsilon_i + (1-\epsilon_i)e^{-\zeta_i n R(n,\epsilon_i)}\right\}\right)}{n\zeta_i} \\ &\overset{(a)}{=} -\lim_{\zeta_i\to 0}\frac{\frac{\mathrm{d}\left[\log\left(\mathbb{E}\left\{\epsilon_i+(1-\epsilon_i)e^{-\zeta_i n R(n,\epsilon_i)}\right\}\right)\right]}{\mathrm{d}\zeta_i}}{\mathrm{d}[n\zeta_i]/\mathrm{d}\zeta_i} \\ &= \frac{(1-\epsilon_i)\lim_{\zeta_i\to 0}\mathbb{E}\{R(n,\epsilon_i)e^{-\zeta_i n R(n,\epsilon_i)}\}}{\lim_{\zeta_i\to 0}\mathbb{E}\left\{\epsilon_i + (1-\epsilon_i)e^{-\zeta_i n R(n,\epsilon_i)}\right\}} \\ &= (1-\epsilon_i)\mathbb{E}\{R(n,\epsilon_i)\}, \end{aligned} \tag{16}$$

where (a) follows by the L'Hospital's rule for the indeterminate form $\frac{0}{0}$, when $\zeta_i \to 0$.
(iii) When $\zeta_i \to \infty$,

$$\begin{aligned} &\lim_{\zeta_i\to\infty} EC(\zeta_i, n, \epsilon_i) \\ &= -\lim_{\zeta_i\to\infty}\frac{\log\left(\mathbb{E}\left\{\epsilon_i + (1-\epsilon_i)e^{-\zeta_i n R(n,\epsilon_i)}\right\}\right)}{n\zeta_i} = 0. \end{aligned} \tag{17}$$

(iv) The proof is similar to [48, Proposition 1]. □

According to **Proposition 2**, the effective capacity is no more than $(1 - \epsilon_i)\mathbb{E}\{R(n, \epsilon_i)\}$, for $\zeta_i \geq 0$. When no delay QoS constraint is imposed (i.e., $\zeta_i = 0$) the effective capacity is equal to the average transmission rate averaged over all channel states. In addition, when $\zeta_i = 0$ and $\epsilon_i = 0$ (achieved by Shannon's communication theory, where the decoding error probability can be made arbitrarily small by choosing the packet length sufficiently large), then the effective capacity becomes the ergodic capacity. In contrast, the effective capacity decreases when the QoS delay exponent ζ_i increases from zero to ∞ (i.e., the delay QoS constraints are stricter). In the simulation section, the ergodic capacity can be applied as an upper bound for the effective capacity without decoding error probability and delay QoS constraints.

III. Problem Formulation

For a given delay QoS constraint specified by $\boldsymbol{\zeta} \triangleq \{\zeta_1, \zeta_2\}$, decoding error probability $\boldsymbol{\epsilon} \triangleq \{\epsilon_1, \epsilon_2\}$ and finite blocklength n, we aim to find optimal beamforming vectors $\boldsymbol{w} \triangleq$

[6] By the original concept of the effective capacity (EC) [44], we employed a queueing model where the data packets arrive at the buffer with a constant rate, while the data packets departure at random rate due to the time-varying wireless channel. However, in practice, it can be applied to the random arrival process with changing rate. In this case, its constant arrival rate can be applied by changing it to a constant value based on minimum envelope rate [45] of the arrival process with changing arrival rate, which is compared with the EC we found in this work.

$\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ and optimal IRS phase shift matrix $\boldsymbol{\Theta}$ that maximize *sum effective capacity* (SEC), which is formulated as follows:

$$\textbf{P1}: \max_{\boldsymbol{w},\boldsymbol{\Theta}} \quad SEC(\boldsymbol{\zeta}, n, \boldsymbol{\epsilon}) \triangleq EC(\zeta_1, n, \epsilon_1) + EC(\zeta_2, n, \epsilon_2)$$

$$\text{s.t.} \quad \gamma_1 \leq \gamma_{1\to 2}, \tag{18}$$

$$|\boldsymbol{h}_i^H \boldsymbol{w}_2|^2 \leq |\boldsymbol{h}_i^H \boldsymbol{w}_1|^2, \quad \forall i \in \{1,2\} \tag{19}$$

$$\|\boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2 \leq P_{\max}, \tag{20}$$

$$|e^{j\theta_k}| = 1, \quad \forall k \in \{1, \ldots, K\} \tag{21}$$

where the constraint in (18) ensures successful SIC decoding at U_2 [49]. Typically, to successfully perform SIC at U_2, the SINR at U_2 for decoding U_1's signal should be no less than that at U_1; otherwise U_1's signal cannot be correctly decoded and cannot be completely removed at U_2. In other words, the signal strength of U_1 received at U_2 should be kept sufficiently high so that U_1's signal can be perfectly decoded [6]. The constraint in (19) ensures the rate fairness among users for the given decoding order [49]. Specifically, this constraint indicates that it avoids allocating most of resources to U_2 (with no interference due to SIC), which guarantees that the received signal power of U_2 is lower than that of U_1, leading to a reasonable achievable rate for U_1. The constraint in (20) indicates the transmission power budget. Finally, the constraint in (21) represents the unit-modulus constraint on each IRS phase-shift element.

A. Problem Reformulation

Problem **P1** is equivalently rewritten as follows:

$$\textbf{P1.1}: \min_{\boldsymbol{w},\boldsymbol{\Theta}} \quad -SEC(\boldsymbol{\zeta}, n, \boldsymbol{\epsilon}) \quad \text{s.t.} \quad (18)-(21).$$

Solving the optimization problem **P1.1** is difficult because the objective function contains expectations over channel states. Moreover, the optimization variables are coupled with the objective function and constraints. To address these challenges, we first transform the objective function using the following proposition, providing a tractable upper bound function.

Proposition 3. *The objective function of Problem **P1.1** has the following upper bound:*

$$-SEC(\boldsymbol{\zeta}, n, \boldsymbol{\epsilon}) \leq (\alpha_1 + \alpha_2)\left[\log(\alpha_1 z_1 + \alpha_2 z_2) - \log(\alpha_1 + \alpha_2)\right], \tag{22}$$

where $\alpha_i = \frac{1}{n\zeta_i}, z_i = \mathbb{E}\left\{\epsilon_i + (1-\epsilon_i)e^{-n\zeta_i R(n,\epsilon_i)}\right\}$.

Proof. We obtain the following:

$$-SEC(\boldsymbol{\zeta}, n, \boldsymbol{\epsilon}) = \frac{1}{n\zeta_1}\log\left(\mathbb{E}\left\{\epsilon_1 + (1-\epsilon_1)e^{-n\zeta_1 R(n,\epsilon_1)}\right\}\right) + \frac{1}{n\zeta_2}\log\left(\mathbb{E}\left\{\epsilon_2 + (1-\epsilon_2)e^{-n\zeta_2 R(n,\epsilon_2)}\right\}\right). \tag{23}$$

To determine the upper bound for the right-hand side (RHS) of (23), we apply the well-known Jensen's inequality for the logarithm function, given by

$$\frac{\alpha_1 \log(z_1) + \alpha_2 \log(z_2)}{\alpha_1 + \alpha_2} \leq \log\left(\frac{\alpha_1 z_1 + \alpha_2 z_2}{\alpha_1 + \alpha_2}\right), \tag{24}$$

where the weights are $\alpha_i > 0$. The RHS of (24) can be simplified by applying the rule $\log(a/b) = \log(a) - \log(b)$, and (24) becomes the following:

$$\frac{\alpha_1 \log(z_1) + \alpha_2 \log(z_2)}{\alpha_1 + \alpha_2} \leq \log(\alpha_1 z_1 + \alpha_2 z_2) - \log(\alpha_1 + \alpha_2). \tag{25}$$

By applying the inequality (25), the upper bound of (23) can be obtained by setting $\alpha_i = \frac{1}{n\zeta_i}, z_i = \mathbb{E}\left[\epsilon_i + (1-\epsilon_i)e^{-n\zeta_i R(n,\epsilon_i)}\right]$, completing the proof. □

Remark. The equality in (22) holds when $z_1 = z_2$, that is $\mathbb{E}\left\{\epsilon_1 + (1-\epsilon_1)e^{-n\zeta_1 R(n,\epsilon_1)}\right\} = \mathbb{E}\left\{\epsilon_2 + (1-\epsilon_2)e^{-n\zeta_2 R(n,\epsilon_2)}\right\}$. In the special case when $\epsilon_1 = \epsilon_2$ and $\zeta_1 = \zeta_2$, then the above equality holds if $\mathbb{E}\left\{e^{-n\zeta_1 R(n,\epsilon_1)}\right\} = \mathbb{E}\left\{e^{-n\zeta_2 R(n,\epsilon_2)}\right\}$. This means that the achievable rates $R(n, \epsilon_1)$ and $R(n, \epsilon_2)$ are equal on average (over the channel states). This outcome can be achieved when rate fairness among users is maintained using the constraint in (19).

According to **Proposition 3**, instead of minimizing the objective function of Problem **P1.1**, we minimize its upper bound, as displayed on the RHS of (22). After removing the scalar terms irrelevant to the optimization variables and noting that the function $\log(\cdot)$ is monotonically increasing, we try to solve the following problem

$$\textbf{P1.2}: \min_{\boldsymbol{w},\boldsymbol{\Theta}} \quad \mathbb{E}\left\{f(\boldsymbol{w}, \boldsymbol{\Theta})\right\} \quad \text{s.t.} \quad (18)-(21),$$

where $f(\boldsymbol{w}, \boldsymbol{\Theta}) \triangleq \frac{1}{n\zeta_1}\left\{\epsilon_1 + (1-\epsilon_1)e^{-n\zeta_1 R(n,\epsilon_1)}\right\} + \frac{1}{n\zeta_2}\left\{\epsilon_2 + (1-\epsilon_2)e^{-n\zeta_2 R(n,\epsilon_2)}\right\}$.

However, solving Problem **P1.2** is still challenging because the objective function is stochastic with the expectation operator. To address this difficulty, we employed the concept of *opportunistically minimizing an expectation* (see [50], Section 1.8), which intuitively states that an optimal control for any given channel state is the optimal control that minimizes the expected objective function. This concept is presented in the following proposition.

Proposition 4. *For any given channel state* $\boldsymbol{h} \triangleq (\boldsymbol{h}_1, \boldsymbol{h}_2)$, *we assumed that there exists a solution* $\{\boldsymbol{w}^\star, \boldsymbol{\Theta}^\star\}$ *minimizing* $f(\boldsymbol{w}, \boldsymbol{\Theta})$. *Thus,* $\{\boldsymbol{w}^\star, \boldsymbol{\Theta}^\star\}$ *is also the solution minimizing* $\mathbb{E}\{f(\boldsymbol{w}, \boldsymbol{\Theta})\}$.

Proof. For any given channel state $\boldsymbol{h}$, $f(\boldsymbol{w}^\star, \boldsymbol{\Theta}^\star) \leq f(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\Theta}})$, for any random control $\{\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\Theta}}\}$ taken in response to state $\boldsymbol{h}$. By taking the expectation, we obtain $\mathbb{E}\{f(\boldsymbol{w}^\star, \boldsymbol{\Theta}^\star)\} \leq \mathbb{E}\left\{f(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\Theta}})\right\}$. It follows that $\{\boldsymbol{w}^\star, \boldsymbol{\Theta}^\star\}$ is a minimizer of $\mathbb{E}\{f(\boldsymbol{w}, \boldsymbol{\Theta})\}$, completing the proof. □

Applying this concept, for given channel state $\boldsymbol{h}$, we must determine the optimal solution $\{\boldsymbol{w}^\star, \boldsymbol{\Theta}^\star\}$ minimizing the following problem

$$\textbf{P1.3}: \quad \min_{\boldsymbol{w},\boldsymbol{\Theta}} \quad f(\boldsymbol{w}, \boldsymbol{\Theta}) \quad \text{s.t.} \quad (18)-(21).$$

To solve Problem **P1.3**, we adopt the AO technique that addresses the coupling of optimization variables. Specifically, the variables $\boldsymbol{w}$ and $\boldsymbol{\Theta}$ are alternately optimized while the others are fixed. The solution approach is presented in the subsequent section.

IV. Proposed Solution

A. Optimization of Active Beamforming at the Base Station

In this section, active beamforming at the BS is optimized for a given IRS phase shift. Problem **P1.3** with respect to $\boldsymbol{w}$ can be rewritten as follows:

$$\begin{aligned}\mathbf{P2}: \min_{\boldsymbol{w}} \quad & \frac{1}{n\zeta_1}\left[\epsilon_1 + (1-\epsilon_1)\delta_1 e^{-n\zeta_1 \log_2\left(1+\frac{|\boldsymbol{h}_1^H \boldsymbol{w}_1|^2}{|\boldsymbol{h}_1^H \boldsymbol{w}_2|^2+\sigma^2}\right)}\right] \\ & + \frac{1}{n\zeta_2}\left[\epsilon_1 + (1-\epsilon_2)\delta_2 e^{-n\zeta_2 \log_2\left(1+\frac{|\boldsymbol{h}_2^H \boldsymbol{w}_2|^2}{\varpi|\boldsymbol{h}_2^H \boldsymbol{w}_1|^2+\sigma^2}\right)}\right] \\ \text{s.t.} \quad & (18), (19), (20),\end{aligned}$$

where $\delta_1 \triangleq e^{Q^{-1}(\epsilon_1)\sqrt{n}\zeta_1}$, and $\delta_2 \triangleq e^{Q^{-1}(\epsilon_2)\sqrt{n}\zeta_2}$ are constants. Then, Problem **P2** is rewritten as follows:

$$\begin{aligned}\mathbf{P2.1}: \min_{\boldsymbol{w},\boldsymbol{\alpha}} f_{2.1}(\boldsymbol{w},\boldsymbol{\alpha}) \triangleq & \frac{1}{n\zeta_1}\left[\epsilon_1 + (1-\epsilon_1)\delta_1 e^{-n\zeta_1 \log_2(1+\alpha_1)}\right] \\ & + \frac{1}{n\zeta_2}\left[\epsilon_2 + (1-\epsilon_2)\delta_2 e^{-n\zeta_2 \log_2(1+\alpha_2)}\right]\end{aligned}$$

$$\text{s.t.} \quad \frac{|\boldsymbol{h}_1^H \boldsymbol{w}_1|^2}{\beta_1} \geq \alpha_1, \tag{26}$$

$$|\boldsymbol{h}_1^H \boldsymbol{w}_2|^2 + \sigma^2 \leq \beta_1, \tag{27}$$

$$\frac{|\boldsymbol{h}_2^H \boldsymbol{w}_2|^2}{\beta_2} \geq \alpha_2, \tag{28}$$

$$\varpi|\boldsymbol{h}_2^H \boldsymbol{w}_1|^2 + \sigma^2 \leq \beta_2, \tag{29}$$

$$(18), (19), (20) \tag{30}$$

where α_1, α_2,β_1, and β_2 are the new variables, and $\boldsymbol{\alpha} \triangleq \{\alpha_1, \alpha_2, \beta_1, \beta_2\}$ represents the set of these variables. The equivalence of the two problems, **P2** and **P2.1**, can be demonstrated by the following lemma.

Proposition 5. *Problems **P2** and **P2.1** are equivalent.*

Proof. We demonstrate that $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star)$ is optimal for Problem **P2.1** if and only if $\boldsymbol{w}^\star$ is optimal for Problem **P2** and the constraints in (26) to (29) hold with equality at $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star)$. We first prove that for the given optimal solution $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star)$ for Problem **P2.1**, the constraints in (26) to (29) hold with equality at $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star)$. By contradiction, we suppose that, without loss of generality, at least one constraint (e.g., constraint (26)) holds with strict inequality. Hence, we can increase $\alpha_1^\star$ to $\tilde{\alpha}_1 = \alpha_1^\star + \Delta\alpha_1^\star$ with $\Delta\alpha_1^\star > 0$ such that $\frac{|\boldsymbol{h}_1^H \boldsymbol{w}_1^\star|^2}{\beta_1}^\star = \tilde{\alpha}_1$. We can see that $\tilde{\alpha}_1$ is feasible since it satisfies (26). Substituting $\tilde{\alpha}_1$ into the first term of the objective of Problem **P2.1** implies

$$\begin{aligned}& \frac{1}{n\zeta_1}\left[\epsilon_1 + (1-\epsilon_1)\delta_1 e^{-n\zeta_1 \log_2(1+\tilde{\alpha}_1)}\right] \\ & = \frac{1}{n\zeta_1}\left[\epsilon_1 + (1-\epsilon_1)\delta_1 e^{-n\zeta_1 \log_2(1+\alpha_1^\star+\Delta\alpha_1^\star)}\right] \\ & < \frac{1}{n\zeta_1}\left[\epsilon_1 + (1-\epsilon_1)\delta_1 e^{-n\zeta_1 \log_2(1+\alpha_1^\star)}\right],\end{aligned} \tag{31}$$

which implies that the feasible point $(\boldsymbol{w}^\star, \tilde{\alpha}_1, \alpha_2^\star, \beta_1^\star, \beta_2^\star)$ has a lower objective value than $(\boldsymbol{w}^\star, \alpha_1^\star, \alpha_2^\star, \beta_1^\star, \beta_2^\star)$. Thus, $(\boldsymbol{w}^\star, \alpha_1^\star, \alpha_2^\star, \beta_1^\star, \beta_2^\star)$ is not an optimal solution, contradicting the optimality assumption. The rest of the constraints can be proved using the same argument. Therefore, the optimal solution must satisfy the constraints in (26) to (29) with equality. In addition, the optimal $\boldsymbol{w}^\star$ for Problem **P2.1** yields the same objective value for Problem **P2**. Thus, it is also optimal for Problem **P2**. Conversely, if $\boldsymbol{w}^\star$ is optimal for Problem **P2** and the constraints (26) to (28) hold with equality, then $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star)$ is optimal for Problem **P2.1**, where $\boldsymbol{\alpha}^\star$ is obtained from the equalities (26) to (29), which completes the proof. □

It is observed that the constraints in (26), (28), (18), and (19) in Problem **P2.1** are non-convex. By applying SCA method, we solve the nonconvex problem by successively solving the approximated convex problems. We start with the following lemma.

Lemma 1. ***(i)*** *With $x \in \mathbb{R}, y > 0$, the lower bound for the convex function $f_1(x,y) = x^2/y$ around $(\bar{x}, \bar{y})$ is given as follows:*

$$\frac{x^2}{y} \geq \frac{2\bar{x}}{\bar{y}}x - \frac{\bar{x}^2}{\bar{y}^2}y. \tag{32}$$

(ii) *With $z \in \mathbb{C}, y > 0$, the lower bound for the function $f_2(z,y) = |z|^2/y$ around $(\bar{z}, \bar{y})$ is given as follows:*

$$\frac{|z|^2}{y} \geq \frac{2\Re\{\bar{z}^* z\}}{\bar{y}} - \frac{|\bar{z}|^2}{\bar{y}^2}y, \tag{33}$$

where $|z|$, z^*, $\Re\{z\}$, and $\Im\{z\}$ are the modulus, complex conjugate, real part, and imaginary part of z, respectively.
(iii) *With $z \in \mathbb{C}$, the lower bound for the function $f_3(z) = |z|^2$ around $\bar{z}$ is given as follows:*

$$|z|^2 \geq 2\Re\{\bar{z}^* z\} - |\bar{z}|^2. \tag{34}$$

(iv) *With $x \in \mathbb{R}, y \in \mathbb{R}$, the upper bound for the function $f_4(x,y) = xy$ around the point $(\bar{x}, \bar{y})$ is given as*

$$xy = \frac{1}{4}[(x+y)^2 - (x-y)^2] \tag{35}$$

$$\leq \frac{1}{4}[(x+y)^2 - 2(x-y)(\bar{x}-\bar{y}) + (\bar{x}-\bar{y})^2]. \tag{36}$$

Proof. (i) The result follows by applying the first-order Taylor expansion of the function $f_1(x,y)$ around $(\bar{x}, \bar{y})$, as follows:

$$f_1(x,y) \geq f_1(\bar{x},\bar{y}) + \frac{\partial f_1(\bar{x},y)}{\partial x}(x-\bar{x}) + \frac{\partial f_1(x,\bar{y})}{\partial y}(y-\bar{y}). \tag{37}$$

(ii) The result follows by applying identity $|z|^2 = (\Re\{z\})^2 + (\Im\{z\})^2$ and Property (i).
(iii) The result follows by applying Property (ii) with $y = \bar{y} = 1$.
(iv) The result follows by applying inequality $-(x-y)^2 \leq -2(x-y)(\bar{x}-\bar{y}) + (\bar{x}-\bar{y})^2$. □

Approximation for (26)*:* Because the RHS of (26) is convex, we need to obtain the concave lower bound for the left-hand side (LHS). Using **Lemma 1** (ii), we obtain

$$\frac{|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2}{\beta_1} \geq \frac{2\Re\left\{\left(\boldsymbol{h}_1^H\boldsymbol{w}_1^{(\kappa_1)}\right)^*\boldsymbol{h}_1^H\boldsymbol{w}_1\right\}}{\beta_1^{(\kappa_1)}} - \frac{|\boldsymbol{h}_1^H\boldsymbol{w}_1^{(\kappa_1)}|^2}{\left(\beta_1^{(\kappa_1)}\right)^2}\beta_1, \tag{38}$$

where $\boldsymbol{w}_1^{(\kappa_1)}$ and $\beta_1^{(\kappa_1)}$ represent the values of $\boldsymbol{w}$ and β_1, respectively, obtained at the (κ_1-1)-th iteration of the SCA. Thus, at $(\boldsymbol{w}_1^{(\kappa_1)}, \beta_1^{(\kappa_1)})$, the constraint in (26) can be approximated as follows:

$$\frac{2\Re\left\{\left(\boldsymbol{h}_1^H\boldsymbol{w}_1^{(\kappa_1)}\right)^*\boldsymbol{h}_1^H\boldsymbol{w}_1\right\}}{\beta_1^{(\kappa_1)}} - \frac{|\boldsymbol{h}_1^H\boldsymbol{w}_1^{(\kappa_1)}|^2}{\left(\beta_1^{(\kappa_1)}\right)^2}\beta_1 \geq \alpha_1. \tag{39}$$

Approximation for (28)*:* Using **Lemma 1** (iii), the constraint in (28) can be approximated at $\boldsymbol{w}_2^{(\kappa_1)}$:

$$\frac{2\Re\left\{\left(\boldsymbol{h}_2^H\boldsymbol{w}_2^{(\kappa_1)}\right)^*\boldsymbol{h}_2^H\boldsymbol{w}_2\right\}}{\beta_2^{(\kappa_1)}} - \frac{|\boldsymbol{h}_2^H\boldsymbol{w}_2^{(\kappa_1)}|^2}{\left(\beta_2^{(\kappa_1)}\right)^2}\beta_2 \geq \alpha_2. \tag{40}$$

Approximation for (18)*:* We rewrite the constraint as

$$|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2\left(|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2+\sigma^2\right) \leq |\boldsymbol{h}_2^H\boldsymbol{w}_1|^2\left(|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2+\sigma^2\right), \tag{41}$$

which is equivalent to

$$|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2 \leq a, \tag{42}$$
$$|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2 + \sigma^2 \leq b, \tag{43}$$
$$|\boldsymbol{h}_2^H\boldsymbol{w}_1|^2 \geq c, \tag{44}$$
$$|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2 + \sigma^2 \geq d, \tag{45}$$
$$ab \leq cd, \tag{46}$$

where a, b, c, and d are the new variables, and $\boldsymbol{\beta} \triangleq \{a,b,c,d\}$ represents the set of these variables. The first two constraints are convex, and the remaining are nonconvex. These nonconvex constraints can be approximated using **Lemma 1** (iii) and (iv), respectively, as follows:

$$2\Re\left\{\left(\boldsymbol{h}_2^H\boldsymbol{w}_1^{(\kappa_1)}\right)^*\boldsymbol{h}_2^H\boldsymbol{w}_1\right\} - |\boldsymbol{h}_2^H\boldsymbol{w}_1^{(\kappa_1)}|^2 \geq c, \tag{47}$$
$$2\Re\left\{\left(\boldsymbol{h}_1^H\boldsymbol{w}_2^{(\kappa_1)}\right)^*\boldsymbol{h}_1^H\boldsymbol{w}_2\right\} - |\boldsymbol{h}_1^H\boldsymbol{w}_2^{(\kappa_1)}|^2 \geq d-\sigma^2, \tag{48}$$
$$\begin{aligned}&(a+b)^2 - 2(a-b)(a^{(\kappa_1)}-b^{(\kappa_1)}) + (a^{(\kappa_1)}-b^{(\kappa_1)})^2 \\ &+ (c-d)^2 - 2(c+d)(c^{(\kappa_1)}+d^{(\kappa_1)}) + (c^{(\kappa_1)}+d^{(\kappa_1)})^2 \leq 0,\end{aligned} \tag{49}$$

where $a^{(\kappa_1)}, b^{(\kappa_1)}, c^{(\kappa_1)}$, and $d^{(\kappa_1)}$ represent the values of $a, b, c,$ and d, respectively, obtained at the (κ_1-1)-th iteration.

Approximation for (19)*:* Using **Lemma 1** (iii) yields convex approximations, as follows:

$$|\boldsymbol{h}_i^H\boldsymbol{w}_2|^2 \leq 2\Re\left\{\left(\boldsymbol{h}_i^H\boldsymbol{w}_1^{(\kappa_1)}\right)^*\boldsymbol{h}_i^H\boldsymbol{w}_1\right\} - |\boldsymbol{h}_i^H\boldsymbol{w}_1^{(\kappa_1)}|^2, \quad \forall i \in \{1,2\}. \tag{50}$$

Therefore, the approximated convex problem for the Problem **P2.1** at the κ_1-th iteration of the SCA is

$$\begin{aligned}\textbf{P2.2}: \min_{\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & f_{2.2}(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}) \triangleq f_{2.1}(\boldsymbol{w},\boldsymbol{\alpha}) \\ \text{s.t.} \quad & (20),(27),(29),(39),(40),(42),(43),(47)-(50).\end{aligned}$$

Because Problem **P2.2** is convex, it can be solved using the CVX solver [51]. The active beamforming design under the SCA framework is summarized in **Algorithm 1**.

Algorithm 1 Active Beamforming Design

```
Initialization: Initialize feasible points $(\boldsymbol{w}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)})$, set
    the iteration index $\kappa_1 = 0$, and tolerance $\varepsilon_1 > 0$.
Repeat
    Solve Problem P2.2 to obtain the solution $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$;
    Update $\boldsymbol{w}^{(\kappa_1+1)} \leftarrow \boldsymbol{w}^\star, \boldsymbol{\alpha}^{(\kappa_1+1)} \leftarrow \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^{(\kappa_1+1)}) \leftarrow \boldsymbol{\beta}^\star$;
    $\kappa_1 \leftarrow \kappa_1 + 1$;
Until Convergence for the given tolerance $\varepsilon_1$.
```

The feasibility and convergence of **Algorithm 1** are analyzed in the following proposition.

Proposition 6. *We let $\boldsymbol{y} \triangleq \{\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}\}$, $\mathcal{X}_{\kappa_1}$, and $\mathcal{X}$ denote the set of optimization variables, feasible set for Problem* ***P2.2****, and feasible set for Problem* ***P2.1****, respectively. In addition, $\{\boldsymbol{y}^{(\kappa_1)}\}$ represents the sequence produced by* ***Algorithm 1****. Hence,*
(i) *$\mathcal{X}_{\kappa_1} \subseteq \mathcal{X}$.*
(ii) *For any $\kappa_1 \geq 0$, $\boldsymbol{y}^{(\kappa_1)}$ is a feasible point of Problem* ***P2.1****.*
(iii) *The sequence of the objective obtained from* ***Algorithm 1*** *is non-increasing and bounded below; thus, it is convergent.*
(iv) *Let $\bar{\boldsymbol{y}}$ be the limit point of the sequence $\{\boldsymbol{y}^{(\kappa_1)}\}$, and assume that $\mathcal{X}_{\kappa_1}$ satisfies the Slater condition, then $\bar{\boldsymbol{y}}$ is a Karush-Kuhn-Tucker (KKT) point of Problem* ***P2.1****.*

Proof. (i) We focus on the constraint in (26) and its convex approximation (39). Similar arguments are applied to the other constraints. For ease of notation, let define the LHS of (26) and (39) as $\Gamma_1(\boldsymbol{y})$, and $\Gamma_1^{(\kappa_1)}(\boldsymbol{y})$, respectively. Thus, these constraints are rewritten as follows:

$$\Gamma_1(\boldsymbol{y}) \geq \alpha_1, \tag{51}$$
$$\Gamma_1^{(\kappa_1)}(\boldsymbol{y}) \geq \alpha_1. \tag{52}$$

From (38), $\Gamma_1(\boldsymbol{y}) \geq \Gamma_1^{(\kappa_1)}(\boldsymbol{y}), \forall \boldsymbol{y}$. Therefore, for any feasible point $\hat{\boldsymbol{y}}$ satisfying (52), we consistently have $\Gamma_1(\hat{\boldsymbol{y}}) \geq \Gamma_1^{(\kappa_1)}(\hat{\boldsymbol{y}}) \geq \alpha_1$. It follows that $\hat{\boldsymbol{y}}$ is also feasible for (51).
(ii) This result immediately follows from (i) because $\boldsymbol{y}^{(\kappa_1)} \in \mathcal{X}_{\kappa_1} \subseteq \mathcal{X}$.
(iii) The LHS of (28) and (40) are denoted as $\Gamma_2(\boldsymbol{y})$ and $\Gamma_2^{(\kappa_1)}(\boldsymbol{y})$, respectively. First, we prove that the sequence of

the objective is non-increasing. We obtain the following:

$$
\begin{aligned}
f_{2.2}(\boldsymbol{y}^{(\kappa_1+1)}) &= \frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_1)\delta_1 e^{-n\zeta_1\log_2\left(1+\Gamma_1(\boldsymbol{y}^{(\kappa_1+1)})\right)}\right]\\
&\quad+\frac{1}{n\zeta_2}\left[\epsilon_1+(1-\epsilon_2)\delta_2 e^{-n\zeta_2\log_2\left(1+\Gamma_2(\boldsymbol{y}^{(\kappa_1+1)})\right)}\right]\\
&\overset{(a)}{\leq}\frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_1)\delta_1 e^{-n\zeta_1\log_2\left(1+\Gamma_1^{(\kappa_1)}(\boldsymbol{y}^{(\kappa_1+1)})\right)}\right]\\
&\quad+\frac{1}{n\zeta_2}\left[\epsilon_1+(1-\epsilon_2)\delta_2 e^{-n\zeta_2\log_2\left(1+\Gamma_2^{(\kappa_1)}(\boldsymbol{y}^{(\kappa_1+1)})\right)}\right]\\
&\overset{(b)}{\leq}\frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_1)\delta_1 e^{-n\zeta_1\log_2\left(1+\Gamma_1^{(\kappa_1)}(\boldsymbol{y}^{(\kappa_1)})\right)}\right]\\
&\quad+\frac{1}{n\zeta_2}\left[\epsilon_1+(1-\epsilon_2)\delta_2 e^{-n\zeta_2\log_2\left(1+\Gamma_2^{(\kappa_1)}(\boldsymbol{y}^{(\kappa_1)})\right)}\right]\\
&\overset{(c)}{=}\frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_1)\delta_1 e^{-n\zeta_1\log_2\left(1+\Gamma_1(\boldsymbol{y}^{(\kappa_1)})\right)}\right]\\
&\quad+\frac{1}{n\zeta_2}\left[\epsilon_1+(1-\epsilon_2)\delta_2 e^{-n\zeta_2\log_2\left(1+\Gamma_2(\boldsymbol{y}^{(\kappa_1+1)})\right)}\right]\\
&= f_{2.2}(\boldsymbol{y}^{(\kappa_1)}),
\end{aligned}
\tag{53}
$$

where (a) is obtained because $\Gamma_m(\boldsymbol{y}) \geq \Gamma_m^{(\kappa_1)}(\boldsymbol{y}), m \in \{1,2\}$, (b) follows because $(\boldsymbol{w}^{(\kappa_1+1)}, \boldsymbol{\alpha}^{(\kappa_1+1)})$ is the optimal solution to Problem **P2.2** at the κ_1-th iteration, and (c) is obtained because $\Gamma_m(\boldsymbol{y}^{(\kappa_1)}) = \Gamma_m^{(\kappa_1)}(\boldsymbol{y}^{(\kappa_1)}), m \in \{1,2\}$. In addition, the sequence $\{f_{2.2}(\boldsymbol{y}^{(\kappa_1)})\}$ is bounded below by (20). Therefore, it is convergent.
(iv) From (iii), we obtain $\lim_{\kappa_1\to\infty} f_{2.2}(\boldsymbol{y}^{(\kappa_1)}) = f_{2.2}(\bar{\boldsymbol{y}})$. There exists a subsequence of $\left\{\boldsymbol{y}^{(\kappa_1)}\right\}$ converging to the limit point $\bar{\boldsymbol{y}}$. According to [52, Theorem 1], $\bar{\boldsymbol{y}}$ is a KKT point for **P2.2**, then it is also a KKT point for **P2.1**, which completes the proof.

□

Algorithm 2 Feasible Point Search Algorithm

```
Initialization: Randomly initialize points $(\boldsymbol{w}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)})$,
    set the iteration index $\kappa' = 0$, and tolerance $\varepsilon' > 0$.
Repeat
    Solve Problem P2.3 to obtain the solution $(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, e^\star)$;
    Update $\boldsymbol{w}^{(\kappa'+1)} \leftarrow \boldsymbol{w}^\star, \boldsymbol{\alpha}^{(\kappa'+1)} \leftarrow \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^{(\kappa'+1)}) \leftarrow \boldsymbol{\beta}^\star$;
    $\kappa' \leftarrow \kappa' + 1$;
Until $e^\star$ below the given tolerance $\varepsilon'$.
```

It is important to initialize a feasible point for the iterative algorithm, which is a nontrivial task. In the following, we propose an initialization scheme that provides a feasible point, inspired the feasibility search algorithm adopted in [53]. Let $e \geq 0$ denote an *infeasibility indicator*, which indicates how far the constraints in Problem **P2.2** are from being satisfied. The feasible point search problem can be stated at the κ'-th iteration, as follows:

$$
\begin{aligned}
\textbf{P2.3}: \min_{\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta},e} \quad & e\\
\text{s.t.} \quad & \mathcal{C}_m \leq e, \forall m, \text{ and } e \geq 0.
\end{aligned}
\tag{54}
$$

where $\mathcal{C}_m$ represents the m-th constraint of Problem **P2.2** with all terms being moved to the LHS. The above optimization problem is convex, therefore it can be solved efficiently using CVX solver. The proposed feasible point search algorithm is summarized in **Algorithm 2**. Interestingly, the initial points in **Algorithm 2** can be generated randomly. When $e = 0$, two problems **P2.2** and **P2.3** have the same feasible set. Then, the output of this algorithm can be considered as an initial feasible input for **Algorithm 1**.

B. IRS Phase-Shift Optimization

This section solves the IRS phase-shift matrix for a given $\boldsymbol{w}$. First, the combined channel $\boldsymbol{h}_i$ can be rewritten as follows:

$$
\boldsymbol{h}_i^H = \boldsymbol{h}_{d,i}^H + \boldsymbol{\phi}^H \boldsymbol{H}_i \in \mathbb{C}^{1\times M}, \tag{55}
$$

where $\boldsymbol{H}_i \triangleq \text{diag}\left(\boldsymbol{h}_{r,i}^H\right)\boldsymbol{A} \in \mathbb{C}^{K\times M}$ and $\boldsymbol{\phi} \triangleq [\phi_1, ..., \phi_K]^H \in \mathbb{C}^{K\times 1}$ denotes a vector whose elements are collected from the diagonal entries of the matrix $\boldsymbol{\Theta}$ (i.e., $\phi_k = e^{j\theta_k}$). We defined $\boldsymbol{a}_{i,l} \triangleq \boldsymbol{H}_i\boldsymbol{w}_l \in \mathbb{C}^{K\times 1}, b_{i,l} \triangleq \boldsymbol{h}_{d,i}^H\boldsymbol{w}_l \in \mathbb{C}, \forall i,l \in \{1,2\}$. Then,

$$
|\boldsymbol{h}_i^H\boldsymbol{w}_l|^2 = \left|\left(\boldsymbol{h}_{d,i}^H + \boldsymbol{\phi}^H\boldsymbol{H}_i\right)\boldsymbol{w}_l\right|^2 = |b_{i,l} + \boldsymbol{\phi}^H\boldsymbol{a}_{i,l}|^2. \tag{56}
$$

The Problem **P1.3** with respect to $\boldsymbol{\phi}$ is written as follows:

$$
\begin{aligned}
\textbf{P3}: \min_{\boldsymbol{\phi}} \ & \frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_1)\delta_1 e^{-n\zeta_1\log_2\left(1+\frac{|\boldsymbol{\phi}^H\boldsymbol{a}_{1,1}+b_{1,1}|^2}{|\boldsymbol{\phi}^H\boldsymbol{a}_{1,2}+b_{1,2}|^2+\sigma^2}\right)}\right]\\
&+\frac{1}{n\zeta_2}\left[\epsilon_1+(1-\epsilon_2)\delta_2 e^{-n\zeta_2\log_2\left(1+\frac{|\boldsymbol{\phi}^H\boldsymbol{a}_{2,2}+b_{2,2}|^2}{\varpi|\boldsymbol{\phi}^H\boldsymbol{a}_{2,1}+b_{2,1}|^2+\sigma^2}\right)}\right]\\
\text{s.t.} \quad & \frac{|\boldsymbol{\phi}^H\boldsymbol{a}_{1,1}+b_{1,1}|^2}{|\boldsymbol{\phi}^H\boldsymbol{a}_{1,2}+b_{1,2}|^2+\sigma^2} \leq \frac{|\boldsymbol{\phi}^H\boldsymbol{a}_{2,1}+b_{2,1}|^2}{|\boldsymbol{\phi}^H\boldsymbol{a}_{2,2}+b_{2,2}|^2+\sigma^2}, \qquad (57)\\
& |\boldsymbol{\phi}^H\boldsymbol{a}_{i,2}+b_{i,2}|^2 \leq |\boldsymbol{\phi}^H\boldsymbol{a}_{i,1}+b_{i,1}|^2, \quad \forall i \in \{1,2\} \qquad (58)\\
& |\phi_k| = 1, \quad \forall k \in \{1,\dots,K\} \qquad (59)
\end{aligned}
$$

Problem **P3** is equivalently formulated as follows:

$$
\begin{aligned}
\textbf{P3.1}: \min_{\boldsymbol{\phi},\boldsymbol{\vartheta}} \quad & f_{3.1}(\boldsymbol{\phi},\boldsymbol{\vartheta}) \triangleq\\
& \frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_1)\delta_1 e^{-n\zeta_1\log_2(1+\vartheta_1)}\right]\\
& +\frac{1}{n\zeta_1}\left[\epsilon_1+(1-\epsilon_2)\delta_2 e^{-n\zeta_2\log_2(1+\vartheta_2)}\right]\\
\text{s.t.} \quad & \frac{|\boldsymbol{\phi}^H\boldsymbol{a}_{1,1}+b_{1,1}|^2}{\eta_1} \geq \vartheta_1, \qquad (60)\\
& |\boldsymbol{\phi}^H\boldsymbol{a}_{1,2}+b_{1,2}|^2+\sigma^2 \leq \eta_1, \qquad (61)\\
& \frac{|\boldsymbol{\phi}^H\boldsymbol{a}_{2,2}+b_{2,2}|^2}{\eta_2} \geq \vartheta_2, \qquad (62)\\
& \varpi|\boldsymbol{\phi}^H\boldsymbol{a}_{2,1}+b_{2,1}|^2+\sigma^2 \leq \eta_2, \qquad (63)\\
& (57), (58), (59), \qquad (64)
\end{aligned}
$$

where $\boldsymbol{\vartheta} \triangleq \{\vartheta_1, \vartheta_2, \eta_1, \eta_2\}$ is the set of new variables. Similarly, at the κ_2-th iteration of the SCA, the nonconvex constraints in (60), (62), and (58) can be respectively approximated:

$$\frac{2\Re\left\{\left((\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{1,1}+b_{1,1}\right)^*\left(\boldsymbol{\phi}^H\boldsymbol{a}_{1,1}+b_{1,1}\right)\right\}}{\eta_1^{(\kappa_2)}} - \frac{|(\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{1,1}+b_{1,1}|^2}{\left(\eta_1^{(\kappa_2)}\right)^2}\eta_1 \geq \vartheta_1, \tag{65}$$

$$\frac{2\Re\left\{\left((\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{2,2}+b_{2,2}\right)^*\left(\boldsymbol{\phi}^H\boldsymbol{a}_{2,2}+b_{2,2}\right)\right\}}{\eta_2^{(\kappa_2)}} - \frac{|(\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{2,2}+b_{2,2}|^2}{\left(\eta_2^{(\kappa_2)}\right)^2}\eta_2 \geq \vartheta_2, \tag{66}$$

$$\begin{aligned}&|\boldsymbol{\phi}^H\boldsymbol{a}_{i,2}+b_{i,2}|^2\\ &\leq 2\Re\left\{\left((\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{i,1}+b_{i,1}\right)^*\left(\boldsymbol{\phi}^H\boldsymbol{a}_{i,1}+b_{i,1}\right)\right\}\\ &-|(\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{i,1}+b_{i,1}|^2, \quad \forall i\in\{1,2\}\end{aligned} \tag{67}$$

The constraint (57) can be approximated as follows:

$$|\boldsymbol{\phi}^H\boldsymbol{a}_{1,1}+b_{1,1}|^2 \leq p, \tag{68}$$

$$|\boldsymbol{\phi}^H\boldsymbol{a}_{2,2}+b_{2,2}|^2+\sigma^2 \leq q, \tag{69}$$

$$\begin{aligned}&2\Re\left\{\left((\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{2,1}+b_{2,1}\right)^*\left(\boldsymbol{\phi}^H\boldsymbol{a}_{2,1}+b_{2,1}\right)\right\}\\ &-|(\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{2,1}+b_{2,1}|^2 \geq r,\end{aligned} \tag{70}$$

$$\begin{aligned}&2\Re\left\{\left((\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{1,2}+b_{1,2}\right)^*\left(\boldsymbol{\phi}^H\boldsymbol{a}_{1,2}+b_{1,2}\right)\right\}\\ &-|(\boldsymbol{\phi}^{(\kappa_2)})^H\boldsymbol{a}_{1,2}+b_{1,2}|^2 \geq s-\sigma^2,\end{aligned} \tag{71}$$

$$\begin{aligned}&(p+q)^2-2(p-q)(p^{(\kappa_2)}-q^{(\kappa_2)})+(p^{(\kappa_2)}-q^{(\kappa_2)})^2\\ &+(r-s)^2-2(r+s)(r^{(\kappa_2)}+s^{(\kappa_2)})+(r^{(\kappa_2)}+s^{(\kappa_2)})^2 \leq 0.\end{aligned} \tag{72}$$

where p, q, r, and s are the new variables, and $\boldsymbol{\chi} \triangleq \{p, q, r, s\}$ represents the set of these variables. Hence, the approximate convex problem for Problem **P3.1** at the κ_2-th iteration is:

$$\begin{aligned}\textbf{P3.2}: \min_{\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\chi}} \quad & f_{3.2}(\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\chi}) \triangleq f_{3.1}(\boldsymbol{\phi},\boldsymbol{\vartheta})\\ \text{s.t.} \quad & (59),(61),(63),(65)-(72),\end{aligned}$$

This problem is still non-convex because the constraint in (59) is non-convex. Therefore, the penalty method is employed to address the nonconvexity of the constraint in (59). We first relax this constraint to an inequality; thus, it is convex. Next, we introduce a penalty term $-\mu\sum_{k=1}^{K}(|\phi_k|^2-1)$ to enforce the inequality to the equality, where $\mu > 0$ represents the penalty factor. Therefore, Problem **P3.2** becomes

$$\begin{aligned}\textbf{P3.3}: \min_{\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\chi}} \quad & f_{3.2}(\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\chi})-\mu\sum_{k=1}^{K}(|\phi_k|^2-1)\\ \text{s.t.} \quad & (61),(63),(65)-(72), \qquad (73)\\ & |\phi_k|^2 \leq 1, \forall k\in\{1,...,K\}. \qquad (74)\end{aligned}$$

The term $-\mu|\phi_k|^2$ in the objective of Problem **P3.3** is not convex. Thus, we again apply **Lemma 1** (iii) to convexify this term, resulting the following problem

$$\begin{aligned}\textbf{P3.4}: \min_{\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\chi}} \; & f_{3.2}(\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\chi})-\mu\sum_{k=1}^{K}(2\Re\{(\phi_k^{(\kappa_2)})^*\phi_k\}-|\phi_k^{(\kappa_2)}|^2)\\ \text{s.t.} \; & (73),(74).\end{aligned} \tag{75}$$

This problem is convex, hence it can be solved using the CVX solver [51]. The algorithm for solving Problem **P3** is presented in **Algorithm 3**. The convergence analysis for this algorithm is similar to the previous part, which is omitted.

Algorithm 3 IRS Phase Shift Design

```
Initialization: Initialize feasible points $(\boldsymbol{\phi}^{(0)}, \boldsymbol{\vartheta}^{(0)}, \boldsymbol{\chi}^{(0)})$, set the iteration index $\kappa_2 = 0$, and tolerance $\varepsilon_2 > 0$.
Repeat
    Solve Problem P3.4 to obtain solution $(\boldsymbol{\phi}^\star, \boldsymbol{\vartheta}^\star, \boldsymbol{\chi}^\star)$;
    Update $\boldsymbol{\phi}^{(\kappa_2+1)} \leftarrow \boldsymbol{\phi}^\star, \boldsymbol{\vartheta}^{(\kappa_2+1)} \leftarrow \boldsymbol{\vartheta}^\star, \boldsymbol{\chi}^{(\kappa_2+1)}) \leftarrow \boldsymbol{\chi}^\star$;
Until Convergence for the given tolerance $\varepsilon_2$.
```

C. Alternating Optimization-based Algorithm, Convergence and Complexity Analysis

Algorithm 4 presents the unified algorithm based on the AO framework that jointly optimizes the active beamforming and IRS phase shift. In addition, the convergence analysis for **Algorithm 4** is presented in **Proposition 7**.

Algorithm 4 Alternating Optimization-based Algorithm for Unified Solution

```
Initialization: Randomly initialize $\boldsymbol{\phi}^{(0)}$, find a feasible point $\boldsymbol{w}^{(0)}$ using Algorithm 2, set the iteration index $\kappa_0 = 0$ and tolerance $\varepsilon_0 > 0$.
Repeat
    For a given IRS phase-shift matrix $\boldsymbol{\phi}^{(\kappa_0)}$ obtained at the previous iteration, obtain $\boldsymbol{w}^{(\kappa_0+1)}$ using Algorithm 1.
    For a given active beamforming $\boldsymbol{w}^{(\kappa_0+1)}$, obtain the IRS phase shift $\boldsymbol{\phi}^{(\kappa_0+1)}$ using Algorithm 3.
    $\kappa_0 \leftarrow \kappa_0 + 1$;
Until Convergence for the given tolerance $\varepsilon_0$.
```

Proposition 7. *The sequence of the objective obtained from **Algorithm 4** is non-increasing and bounded from below, which converges to a suboptimal point.*

Proof. We have the following:

$$\begin{aligned}f(\boldsymbol{w}^{(\kappa_0)},\boldsymbol{\phi}^{(\kappa_0)}) &= f_{2.1}(\boldsymbol{w}^{(\kappa_0)},\boldsymbol{\alpha}^{(\kappa_0)}) \geq f_{2.1}(\boldsymbol{w}^{(\kappa_0+1)},\boldsymbol{\alpha}^{(\kappa_0+1)})\\ &= f(\boldsymbol{w}^{(\kappa_0+1)},\boldsymbol{\phi}^{(\kappa_0)}) = f_{3.1}(\boldsymbol{\phi}^{(\kappa_0)},\boldsymbol{\vartheta}^{(\kappa_0)})\\ &\geq f_{3.1}(\boldsymbol{\phi}^{(\kappa_0+1)},\boldsymbol{\vartheta}^{(\kappa_0+1)}) = f(\boldsymbol{w}^{(\kappa_0+1)},\boldsymbol{\phi}^{(\kappa_0+1)}).\end{aligned}$$

In addition, the objective function is bounded from below due to the constraints in (20) and (21); thus, the sequence of objectives is convergent. The solution is suboptimal because Problem **P1.3** is nonconvex. □

For the proposed method, the worst-case complexity of **Algorithms 1**, **2**, and **3** is $\mathcal{O}\left(M^{3.5}\log(1/\varepsilon_c)I_1\right)$, $\mathcal{O}\left(M^{3.5}\log(1/\varepsilon_c)I_2\right)$, and $\mathcal{O}\left(K^{3.5}\log(1/\varepsilon_c)I_3\right)$, respectively, where ε_c, I_1, I_2, and I_3 denote the accuracy of the interior point algorithm adopted in the CVX solver [51] and the number of iterations for **Algorithms 1 2**, and **3**, respectively. Finally, the overall complexity of the joint active beamforming and IRS phase-shift algorithm (i.e., **Algorithm 4)** is $\mathcal{O}\left(M^{3.5}\log(1/\varepsilon_c)I_2 + \left(M^{3.5}I_1 + K^{3.5}I_3\right)\log(1/\varepsilon_c)I_0\right)$, where I_0 denotes the number of iterations in **Algorithm 4**.

In existing studies, the SDR method is common for solving active beamforming and IRS phase-shift subproblems, which can be conducted by formulating Problems **P2** and **P3** as semidefinite programming problems. The rank-one constraints are relaxed, followed by Gaussian randomization to recover the rank-one solution. According to [54], the total complexity of the joint optimization algorithm is $\mathcal{O}\Bigg(M^{4.5}\log(1/\varepsilon_c)I_2 + \left(M^{4.5}I_1 + K^{4.5}I_3\right)\log(1/\varepsilon_c)I_0\Bigg)$, which is higher than that of the proposed algorithm. The complexity can be prohibitively high when K and M are large. Moreover, Gaussian randomization technique is not guaranteed to have a rank-one solution and fails to generate a feasible solution in specific scenarios [55]. For comparison, in the simulation, the SDR-based algorithm is added as a benchmark scheme.

V. Numerical Results

This section provides numerical examples to validate the performance of the proposed algorithm in IRS-assisted NOMA-URLLC networks under statistical delay QoS constraints. For the evaluation, we consider a BS with $M = 6$ antennas located at $(0\text{ m}, 0\text{ m})$, a single-antenna U_2 located at $(10\text{ m}, 1\text{ m})$, and a single-antenna U_1 at $(15\text{ m}, 5\text{ m})$. An IRS with $K = 30$ reflection elements is located at $(12\text{ m}, 5\text{ m})$ to assist in downlink communication between the BS and users. We assume that the direct channel between the BS and user (i.e., $\mathbf{h}_{d,i}$) follows Rayleigh fading[7] , which is determined by $\boldsymbol{h}_{d,i} = L_{d,i}(d)\bar{\boldsymbol{h}}_{d,i}$ with the distance-dependent pathloss $L_{d,i}(d) = 32.6 + 36.7\log_{10}(d)$ dB and non-line-of-sight (NLoS) component $\bar{\boldsymbol{h}}_{d,i} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I})$. The IRS-assisted channels $\mathbf{A}$, $\mathbf{h}_{r,1}$, and $\mathbf{h}_{r,2}$ were assumed to follow Rician fading. We also assumed that the BS and IRS were equipped with a uniform linear array of antennas and IRS elements, respectively. Therefore, they can be modeled as follows:

$$\mathbf{A} = L_A(d)\left(\sqrt{\frac{\delta}{\delta+1}}\mathbf{a}_K(\psi)\mathbf{a}_M(\varsigma)^H + \sqrt{\frac{1}{\delta+1}}\bar{\mathbf{A}}\right), \quad (76)$$

$$\mathbf{h}_{r,i} = L_{r,i}(d)\left(\sqrt{\frac{\delta}{\delta+1}}\mathbf{a}_K(\varrho_i) + \sqrt{\frac{1}{\delta+1}}\bar{\mathbf{h}}_{r,i}\right), \forall i \quad (77)$$

where $\mathbf{L}(d) = 35.6 + 22.0\log_{10}(d)$ dB, and $\mathbf{L}(d) = \{L_A(d), L_{r,i}(d)\}$ are the corresponding pathlosses. In addition, δ represents the Rician factor $\delta = 10$, $(\mathbf{a}_K, \mathbf{a}_M)$ denotes the set of steering vectors, $(\psi, \varsigma, \varrho_i)$ indicate the angular parameters, and $\bar{\mathbf{A}}$ and $\bar{\mathbf{h}}_{r,i}$ denote the NLoS components whose elements are distributed as $\mathcal{CN}(0, 1)$. The transmission bandwidth and noise power are set to 1 MHz and $\sigma^2 = -70$ dBm [56], respectively. In addition, unless otherwise stated, we assume that the level of imperfect SIC is $\varpi = 0$. We compare the performance of the proposed algorithm with the following benchmark schemes:

- **Upper bound 1 (also referred to as "UB 1 ($\zeta_i = 0$, $n = \infty$, $\epsilon_i = 0$)"):** In this scheme, Shannon's capacity for U_i is adopted (i.e., no decoding error ($\epsilon_i = 0$) and has an infinite blocklength ($n = \infty$)). In addition, we assumed no delay QoS constraints (i.e., $\zeta_i = 0$). This approach provides an upper bound on the effective capacity according to **Proposition 2** (ii).
- **Upper bound 2 (also referred to as "UB 2 ($\zeta_i = 0$)"):** In this scheme, we consider communication with finite blocklength, therefore, we adopt the channel coding rate in FBC (i.e., $R(n, \epsilon_i)$ given in Eq. (3)). According to **Proposition 2** (ii), when $\zeta_i = 0$ (i.e., statistical delay QoS guarantee is not considered), the effective capacity is equal to $(1 - \epsilon_i)\mathbb{E}\{R(n, \epsilon_i)\}$.
- **SDR-based algorithm [17]:** This scheme also employed the AO algorithm to decompose the primal problem into subproblems to address the coupling of variables. Each subproblem was converted into the rank-constrained semidefinite programming form. Then, the SDR method was adopted to solve the problem.
- **Proposed without IRS (also referred to as "Proposed w/o IRS"):** This scheme adopts the proposed scheme without the IRS.
- **FDMA:** This scheme adopts the frequency division multiple access (FDMA), where each user is allocated half of the bandwidth. Two users transmit their signal simultaneously to the BS over two equal adjacent frequency resource blocks. Accordingly, the achievable rate for user i is formulated as follows:

$$R_i^{\text{FDMA}} = \frac{1}{2}\log_2\left(1 + \frac{|\boldsymbol{h}_i^H\boldsymbol{w}_i|^2}{\frac{1}{2}\sigma^2}\right), i \in \{1, 2\}. \quad (78)$$

- **TDMA:** This scheme adopts the time division multiple access (TDMA), where each user is allocated half of the transmission time. Two users transmit their signal to the BS consecutively over two equal adjacent time slots. It is noted that by considering the "time-selective" property of the IRS [57], the phase shift matrix can be set differently for two users over two time slots, denoted as $\boldsymbol{\Theta}_i$. Accordingly, the achievable rate for user i is formulated as follows:

$$R_i^{\text{TDMA}} = \frac{1}{2}\log_2\left(1 + \frac{2|\tilde{\boldsymbol{h}}_i^H\boldsymbol{w}_i|^2}{\sigma^2}\right), i \in \{1, 2\}, \quad (79)$$

where $\tilde{\boldsymbol{h}}_i^H \triangleq \boldsymbol{h}_{d,i}^H + \boldsymbol{h}_{r,i}^H\boldsymbol{\Theta}_i\boldsymbol{A} \in \mathbb{C}^{1\times M}$.

[7] In this study, we assume that the direct links between the BS and users are severely affected by the blockages, such as buildings and trees in the urban areas. In this case, NLoS links are dominant compared to the LoS ones. Therefore, it is practical to assume that there are only NLoS direct links, which is an ideal case for the IRS implementation that provides the configuration links to enhance the transmission. Moreover, our system model can be easily extended to the case where both NLoS and LoS direct links coexist, which is modeled by the well-known Rician fading model.

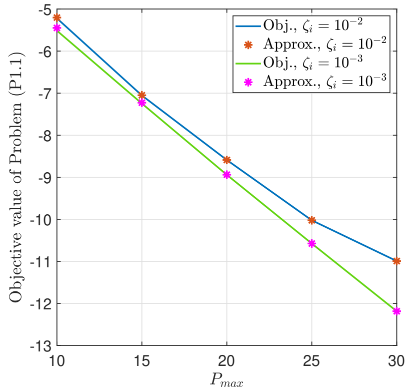

Fig. 3. Tightness evaluation for the approximation in (22) with different values of delay QoS exponent.

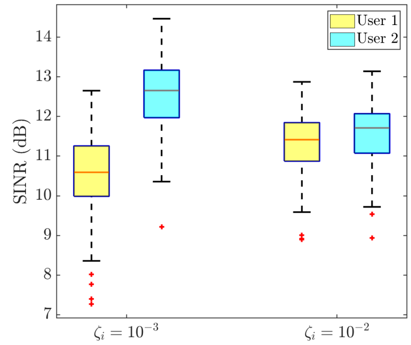

Fig. 4. Box plot verifying received SINR of two users with different values of delay QoS exponent.

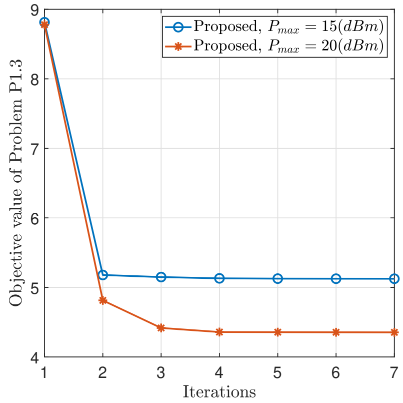

Fig. 5. Convergence behavior of the proposed algorithm.

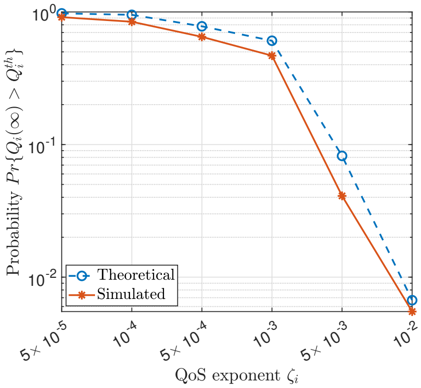

Fig. 6. Overflow probability vs. QoS exponent ζ_i.

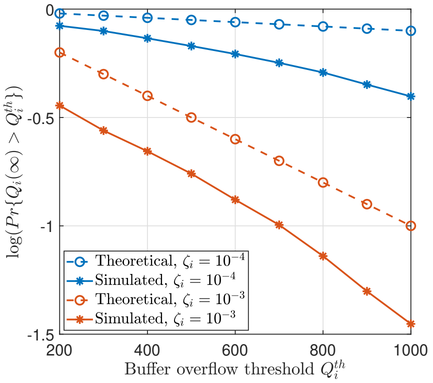

Fig. 7. Overflow probability vs. buffer overflow threshold Q_i^{th} for different values of QoS exponent.

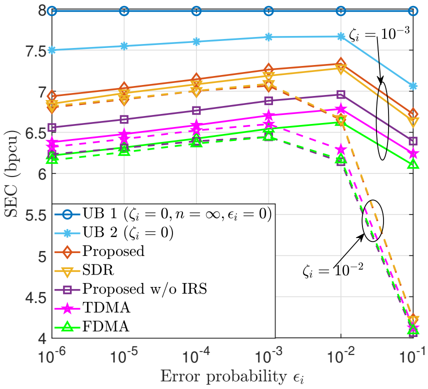

Fig. 8. SEC vs. decoding error probability.

A. Accuracy of Approximation in (22)

In Fig. 3, we investigate the tightness of the approximation derived in (22), where the objective function of Problem **P1.1** is approximated by the function on the RHS of the inequality in (22). In this figure, the curves "Obj." and "Approx." indicate the objective function and the approximation function, respectively. By varying maximum transmit power with delay QoS exponent $\zeta_i \in \{10^{-2}, 10^{-3}\}$, it can be seen that the two curves overlap each other. Therefore, this figure numerically verifies that the approximation function is suitable for optimization instead of the complicated objective function in Problem **P1.1**.

B. Evaluation of SINR Values

In this part, we verify that the attained SINR values are above the threshold of 5dB, which confirms the approximation in (3). In the numerical experiment, we set $\epsilon_i = 10^{-5}, n = 200$. As shown in Fig. 4, the distribution of the SINR values for 2 users over two values of delay QoS parameter $\zeta_i \in \{10^{-3}, 10^{-2}\}$. It can be seen that all SINR values of 2 users are above 5dB, confirming the approximation in (3).

C. Convergence Behavior of AO-based Algorithm 4

Fig. 5 shows the convergence behavior of the proposed algorithm for different values of power budget $P_{max} \in \{15, 20\}$ (dBm). The objective function of the overall algorithm is the objective function of Problem **P1.3**. The convergence tolerance is set as 10^{-3}. As shown in Fig. 5, the objective value, obtained by **Algorithms 4**, decreases rapidly and saturates as the number of iterations increases. Specifically, the algorithm converges after 5 iterations regardless of the BS transmit power.

D. Influence of Delay QoS Exponent on the Buffer Overflow Probability

This part presents the influence of QoS exponent on the buffer overflow probability. In the numerical experiment, we set $\epsilon_i = 10^{-5}, n = 200$. First, we repeat the simulation 1000 times. For each simulation, we randomly generate channel realizations for each time block and compute the constant arrival rate based on EC expression in (14). Then, we simulate the queue using the queueing model given in (9). Next, for a given Q_i^{th}, the probability $\Pr\{Q_i(\infty) \geq Q_i^{\text{th}}\}$ is estimated by counting frequency of the event $(Q_i(\infty) \geq Q_i^{th})$ over 1000

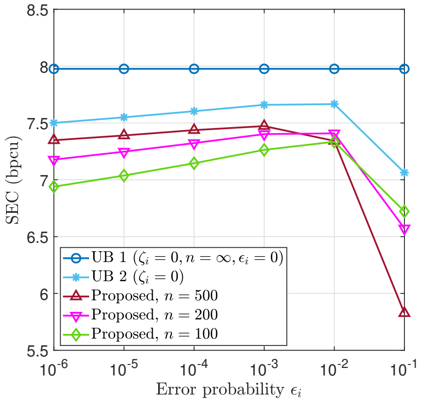

Fig. 9. SEC vs. decoding error probability with different values of blocklength.

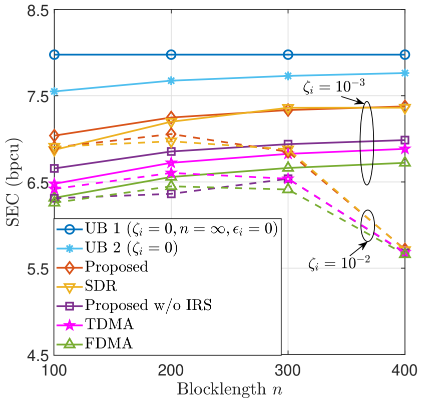

Fig. 10. SEC vs. blocklength.

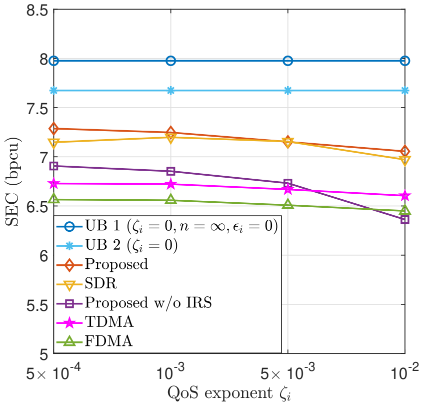

Fig. 11. SEC vs. delay QoS exponent.

times of simulation. Theoretically, for large Q_i^{th}, we rewrite the approximation in (10), as follows:

$$\Pr\{Q_i(\infty) \geq Q_i^{th}\} \approx e^{-\zeta_i Q_i^{\text{th}}}, \tag{80}$$

To verify the accuracy of the approximation given in (80), the obtained simulated overflow probability is compared with the theoretical one given in the RHS of (80). In Figs. 6 and 7, we denote simulated overflow probability by 'Simulated', while the theoretical one given by $e^{-\zeta_i Q_i^{\text{th}}}$ is denoted by 'Theoretical'. From both figures, we observe that the 'Theoretical' curve is an upper bound. The reason for this is explained as follows. According to [44], for smaller Q_i^{th}, the following approximation is more accurate than (80), which is given by $\Pr\{Q_i(\infty) \geq Q_i^{th}\} \approx \lambda e^{-\zeta_i Q_i^{\text{th}}}$, where $\lambda = \Pr\{Q_i(\infty) > 0\}$ is the probability that the queue is not empty. Because $\lambda < 1$, hence $\lambda e^{-\zeta_i Q_i^{\text{th}}} < e^{-\zeta_i Q_i^{\text{th}}}$. Therefore, the simulation result is lower than the theoretical one. In addition, because $\lambda < 1$, we can observe that there exists a gap between two curves. Fig. 6 illustrates the overflow probability against the delay QoS exponent for $Q_i^{th} = 500$. It can be seen that the overflow probability decreases as the delay QoS exponent increases. Fig. 7 illustrates the overflow probability (in logarithm scale) against the queuelength threshold Q_i^{th}. According to (80), we know that $\log(\Pr\{Q_i(\infty) \geq Q_i^{\text{th}}\}) = -\zeta_i Q_i^{\text{th}}$, which is theoretically linear in Q_i^{th} with slope $-\zeta_i$. We can see that the logarithmic buffer overflow probabilities decrease almost linearly and close to the theoretical value. Therefore, both results show the consistency with the theoretical analysis.

E. Influence of the Delay QoS Exponent, Decoding Error Probability, and Blocklength on the SEC

Fig. 8 presents the SEC as a function of the decoding error probability with the delay QoS exponent $\zeta_i \in \{10^{-2}, 10^{-3}\}$. The blocklength is fixed at $n = 100$. As expected, with the ideal scenario ($\zeta_i = 0, n = \infty, \epsilon_i = 0$), the upper bound performs the best because no delay QoS requirement exists, and the transmission is performed under an infinite blocklength with zero decoding probability. According to **Proposition 2** (ii), UB1 scheme is the ergodic capacity, which does not depend on the value of the error probability; hence, its performance remains unchanged over all values of ϵ_i. In contrast, for UB2 scheme, we considered FBC without delay QoS guarantees (i.e., $\zeta_i = 0$). In fact, the performance achieved by the UB2 scheme is lower than UB1 scheme and varies with error probability. Because there is no delay QoS constraint, UB 2 scheme can achieves better performance than the other schemes, which can be viewed as an upper bound. Furthermore, for the cases $\zeta_i \in \{10^{-2}, 10^{-3}\}$, as the error probability increases, SEC increases then decreases after a threshold. This is because using a robust coding scheme with a small decoding error probability requires a small data transmission rate, leading to a small effective capacity. In contrast, if a higher transmission rate with relatively weak channel coding is favored, then the error probability can be increased, leading to retransmission, which again reduces the effective capacity. Second, SEC with $\zeta_i = 10^{-3}$ is higher than that with $\zeta_i = 10^{-2}$ because a lower effective capacity can be achieved when a stricter QoS delay constraint is applied. Third, the proposed scheme can achieve better performance than the proposed one without the IRS scheme because the proposed approach exploits the IRS, which improves the signal quality through passive beamforming. In addition, the proposed scheme is superior to the TDMA and FDMA, because NOMA is employed to serve multiple users in the same resource block, offering higher spectral efficiency. Moreover, we can see that TDMA scheme achieves better performance than its counterpart FDMA, because the time selective property of the IRS is exploited in TDMA. This result reveals the importance of integrating IRS and NOMA in URLLC services. Moreover, the proposed scheme slightly outperforms the SDR-based scheme.

Fig. 9 illustrates the SEC against the decoding error probability with various blocklength values when $n \in \{100, 200, 500\}$, whereas the delay QoS exponent is fixed at $\zeta_i = 10^{-3}$. It can be seen that all curves are quasiconcave and each curve achieves the maximum value at a unique ϵ_i^*, as theoretically demonstrated in **Proposition 2** (iv). When blocklength is longer, ϵ_i^* is smaller, and the SEC attained at ϵ_i^* becomes higher. Therefore, if the channel code lowers error decoding, then increasing blocklength is a useful way to

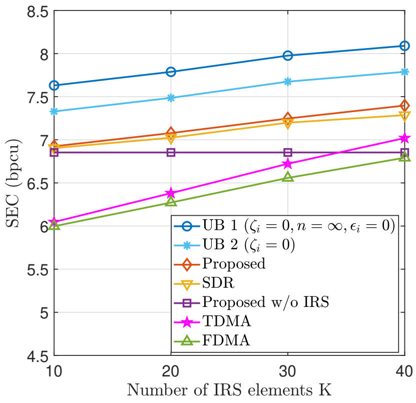

Fig. 12. SEC vs. the number of IRS elements.

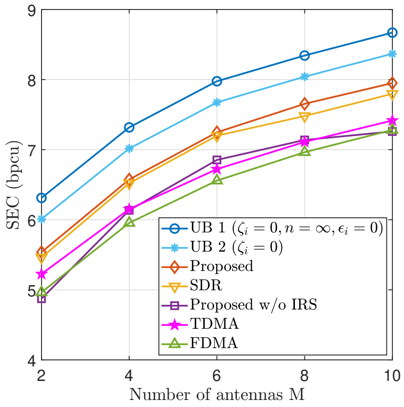

Fig. 13. SEC vs. the number of antennas.

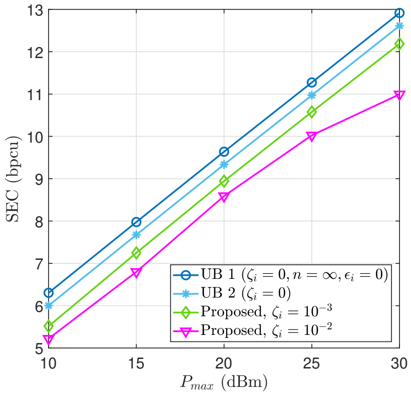

Fig. 14. SEC vs. the maximum transmit power.

increase the SEC.

Fig. 10 illustrates the SEC as a function of the blocklength for $\zeta_i \in \{10^{-2}, 10^{-3}\}$, whereas the decoding error probability is fixed at $\epsilon_i = 10^{-5}$. When $\zeta_i = 10^{-2}$, the SEC initially increases as the blocklength increases. However, as the blocklength increases beyond a specific threshold, the SEC decreases. This is because the blocklength is the coherence duration on which the fading state remains constant (i.e., a longer blocklength corresponds to slower fading). In the case of slow-fading, strong attenuation may persist, causing a long duration of low-rate transmission, resulting in a buffer overflow [47]. In such cases, the systems are more conservative and support lower arrival rates to avoid buffer overflow. In contrast, when $\zeta_i = 10^{-3}$ (i.e., a looser QoS delay constraint), the SEC increases as the blocklength increases.

Fig. 11 illustrates the influence of the delay QoS exponent on the SEC, where $n = 200$ and $\epsilon_i = 10^{-5}$. The SEC decreases as the delay QoS exponent increases, which is consistent with the results in **Proposition 2** (i). Under a stringent delay QoS requirement, a small effective capacity can be supported by a low-rate transmission. In addition, the proposed scheme outperforms the SDR, proposed scheme without the IRS, TDMA, and FDMA schemes under all values for the delay QoS exponents.

Fig. 12 presents the SEC against the number of IRS elements when $n = 200, \epsilon_i = 10^{-5}, \zeta_i = 10^{-3}$. A larger number of IRS elements leads to a higher SEC for the IRS-aided schemes, whereas the SEC of the proposed scheme without the IRS remains unchanged. This is because, as the number of IRS elements increases, more phase shifters for the incoming signals can improve the spatial diversity. The system can exploit the additional DoFs to enhance the signal power, which ultimately improves the performance system. In addition, TDMA and FDMA schemes offer the lowest performance for a low value of K. However, as $K \geq 40$, the TDMA scheme outperforms the proposed scheme without the IRS, highlighting the benefit of IRS phase shift optimization with a large number of IRS elements. The proposed scheme outperforms the SDR, proposed without the IRS, TDMA, and FDMA schemes for the overall values of K. Moreover, the performance gain of the proposed scheme over the SDR is pronounced with a higher number of IRS elements because with a higher K, the probability of obtaining a rank-one solution in the Gaussian randomization is lower.

Fig. 13 depicts the SEC as a function of the number of antennas when $n = 200, \epsilon_i = 10^{-5}, \zeta_i = 10^{-3}$. The SEC of the proposed scheme is higher than that of the SDR, proposed scheme without the IRS, TDMA and FDMA schemes. In addition, we notice that more antennas provide higher diversity gain, leading to a higher transmission rate. Thus, this result implies a higher effective capacity. It is also observed that the performance gap between the proposed and SDR schemes is greater with an increase in the number of antennas.

Fig. 14 illustrates the effect of the maximum transmit power on the SEC for different values of delay QoS exponent $\zeta_i \in \{10^{-2}, 10^{-3}\}$, whereas the blocklength and decoding error probability are fixed at $n = 200$ and $\epsilon_i = 10^{-5}$, respectively. In all cases of QoS constraints, the SEC increases as the maximum transmit power increases. As expected, UB1 perform the best. In addition, with a higher QoS exponent (i.e., stricter QoS constraint), lower SEC can be achieved.

Fig. 15 illustrates the effect of of imperfect SIC process on the SEC. As can be seen from the figure, when the imperfect SIC level increase, the performance of our proposed scheme is lower than the OMA schemes, which highlights the importance of reducing the interference caused by imperfect SIC.

Moreover, we also verify the performance of the proposed scheme compared with the SDMA scheme under a multi-user setup. Specifically, we consider the scenario, where the number of users $N \geq 2$ and the number of antennas $M = 3$. It is noted that our proposed scheme can be easily extended to the hybrid NOMA-TDMA scheme, where the users are grouped into clusters with two users. We specifically investigate the channel correlation in each cluster, where the channel correlation coefficient between two channel vectors of two users, denoted as $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, can be measured by $\rho = \frac{|\boldsymbol{h}_1^H \boldsymbol{h}_2|}{\|\boldsymbol{h}_1\|\|\boldsymbol{h}_2\|}$, where $\rho = 0$, $0 < \rho < 1$, and $\rho = 1$ indicate that the two channel vectors are mutually orthogonal, correlated at a certain level, and highly correlated (i.e., same direction), respectively [58]. As can be seen in Fig. 16, for the case $\rho = 0$, which is favorable for SDMA, the performance is superior to the proposed scheme as $N \leq 6$. In contrast,

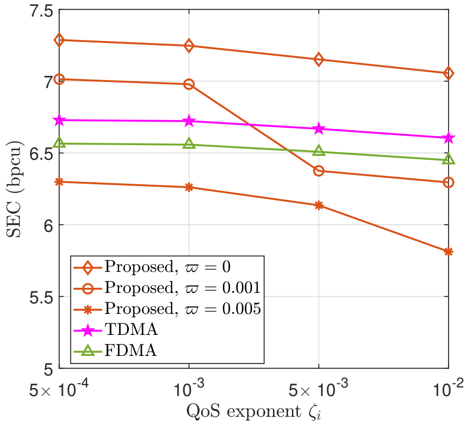

Fig. 15. SEC vs. delay QoS exponent with different values of imperfect SIC levels.

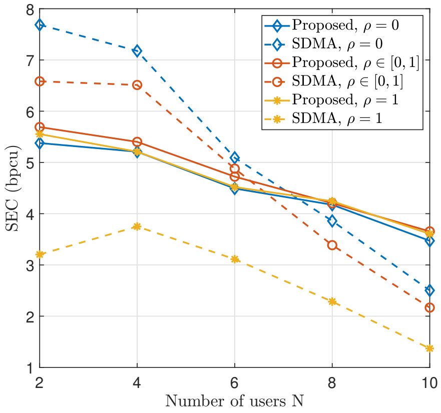

Fig. 16. SEC vs. number of users in multi-user scenario.

as $N \geq 8$ becomes large, the proposed scheme outperforms the SDMA scheme. For the case $\rho = 1$, which is favorable for NOMA, the proposed scheme outperforms the SDMA scheme over all values of N. Next, for the case where ρ is randomly taken in the interval $[0,1]$, the proposed scheme outperforms the SDMA scheme as $N \geq 8$. For all cases, it can be seen that the proposed scheme shows better performance than the SDMA scheme as the number of users N is much larger than the number of antennas, i.e., $N > M$. This is because SDMA users suffer from intense interference with insufficient spatial DoFs. In contrast, for NOMA, users can be grouped into clusters. In each cluster, the superimposed of two users is transmitted through beamforming to carry out NOMA. Intra-cluster interference can be canceled out using SIC and inter-cluster interference can be mitigated by transmitting over orthogonal time slots.

VI. Conclusions

This work developed optimal transmission schemes maximizing the SEC in the FBL regime, which can statistically guarantee URLLC QoS in IRS-assisted NOMA networks. We formulated a nonconvex problem that jointly optimizes active beamforming and the IRS phase shift while ensuring the delay QoS constraints. To make the problem tractable, we derived a tight upper bound of the objective function and employed the concept of opportunistically minimizing an expectation. Then, we decomposed the problem into subproblems: active beamforming at the BS and phase-shift optimization at the IRS. The subproblems were solved using the SCA and AO techniques until convergence. The convergence to a suboptimal stationary solution and the computing complexity of the proposed algorithm were rigorously analyzed. Finally, through extensive numerical experiments, we evaluated the proposed control in the FBL regime and confirmed significant improvement in terms of SEC compared to the existing benchmark schemes. As the number of antennas and IRS elements increases, the performance and complexity gain becomes clearer. As future works, we will investigate more enhanced and lower complexity EC maximization transmission controls that consider imperfect CSI, various IRS types such as active IRS and STAR-IRS [59], and new access schemes such as RSMA and hybrid OMA-NOMA/RSMA.

References

[1] S. He, Z. An, J. Zhu, J. Zhang, Y. Huang, and Y. Zhang, "Beamforming design for multiuser URLLC with finite blocklength transmission," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8096–8109, 2021.

[2] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 2019.

[3] T. T. H. Pham, W. Noh, and S. Cho, "Multi-agent reinforcement learning based optimal energy sensing threshold control in distributed cognitive radio networks with directional antenna," *ICT Express*, 2024.

[4] S. H. Lim, J. Han, W. Noh, Y. Song, and S.-W. Jeon, "Hybrid neural coded modulation: Design and training methods," *ICT Express*, vol. 8, no. 1, pp. 25–30, 2022.

[5] U. Ghafoor, M. Ali, H. Z. Khan, A. M. Siddiqui, and M. Naeem, "NOMA and future 5G & B5G wireless networks: A paradigm," *J. Netw. Comput. Appl.*, vol. 204, p. 103413, 2022.

[6] 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, 2022.

[7] S. N. Sur and R. Bera, "Intelligent reflecting surface assisted MIMO communication system: A review," *Phys. Commun.*, vol. 47, p. 101386, 2021.

[8] Y. Xu, C. Shen, T.-H. Chang, S.-C. Lin, Y. Zhao, and G. Zhu, "Transmission energy minimization for heterogeneous low-latency NOMA downlink," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1054–1069, 2020.

[9] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson, and J. Ansari, "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G," in *IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2016, pp. 1–8.

[10] J. Cheng, C. Shen, Z. Chen, and N. Pappas, "Robust beamforming design for IRS-aided URLLC in D2D networks," *IEEE Trans. Commun.*, 2022.

[11] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550–4564, 2018.

[12] 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, 2022.

[13] H. Ren, K. Wang, and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 707–723, 2021.

[14] R. Hashemi, S. Ali, N. Mahmood, and M. Latva-aho, "Average rate and error probability analysis in short packet communications over RIS-aided URLLC systems," *IEEE Trans. Veh. Technol.*, 2021.

[15] B. Xu, H. Huang, J.-B. Wang, L. Qiu, H. Zhang, and Y. Zhang, "Energy-efficient precoding design for downlink IRS-assisted URLLC system," in *IEEE 2nd International Conference on Information Communication and Software Engineering (ICICSE)*. IEEE, 2022, pp. 141–145.

[16] H. Xie, J. Xu, Y.-F. Liu, L. Liu, and D. W. K. Ng, "User grouping and reflective beamforming for IRS-aided URLLC," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2533–2537, 2021.

[17] J. Wang, J. Yao, D. Chen, J.-B. Wang, L. Ge, and H. Zhang, "Joint power allocation and phase shift design for IRS-aided NOMA-URLLC systems," in *IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2022, pp. 815–820.

[18] C. Zhang, M. Cui, and G. Zhang, "Throughput optimization for IRS-assisted multi-user NOMA URLLC systems," *Wirel. Netw.*, vol. 29, no. 6, pp. 2505–2517, 2023.

[19] S. Lv, X. Xu, S. Han, Y. Liu, P. Zhang, and A. Nallanathan, "STAR-RIS enhanced finite blocklength transmission for uplink NOMA networks," *IEEE Trans. Commun.*, 2023.

[20] T. H. T. Le, Y. K. Tun, N. T. Thu, L. V. Nguyen, and E.-N. Huh, "Min-max decoding error probability optimization in RIS-aided hybrid TDMA-NOMA networks," *IEEE Access*, 2024.

[21] Y. Yang, Y. Hu, and M. C. Gursoy, "Energy efficiency of RIS-assisted NOMA-based MEC networks in the finite blocklength regime," *IEEE Trans. Commun.*, 2023.

[22] M. Katwe, R. Deshpande, K. Singh, C. Pan, P. H. Ghare, and T. Q. Duong, "Spectrally-efficient beamforming design for STAR-RIS-aided URLLC NOMA systems," *IEEE Trans. Commun.*, 2024.

[23] R. Deshpande, M. V. Katwe, K. Singh, and Z. Ding, "Resource allocation design for spectral-efficient URLLC using RIS-aided FD-NOMA system," *IEEE Wireless Commun. Lett.*, vol. 12, no. 7, pp. 1209–1213, 2023.

[24] Y. Zhu, X. Yuan, Y. Hu, T. Wang, M. C. Gursoy, and A. Schmeink, "Low-latency hybrid NOMA-TDMA: QoS-driven design framework," *IEEE Trans. Wireless Commun.*, 2022.

[25] J. Zeng, C. Xiao, Z. Li, W. Ni, and R. P. Liu, "Dynamic power allocation for uplink NOMA with statistical delay QoS guarantee," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8191–8203, 2021.

[26] Z. Ding and H. V. Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, 2020.

[27] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3007–3038, 2019.

[28] M. Amjad, L. Musavian, and S. Aïssa, "Effective capacity of NOMA with finite blocklength for low-latency communications," *arXiv preprint arXiv:2002.07098*, 2020.

[29] A. S. De Sena, F. R. M. Lima, D. B. Da Costa, Z. Ding, P. H. Nardelli, U. S. Dias, and C. B. Papadias, "Massive MIMO-NOMA networks with imperfect SIC: design and fairness enhancement," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6100–6115, 2020.

[30] J. Yao, Q. Zhang, and J. Qin, "Joint decoding in downlink NOMA systems with finite blocklength transmissions for ultrareliable low-latency tasks," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17 705–17 713, 2022.

[31] W. Jiang and H. D. Schotten, "Intelligent reflecting vehicle surface: A novel IRS paradigm for moving vehicular networks," in *MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM)*. IEEE, 2022, pp. 793–798.

[32] B. Sokal, P. R. Gomes, A. L. de Almeida, B. Makki, and G. Fodor, "Reducing the control overhead of intelligent reconfigurable surfaces via a tensor-based low-rank factorization approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6578–6593, 2023.

[33] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, "Channel estimation for reconfigurable intelligent surface aided multi-user mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, 2023.

[34] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2621–2636, 2020.

[35] C. Liu, X. Liu, D. W. K. Ng, and J. Yuan, "Deep residual learning for channel estimation in intelligent reflecting surface-assisted multi-user communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 898–912, 2021.

[36] M. Amjad, L. Musavian, and S. Aissa, "NOMA versus OMA in finite blocklength regime: Link-layer rate performance," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 253–16 257, 2020.

[37] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[38] X. Ou, X. Xie, H. Lu, H. Yang, and H. Tang, "Energy-efficient resource allocation for short packet transmission in MISO multicarrier NOMA," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 12 797–12 810, 2022.

[39] Y. Yang, Y. Hu, and M. C. Gursoy, "Energy efficiency of RIS-assisted noma-based MEC networks in the finite blocklength regime," *IEEE Trans. Commun.*, 2023.

[40] X. Chen, R. Jia, and D. W. K. Ng, "On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2539–2551, 2018.

[41] R. Han, Y. Yu, L. Bai, J. Wang, J. Choi, and W. Zhang, "Effective capacity analysis of delay-sensitive communications in NOMA systems," *IEEE Trans. Wireless Commun.*, 2023.

[42] J. Choi, "An effective capacity-based approach to multi-channel low-latency wireless communications," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2476–2486, 2018.

[43] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, 1995.

[44] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, 2003.

[45] C.-S. Chang, *Performance Guarantees in Communication Networks*, 07 2001, vol. 12.

[46] Y. Hu, Y. Li, M. C. Gursoy, S. Velipasalar, and A. Schmeink, "Throughput analysis of low-latency IoT systems with QoS constraints and finite blocklength codes," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3093–3104, 2020.

[47] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Throughput-delay tradeoffs with finite blocklength coding over multiple coherence blocks," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5892–5904, 2019.

[48] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP J. Wirel. Commun. Netw.*, vol. 2013, pp. 1–13, 2013.

[49] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, 2020.

[50] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synth. Lect. Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.

[51] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[52] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, University of Minnesota, 2014.

[53] J. Zuo, Y. Liu, Z. Qin, and N. Al-Dhahir, "Resource allocation in intelligent reflecting surface assisted NOMA systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7170–7183, 2020.

[54] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, 2010.

[55] A. H. Phan, H. D. Tuan, H. H. Kha, and D. T. Ngo, "Nonsmooth optimization for efficient beamforming in cognitive radio multicast transmission," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2941–2951, 2012.

[56] W. Wang, L. Duan, X. Liu, and N. Zhao, "Enhancing MISO-NOMA networks via constructive interference precoding," *IEEE Trans. Commun.*, 2023.

[57] G. Chen, Q. Wu, W. Chen, D. W. K. Ng, and L. Hanzo, "IRS-aided wireless powered mec systems: TDMA or NOMA for computation offloading?" *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1201–1218, 2022.

[58] G. Chen and Q. Wu, "Fundamental limits of intelligent reflecting surface aided multiuser broadcast channel," *IEEE Transactions on Communications*, vol. 71, no. 10, pp. 5904–5919, 2023.

[59] G. Chen, Q. Wu, C. He, W. Chen, J. Tang, and S. Jin, "Active IRS aided multiple access for energy-constrained iot systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1677–1694, 2022.

Thi My Tuyen Nguyen received the B.S. degree in Mathematics from University of Science, Ho Chi Minh City, Viet Nam in 2016, and M.S. degree in Computer Science and Engineering from Soongsil University, South Korea in 2021. She is currently pursuing Ph.D. at Chung-Ang University, South Korea. Her research interests include statistical QoS provisioning, multiple access techniques, and optimization methods in wireless communications.

The Vi Nguyen received the B.S. degree in Mathematics from University of Science, Ho Chi Minh City, Viet Nam in 2016, and M.S. degree in Computer Science and Engineering from Chung-Ang University, South Korea in 2021. He is currently pursuing Ph.D. in School of Computer Science and Engineering at Chung-Ang University, South Korea. His research interests include reconfigurable intelligent surfaces, multiple access techniques, MIMO systems, and optimization methods in wireless communications.

Wonjong Noh received the B.S., M.S., and Ph.D. degrees from the Department of Electronics Engineering, Korea University, Seoul, Korea, in 1998, 2000, and 2005, respectively. From 2005 to 2007, he conducted his postdoctoral research at Purdue University, IN, USA, and the University of California at Irvine, CA, USA. From 2008 to 2014, he was a Principal Research Engineer with Samsung Advanced Institute of Technology, Samsung Electronics, Korea. From 2014 to 2018, he was an Assistant Professor at the Department of Electronics and Communication Engineering, Gyeonggi University of Science and Technology, Korea. He is currently an Professor at the School of Software, College of Information Science, at Hallym University, Korea. His current research interests include fundamental capacity analysis and optimizations in 5G/6G wireless communication and networks, intelligent LEO satellite networks, federated mobile edge computing, machine learning-based systems control, and big-data based healthcare and medical system design. He received a government postdoctoral fellowship from the Ministry of Information and Communication, Korea, in 2005. He was also a recipient of the Samsung Best Paper Gold Award in 2010, the Samsung Patent Bronze Award in 2011, and the Samsung Technology Award in 2013. He has served numerous international conferences as a TPC member or an organizing committee member such as IEEE Globecom, WCNC, ICOIN, ICTC, ICUFN, ICMIC, JCCI and ICAIIC.

Sungrae Cho is a professor with the school of computer sciences and engineering, Chung-Ang University (CAU), Seoul. Prior to joining CAU, he was an assistant professor with the department of computer sciences, Georgia Southern University, Statesboro, GA, USA, from 2003 to 2006, and a senior member of technical staff with the Samsung Advanced Institute of Technology (SAIT), Kiheung, South Korea, in 2003. From 1994 to 1996, he was a research staff member with electronics and telecommunications research institute (ETRI), Daejeon, South Korea. From 2012 to 2013, he held a visiting professorship with the national institute of standards and technology (NIST), Gaithersburg, MD, USA. He received the B.S. and M.S. degrees in electronics engineering from Korea University, Seoul, South Korea, in 1992 and 1994, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2002. He has been a KICS fellow since 2021. He received numerous awards including Haedong Best Researcher of 2022 in Telecommunications and Award of Korean Ministry of Science and ICT in 2021.

His current research interests include wireless networking, network intelligence, and network optimization. He has been an editor-in-chief (EIC) of ICT Express (Elsevier) since 2024, a subject editor of IET Electronics Letter since 2018, an executive editor of Wiley Transactions on Emerging Telecommunications Technologies since 2023, and was an area editor of Ad Hoc Networks Journal (Elsevier) from 2012 to 2017. He has served numerous international conferences as a general chair, TPC chair, or an organizing committee chair, such as IEEE ICC, IEEE SECON, IEEE ICCE, ICOIN, ICTC, ICUFN, APCC, TridentCom, and the IEEE MASS.