Superimposed Pilot and RIS-Aided URLLC: A Joint Design of Phase Shifts and Power Control

Xingguang Zhou, *Graduate Student Member, IEEE*, Wenchao Xia, *Member, IEEE*, Kai-Kit Wong, *Fellow, IEEE*, Hyundong Shin, *Fellow, IEEE*, Hongbo Zhu, *Member, IEEE*

***Abstract*—Suffering from serious rate degradation, how to improve transmission rate with low latency is a challenging issue in ultra-reliable and low-latency communications (URLLC), especially when there is no enough blocklength for data transmission. To handle this issue, we propose to integrate the reconfigurable intelligent surface (RIS) and superimposed pilot (SP) into massive multiple-input multiple-output (mMIMO) systems, where the SP ensures latency by simultaneously sending pilot and data while the RIS improves high transmission rate by reflecting the SP signals. Practically, we derive the finite blocklength ergodic achievable rate lower bound in closed form under imperfect channel estimation and pilot interference removal. Then, we maximize the weighted sum rate of all the users by jointly designing the power control of SP at each user and the phase shifts at the RIS. Due to the highly coupled variables, we first decompose the original problem into the phase shift design subproblem and the power control design subproblem, which are resolved by a genetic algorithm (GA) and an iterative algorithm based on geometric programming (GP). Then, a block coordinate descent algorithm is proposed. Correspondingly, the complexity and convergence of the proposed algorithms are analyzed. Finally, our numerical results demonstrate that the joint design scheme can bring effective rate improvement in stringent latency constraints.**

***Index Terms*—Short-packet transmission, URLLC, massive MIMO, superimposed pilot, reconfigurable intelligent surface (RIS).**

I. Introduction

With the advent of Industry 4.0, smart factories and intelligent manufacturing have become the theme of future industrial development. Undoubtedly, Industrial Internet of Things (IIoT) will be recognized as an important solution to this theme [1], which connects various machines and devices though wireless network and depends on real-time and precise control. Hence, ultra-reliable and low-latency communication (URLLC) is envisioned to establish in the realization of IIoT. As the name suggests, the latency demand of industrial communication is as low as below 1 ms while keeping the ultra reliability between $1-10^{-6}$ and $1-10^{-9}$. Apart from the industrial scenario, other mission-critical applications, e.g., autonomous driving, remote healthcare, and smart grids [2], also have stringent latency and reliability requirements. According to the 3rd generation partnership project (3GPP), the URLLC requirements of a packet with 32 bytes is specified as 1 ms latency and 10^{-5} packet error rate [3].

Subject to the low latency in URLLC, it is impossible to transmit a long data packet as traditional human-to-human communication. Generally, the URLLC applications only exchange short packet data such as measurement data or control commands. Therefore, transmitting short packets in URLLC is called short packet (finite blocklength) transmission.For instance, the packet size of industrial automation is about 10 $\sim$ 300 bytes, and that of smart grid is 80 $\sim$ 1000 bytes [4]. To create a short packet, the information bits are firstly encoded into data symbols. After that, additional symbols are added to the front of data symbols as the overhead for estimating channels and we refer to the total number of symbols as blocklength [5]. However, short packet will inevitably cause a degradation of transmission rate, since the packet error rate becomes non-negligible in the finite blocklength regime. Besides, it should be note that the overhead is also time-consuming, which further decreases the data blocklength in the given latency and aggravates the rate loss.

Reconfigurable intelligent surface (RIS) is a promising technology to support high transmission rate and wide coverage communications, which has drawn extensive attention in academia and industry [6]–[8]. Specifically, composed of numerous passive reflecting elements, RIS is a square panel that can reflect the incident signal. By adjusting each reflecting element with a controller, an independent phase shift is introduced on the incident signal, thus collaboratively reconfiguring the propagation of reflected signal to improve the received signal-to-interference-plus-noise ratio (SINR) level. Additionally, the passive reflecting elements are low-cost and low-energy in comparison with the antennas and radio frequency (RF) chains in traditional multiple-input multiple-output (MIMO) technique. Accordingly, RIS has the potential to compensate the rate loss of in MIMO URLLC systems.

A. Related Work

Presently, only a few works involve the investigation of RIS-aided MIMO URLLC systems [9]–[12]. Particularly, A two-timescale design was proposed for downlink short packet transmission in a RIS-enabled massive MIMO (mMIMO) system [9], where the phase shifts and transmitting power were jointly optimized for maximizing sum rate. Apart from phase

X. Zhou, W. Xia, and H. Zhu are with the Jiangsu Key Laboratory of Wireless Communications, and also with the Engineering Research Center of Health Service System Based on Ubiquitous Wireless Networks, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, (e-mail: 2020010304@njupt.edu.cn, xiawenchao@njupt.edu.cn, zhuhb@njupt.edu.cn).

K. K. Wong is affiliated with the Department of Electronic and Electrical Engineering, University College London, Torrington Place, WC1E 7JE, United Kingdom and he is also affiliated with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17404, Korea, (e-mail: kai-kit.wong@ucl.ac.uk).

H. Shin is affiliated with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104 Korea, (e-mail: hshin@khu.ac.kr).

shifts and transmit power, blocklength optimization was also considered in [10] to study how to maximize sum rate while minimizing the total transmission latency. Moreover, by carefully designing beamforming vector, the sum rate performance can be also improved [11]. Different from [11], [12] studied the reliability performance of URLLC transmission assisted by RIS in device-to-device (D2D) networks.

It is worthy to point out that the above works except for [9] assumed the accurate channel state information (CSI) being available. Indeed, the low-overhead and accurate channel estimation scheme is still an open problem in the field of RIS due to a substantial number of passive elements. Although various methods have been proposed to handle this issue by harnessing the sparsity of RIS channels [13], [14], they are not suitable for URLLC transmission since the complexity will incur high processing delay [15]. The estimation scheme in [9] has low complexity, but the pilot overhead will cause considerable rate degradation in short packet transmission due to the adoption of regular pilot (RP), in which the pilot is transmitted before the data and the pilot length is a linear function concerning user number. Hence, to decrease the pilot overhead while considering imperfect CSI in RIS-aided MIMO URLLC systems, novel channel estimation scheme should be put forward.

Superimposed pilot (SP) can be viewed as an important candidate for MIMO URLLC systems [16]. In the SP scheme, the pilot and data are simultaneously transmitted at the same frequency, which does not spend additional delay sending pilot and thus is very suitable for short packet transmission. Besides, the SP allows sending more pilot and data symbols compared to the RP, which eliminates the pilot overhead and increases the number of connected users. However, the superimposed data and pilot increase the interference in the channel estimation and data detection, respectively. In spite of this, extensive researches have presented the potential of SP on improving spectral efficiency [17]–[19]. However, the merit of low latency in the SP scheme has received little attention. In [20], a low-correlation SP scheme was proposed to enable URLLC communications in the satellite internet of things scenarios. In our previous works [21], it has been verified that the SP outperforms the RP in the finite blocklength transmission, where the number of data symbols in the RP is less than that in the SP under the same blocklength. Nevertheless, as mentioned in Section I, the number of data symbols in a short packet is fixed after encoding. Once the pilot overhead is high and there is no enough blocklength remained for the data transmission, the RP scheme can not finish the transmission in regulated latency. Finally, the benefit of RIS has not been explored in our pervious works, which entails a SP-RIS aided mMIMO system design for URLLC.

B. Contributions

Inspired by the aforementioned research gaps, in this paper, we creatively combine RIS and SP to enable URLLC in mMIMO systems, where the SP is used to ensure latency while the RIS aims to improve transmission rate. Different from the GSP scheme in [22], where the number data symbols was reduced, the SP scheme has higher transmission efficiency. Additionally, in [22], the analysis was based on approximated aggregated channel and infinite blocklength, and the joint design of phase shifts and power control was not considered. The main contributions of this paper are summarized as follows.

- We integrate RIS and SP into mMIMO systems in uplink URLLC scenario, where the blocklength is dedicated for data symbols and there is no extra time-frequency resources reserved for pilot transmission. To reduce processing latency, we estimate the aggregated channel, i.e., the superimposition of direct channel and cascaded RIS channel, instead of estimating them individually, by using the linear minimum mean square error (LMMSE) method, which has a low pilot overhead and complexity. With imperfect CSI, we consider imperfect removal of pilot interference in the data detection process, which further improves the performance of SP.
- We derive the finite blocklength ergodic achievable rate lower bound (LB) of our system in closed form, which is a complicated function of the fixed phase shifts and only depends on the large-scale fading parameters. Based on the derived result, we formulate the weighted sum rate maximization problem, where the SP power control at each user and the phase shifts at the RIS are jointly designed subject to the rate and energy constraints and the range constraint of phase shifts.
- To obtain a solution, we first decouple the original problem into two subproblems. For the phase shift design problem, we propose a method based on genetic algorithm (GA) to figure out it. For the power control design problem, we transform it into a geometric programming (GP) problem by employing the log-function approximation and the successive convex approximation (SCA) methods. Then, an efficient iterative algorithm is proposed to solve a series of GPs. Finally, to alternately optimize the power and the phase shifts, a block coordinate descent (BCD) algorithm is proposed and its convergence and complexity are also analyzed.
- We evaluate numerically the superiority of proposed joint design scheme compared to other baseline schemes. The results indicate that our design scheme can realize high transmission rate in the finite blocklength regime and the joint optimization is indispensable when the RIS is integrated with the SP.

Notation: The column vectors and matrices in this paper are denoted by lower-case bold-face letters and bold-face capital letters, respectively. $\mathbf{x}^H$, $\mathbf{X}^H$, and $\mathbf{X}^{-1}$ represent the conjugate transpose of vector $\mathbf{x}$ and matrix $\mathbf{X}$ and the inverse matrix of $\mathbf{X}$, respectively. $\mathbb{E}\{\cdot\}$, $\|\cdot\|$, $\mathrm{Re}\{\cdot\}$, and $\mathrm{Cov}\{\cdot\}$ denote the expectation, Euclidean norm, real part, and covariance operators, respectively. $\mathcal{CN}(\cdot,\cdot)$ and $\mathbb{C}$ stand for the complex Gaussian distribution and complex number set, respectively.

II. System Model

Considering a uplink URLLC scenario, an RIS-aided mMIMO system as shown in Fig. 1, where a base station (BS)

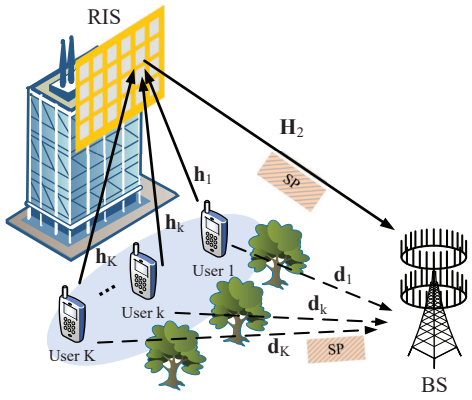

Fig. 1. An SP-RIS-aided mMIMO system with direct links.

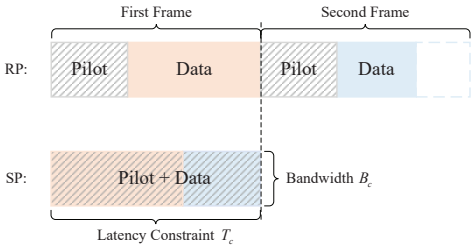

Fig. 2. The frame structure of RP and SP schemes. Given the bandwidth B_c and the latency constraint T_c, the maximal available blocklength is $\tau_c = T_c B_c$.

configured with M antennas simultaneously receives short packets from K ($K < M$) single-antenna users with the assistance of the RIS. Specifically, to enhance the channel strength of the direct links, the RIS reflects the incoming signal waves from the users to the BS by N passive reflecting elements, after which phase shifts are introduced and the corresponding matrix with unit amplitude is given by $\mathbf{\Phi} = diag\left\{e^{j\theta_1}, e^{j\theta_2}, \ldots, e^{j\theta_N}\right\}$, where $\theta_n \in [0, 2\pi)$ is the phase shift of element n and can be adjusted by the BS.

A. Frame Structure

Given the bandwidth B_c, the blocklength determines the transmission latency in URLLC [23]. We assume that the maximal available blocklength is τ_c and the latency equal to $T_c = \frac{\tau_c}{B_c}$, which is insufficient to finish the data transmission sometimes. As shown in Fig. 2, to transmit τ_c data symbols, the conventional RP scheme must use two frames [24], which violates the latency requirement T_c. Thus, the SP scheme is exploited in this paper, where the pilot and data of the same length are transmitted in the same time slot and bandwidth. This scheme saves transmission latency at the expense of a reduction in the SINR, since the data and pilot are regarded as interference in channel estimation and data detection, respectively. Thus, the SP scheme offers a tradeoff between low latency and achievable rate.

B. Channel Model

To establish robust communication link, the RIS is generally installed on a position with a high altitude when a few scatters exist in ground communication links. Therefore, the RIS-related channels include line-of-sight (LoS) components and the Rician fading model is adopted to describe the channels[1]. Let α_k and β denote the large-scale path loss of the user-k-RIS link and RIS-BS link, respectively. Let ϵ_k and δ represent the the Rician factors of the user-k-RIS link and RIS-BS link, respectively. Then, the $N \times 1$ user-k-RIS channel $\mathbf{h}_k$ and $M \times N$ RIS-BS channels $\mathbf{H}_2$ can be respectively expressed as

$$\mathbf{h}_k = \sqrt{\frac{\alpha_k}{\varepsilon_k + 1}}\left(\sqrt{\varepsilon_k}\bar{\mathbf{h}}_k + \tilde{\mathbf{h}}_k\right), \tag{1}$$

$$\mathbf{H}_2 = \sqrt{\frac{\beta}{\delta + 1}}\left(\sqrt{\delta}\bar{\mathbf{H}}_2 + \tilde{\mathbf{H}}_2\right), \tag{2}$$

where $\tilde{\mathbf{h}}_k \in \mathbb{C}^{N\times 1}$ and $\tilde{\mathbf{H}}_2 \in \mathbb{C}^{M\times N}$ represent the non-line-of-sight (NLoS) components with elements following the distribution of $\mathcal{CN}(0,1)$. Without loss of generality, assuming that the uniform square planar arrays (USPA) is employed at the BS and the RIS. Define steering vector as $\mathbf{a}_X(\vartheta^a, \vartheta^e)$ with $0 \le x, y \le \sqrt{X} - 1$ and $X \in \{M, N\}$, which is given by

$$\mathbf{a}_X(\vartheta^a, \vartheta^e) = \Big[1, \ldots, e^{j2\pi\frac{d}{\lambda}(x\sin\vartheta^e\sin\vartheta^a + y\cos\vartheta^e)}, \ldots, e^{j2\pi\frac{d}{\lambda}\left(\left(\sqrt{X}-1\right)\sin\vartheta^e\sin\vartheta^a + \left(\sqrt{X}-1\right)\cos\vartheta^e\right)}\Big]^T, \tag{3}$$

where d and λ are the element spacing and carrier wavelength, respectively. Then, the deterministic LoS components $\bar{\mathbf{h}}_k \in \mathbb{C}^{N\times 1}$ and $\bar{\mathbf{H}}_2 \in \mathbb{C}^{M\times N}$ can be further written as

$$\bar{\mathbf{h}}_k = \mathbf{a}_N\left(\varphi^a_{kr}, \varphi^e_{kr}\right), \tag{4}$$

$$\bar{\mathbf{H}}_2 = \mathbf{a}_M\left(\varphi^a_r, \varphi^e_r\right)\mathbf{a}_N^H\left(\varphi^a_t, \varphi^e_t\right), \tag{5}$$

where φ^a_{kr} and φ^e_{kr} are the azimuth and elevation angles of arrival (AoA) at the RIS from user k, respectively. Similarly, φ^a_r and φ^e_r are the AoA at the BS from the RIS, respectively. φ^a_t and φ^e_t are the azimuth and elevation angles of departure (AoD) from the RIS towards the BS, respectively.[2]

Different from the RIS-related channels, the channels between the BS and users undergo a surrounding with rich scatters. Thus, considering the Rayleigh fading model for the direct links is reasonable. Let us denote $\mathbf{d}_k = \sqrt{\gamma_k}\tilde{\mathbf{d}}_k$ as the channel between the k-th user and the BS, where γ_k and $\tilde{\mathbf{d}}_k$ are the large-scale path loss and the NLoS component of the user-k-BS link following the distribution of $\mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$, respectively.

Based on the above definitions, the aggregated channel of the k-th user can be represented as

$$\mathbf{f}_k = \mathbf{d}_k + \mathbf{g}_k, \tag{6}$$

where $\mathbf{g}_k = \mathbf{H}_2\mathbf{\Phi}\mathbf{h}_k$ is the cascaded channel of the k-th user.

[1]We treat the phase of the LoS component as a deterministic constant that is perfectly known by receiver and included into the Rician factor.

[2]The LoS components are deterministic because they only rely on the large-scale parameters AoA and AoD. In URLLC scenarios, the transmission latency is much shorter than the channel coherence time such that the large-scale parameters hardly change and can be obtained by measuring in advance. Thus, all the angles, path-loss factors, and Rician factors are regarded as known constants in this paper.

III. Channel Estimation and Achievable Rate

Since the RIS does not have the ability of signal processing, the process of channel estimation and data detection must be successively performed at the BS after receiving the SP signals. However, the pilot overhead will sharply increase when estimating the RIS-BS channels due to the extremely high dimension, which causes high system complexity and processing delay. Accordingly, instead of estimating the channels individually, we directly estimate the aggregated channel by using the traditional method in mMIMO systems with fixed phase shifts and power allocation [9], [25]–[28].

A. Channel Estimation

To distinguish the aggregated channels of different users, orthogonal pilot transmission scheme is employed among the K users ($K \leq \tau_c$). Specifically, Let $\boldsymbol{\psi}_i \in \mathbb{C}^{\tau_c \times 1}$ and $\mathbf{s}_i \in \mathbb{C}^{\tau_c \times 1}$ is the pilot and data of the i-th user, respectively, with the orthogonality $\boldsymbol{\psi}_i^H \boldsymbol{\psi}_i = \tau_c$ and $\boldsymbol{\psi}_i^H \boldsymbol{\psi}_j = 0, i \neq j$ satisfied. Then, the received signal at the BS is given by

$$\mathbf{Y} = \sum\nolimits_{i=1}^{K} \mathbf{f}_i \left(\sqrt{q_i} \boldsymbol{\psi}_i^H + \sqrt{p_i} \mathbf{s}_i^H \right) + \mathbf{W}, \tag{7}$$

where q_i and p_i are the normalized transmitting power on the pilot and the data of the i-th user, respectively, $\mathbf{s}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{\tau_c})$, $\mathbf{W} \in \mathbb{C}^{M \times \tau_c}$ is the receiver noise matrix whose entries are independently and identically distributed (i.i.d) random variables following the distribution of $\mathcal{CN}(0,1)$.

By multiplying $\mathbf{Y}$ with $\boldsymbol{\psi}_k / \sqrt{\tau_c}$, the observation vector for the aggregated channel of the k-th user is obtained, as follows

$$\begin{aligned} \mathbf{y}_k &= \mathbf{Y} \frac{\boldsymbol{\psi}_k}{\sqrt{\tau_c}} \\ &= \sqrt{q_k \tau_c} \mathbf{f}_k + \underbrace{\sum\nolimits_{i=1}^{K} \sqrt{\frac{p_i}{\tau_c}} \mathbf{f}_i \mathbf{s}_i^H \boldsymbol{\psi}_k + \mathbf{W} \frac{\boldsymbol{\psi}_k}{\sqrt{\tau_c}}}_{\text{effective noise}}, \end{aligned} \tag{8}$$

where $\mathbf{W} \boldsymbol{\psi}_k / \sqrt{\tau_c} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. It should be noted that both the aggregated channel and effective noise term are not Gaussian owing to the presence of the product of two complex Gaussian variables. Besides, the effective noise term is not independent of the aggregated channel $\mathbf{f}_k$ but only uncorrelated. Accordingly, it is hard to obtain the closed-form expression of the optimal minimum mean square estimate (MMSE) [29]. In this case, the suboptimal estimator LMMSE [29] is used as an alternative, which is given by the following lemma [16].

Lemma 1*: Applying the LMMSE estimator to the observation vector (8), then the estimate of* $\mathbf{f}_k$ *is given by*

$$\hat{\mathbf{f}}_k = c_k \mathbf{y}_k, \tag{9}$$

where

$$c_k = \frac{\sqrt{q_k \tau_c} \left(\xi_k + \gamma_k \right)}{q_k \tau_c \left(\xi_k + \gamma_k \right) + \sum_{i=1}^{K} p_i \left(\xi_i + \gamma_i \right) + 1}, \tag{10}$$

$$\xi_k = \frac{\beta \alpha_k}{(\delta + 1)(\varepsilon_k + 1)} \left(\delta \varepsilon_k \left| f_k(\boldsymbol{\Phi}) \right|^2 + (\delta + \varepsilon_k + 1) N \right), \tag{11}$$

$$f_k(\boldsymbol{\Phi}) \triangleq \mathbf{a}_N^H \left(\varphi_t^a, \varphi_t^e \right) \boldsymbol{\Phi} \bar{\mathbf{h}}_k. \tag{12}$$

Denote the estimation error by $\mathbf{e}_k = \mathbf{f}_k - \hat{\mathbf{f}}_k$*. It can be readily proved that the error* $\mathbf{e}_k$ *is uncorrelated of the estimate* $\hat{\mathbf{f}}_k$*. Then, the covariance matrices of* $\hat{\mathbf{f}}_k$ *and* $\mathbf{e}_k$ *are, respectively,*

$$\mathbb{E}\left\{ \hat{\mathbf{f}}_k \hat{\mathbf{f}}_k^H \right\} = \eta_k \mathbf{I}_M, \tag{13}$$

$$\mathbb{E}\left\{ \mathbf{e}_k \mathbf{e}_k^H \right\} = \left(\xi_k + \gamma_k - \eta_k \right) \mathbf{I}_M, \tag{14}$$

where $\eta_k = \sqrt{q_k \tau_c} \left(\xi_k + \gamma_k \right) c_k$.

Proof: Please refer to Appendix A.

Remark 1*:* When the channels $\mathbf{H}_2$ and $\mathbf{h}_k$ are estimated individually, the pilot overhead grows proportionally with N. From Lemma 1, we can find that only K aggregated channels need to be estimated, which reduces the pilot overhead and processing delay. Moreover, the estimate is not only related to the pilot power but also the data power. Based on (14), the mean square error (MSE) of the k-th user can be calculated as $\mathbf{MSE}_k = \frac{M(\xi_k + \gamma_k)(\sum_{i=1}^{K} p_i (\xi_i + \gamma_i) + 1)}{q_k \tau_c (\xi_k + \gamma_k) + \sum_{i=1}^{K} p_i (\xi_i + \gamma_i) + 1}$, which shows that the increase of data power will impair the channel estimation quality. Consequently, it is essential to optimize the power control in SP.

B. Achievable Rate

To decrease the system complexity, the low-complexity maximal-ratio-combining (MRC) detector is utilized. With the SP context, the pilot interference is inevitable but can be reduced since the BS knows the pilot assignment. Then, after removing the pilot interference, the signal is detected and can be expressed as

$$\begin{aligned} \hat{\mathbf{s}}_k^H &= \hat{\mathbf{f}}_k^H \left(\mathbf{Y} - \sum\nolimits_{i=1}^{K} \sqrt{q_i} \hat{\mathbf{f}}_i \boldsymbol{\psi}_i^H \right) \\ &= \sum\nolimits_{i=1}^{K} \sqrt{q_i} \hat{\mathbf{f}}_k^H \mathbf{e}_i \boldsymbol{\psi}_i^H + \sum\nolimits_{i=1}^{K} \sqrt{p_i} \hat{\mathbf{f}}_k^H \mathbf{f}_i \mathbf{s}_i^H + \hat{\mathbf{f}}_k^H \mathbf{W}. \end{aligned} \tag{15}$$

Obviously, the residual pilot interference results from the channel estimation error. Next, we derive the closed-form SINR expression based on (15)[3]. Using the use-and-then-forget technique [30], (15) can be rewritten as

$$\begin{aligned} \hat{\mathbf{s}}_k^H = &\sqrt{p_k} \eta_k \left(\xi_k + \gamma_k \right)^{-1} \mathbb{E}\left\{ \mathbf{f}_k^H \mathbf{f}_k \right\} \mathbf{s}_k^H \\ &+ \sqrt{p_k} \eta_k \left(\xi_k + \gamma_k \right)^{-1} \left(\mathbf{f}_k^H \mathbf{f}_k - \mathbb{E}\left\{ \mathbf{f}_k^H \mathbf{f}_k \right\} \right) \mathbf{s}_k^H + \mathbf{z}_k^H, \end{aligned} \tag{16}$$

where the non-Gaussian noise term $\mathbf{z}_k^H$ is given by

$$\begin{aligned} \mathbf{z}_k^H = &\sqrt{p_k} \bar{\mathbf{f}}_k^H \mathbf{f}_k \mathbf{s}_k^H + \sum\nolimits_{i \neq k}^{K} \sqrt{p_i} \hat{\mathbf{f}}_k^H \mathbf{f}_i \mathbf{s}_i^H \\ &+ \sum\nolimits_{i=1}^{K} \sqrt{q_i} \hat{\mathbf{f}}_k^H \mathbf{e}_i \boldsymbol{\psi}_i^H + \hat{\mathbf{f}}_k^H \mathbf{W}, \end{aligned} \tag{17}$$

with $\bar{\mathbf{f}}_k = \hat{\mathbf{f}}_k - c_k \sqrt{q_k \tau_c} \mathbf{f}_k$. Note that $\mathbf{z}_k^H$ is uncorrelated with the other terms in (16). Based on the [30, Chapter. 2], the RIS-aided MIMO system in Section II can be regarded as a single-input single-output (SISO) system with fading channel and non-Gaussian effective noise at each user, where the effective SINR at the k-th user is formulated as

$$\Gamma_k = \frac{p_k \tau_c \left| \frac{\eta_k}{\xi_k + \gamma_k} \mathbb{E}\{ \mathbf{f}_k^H \mathbf{f}_k \} \right|^2}{p_k \tau_c \mathrm{Var}\left(\frac{\eta_k}{\xi_k + \gamma_k} \mathbf{f}_k^H \mathbf{f}_k \right) + \mathbb{E}\{ \| \mathbf{z}_k^H - \mathbb{E}\{ \mathbf{z}_k^H \} \|^2 \}}. \tag{18}$$

[3]Since the non-Gaussian nature, it is difficult to acquire the closed-form SINR expressions for MMSE and zero-forcing (ZF) detectors. Thus, we focus on MRC detector and the cases of MMSE and ZF are left for future work.

According to [31], the URLLC ergodic achievable rate can be given by

$$R_k = \mathbb{E}\left\{\log_2\left(1+\varpi_k\right) - \frac{Q^{-1}\left(\epsilon\right)}{\ln 2\sqrt{\tau_c}}\sqrt{V\left(\varpi_k\right)}\right\}, \tag{19}$$

where the channel variations between different fading blocks are averaged, ϖ_k, ϵ, and $Q^{-1}\left(\cdot\right)$ denote the instantaneous SINR for the k-th user, decoding error rate, and the inverse of the Gaussian Q-function, respectively, and $V\left(\varpi_k\right) = 1 - \frac{1}{\left(1+\varpi_k\right)^2}$. However, deriving the exact expression of (19) is difficult. Instead, we use Jensen's inequality and Lemma 1 in [32] to obtain the LB of (19), which is shown in the following theorem:

Theorem 1*: The URLLC ergodic achievable rate of the k-th user in the SP-RIS-aided mMIMO system is lower bounded by*

$$R_k \geq \hat{R}_k = \frac{1}{\ln 2} f\left(\mathbb{E}\left\{\left(\varpi_k\right)^{-1}\right\}\right) = \frac{1}{\ln 2} f\left(\frac{1}{\Gamma_k}\right), \tag{20}$$

where $f\left(x\right) = \ln\left(1+\frac{1}{x}\right) - \frac{Q^{-1}(\epsilon)}{\sqrt{\tau_c}}\sqrt{V\left(\frac{1}{x}\right)} \geq 0, x > 0$, and the closed-form SINR expression Γ_k is shown by (21) at the top of next page with

$$\begin{aligned}\Upsilon_k = &\left(\frac{\beta\alpha_k}{\left(\delta+1\right)\left(\varepsilon_k+1\right)}\right)^2 \\ &\times\left\{2\delta\varepsilon_k\left|f_k\left(\mathbf{\Phi}\right)\right|^2\left(\delta MN + MN + 2\right)\right. \\ &+ MN^2\left(\delta^2 + 2\delta\varepsilon_k + 2\delta + 2\varepsilon_k + 1\right) \\ &\left.+N\left(2\delta + 2\varepsilon_k + 1\right)\right\},\end{aligned} \tag{22}$$

$$\begin{aligned}\chi_k = &\left(\frac{\beta\alpha_k}{\left(\delta+1\right)\left(\varepsilon_k+1\right)}\right)^2 \\ &\times\left\{2\delta\varepsilon_k\left|f_k\left(\mathbf{\Phi}\right)\right|^2\left(\delta MN + 2M + N\varepsilon_k + N\right)\right. \\ &+ N^2\left(2\delta\varepsilon_k + 2\delta + 2\varepsilon_k + M\delta^2 + 1\right) \\ &\left.+MN\left(2\delta + 2\varepsilon_k + 1\right)\right\},\end{aligned} \tag{23}$$

$$\begin{aligned}\Xi'_{ki} = &\frac{\beta^2\alpha_i\alpha_k}{\left(\delta+1\right)^2\left(\varepsilon_i+1\right)\left(\varepsilon_k+1\right)} \\ &\times\left(\varepsilon_k\varepsilon_i\left|\bar{\mathbf{h}}_i^H\bar{\mathbf{h}}_k\right|^2 + N\left(\varepsilon_i+\varepsilon_k+1\right)\right. \\ &\left.+2\delta\varepsilon_k\left|f_k\left(\mathbf{\Phi}\right)\right|^2 + 2\delta\varepsilon_i\left|f_i\left(\mathbf{\Phi}\right)\right|^2 + 2N\delta\right),\end{aligned} \tag{24}$$

$$\begin{aligned}\Omega_{ki} = &\frac{\beta^2\alpha_i\alpha_k}{(\delta+1)^2(\varepsilon_i+1)(\varepsilon_k+1)} \\ &\times\left\{M\delta^2\varepsilon_k\varepsilon_i|f_k(\mathbf{\Phi})|^2|f_i(\mathbf{\Phi})|^2\right. \\ &+\delta\varepsilon_k|f_k(\mathbf{\Phi})|^2(\delta MN + N\varepsilon_i + N + 2M) \\ &+\delta\varepsilon_i|f_i(\mathbf{\Phi})|^2(\delta MN + N\varepsilon_k + N + 2M) \\ &+N^2\left(M\delta^2 + \delta\left(\varepsilon_i+\varepsilon_k+2\right) + \left(\varepsilon_k+1\right)\left(\varepsilon_i+1\right)\right) \\ &+MN\left(2\delta+\varepsilon_i+\varepsilon_k+1\right) + M\varepsilon_k\varepsilon_i\left|\bar{\mathbf{h}}_k^H\bar{\mathbf{h}}_i\right|^2 \\ &\left.+2M\delta\varepsilon_k\varepsilon_i\mathrm{Re}\left\{f_k^H(\mathbf{\Phi})f_i(\mathbf{\Phi})\bar{\mathbf{h}}_i^H\bar{\mathbf{h}}_k\right\}\right\},\end{aligned} \tag{25}$$

$$\Psi_{ij} = \left(\xi_i+\gamma_i\right)\left(\xi_j+\gamma_j\right) - \xi_i\xi_j. \tag{26}$$

Proof: Please refer to Appendix B.

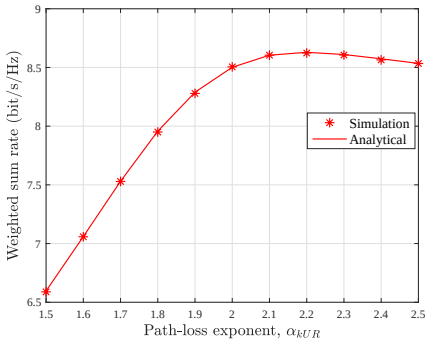

Fig. 3. Accuracy of the derived SINR expression in (21): $M = 64$, $N = 16$, $K = 4$, $\tau_c = 100$, $E = 13$dB, $\kappa = 0.3$, other parameters see Section V.

We validate the accuracy of the derived SINR expression in Fig. 3, where 10000 random channel generations are averaged based on the Monte-Carlo method and the analytical result is obtained by (21). We find that the analytical result is tight for the simulation result for any path-loss exponent, which verifies the SINR expression derived in Theorem 1 is accurate.

Remark 2*:* With SP, it can be seen that there is no rate loss due to pre-log overhead in (20). Theorem 1 provides the closed-form rate expression determined by the large-scale fading parameters. As discussed in Section II, these parameters are almost unchanged during the transmission time, which suffices long-term optimization for the RIS phase shifts and the SP power control.

IV. Optimization Problem and Proposed Solution

In this section, we jointly design the phase shifts $\mathbf{\Phi}$ and pilot power $\mathbf{q} = [q_1, q_2, \ldots q_K]$ and data power $\mathbf{p} = [p_1, p_2, \ldots p_K]$ for the SP-RIS-aided mMIMO system based on the closed-form results given in Section III.

A. Problem Formulation

As mentioned in Section II.A, by using SP scheme, the blocklength (latency) is restricted to τ_c symbols. Thus, under given decoding error rate ϵ, the URLLC is ensured. With the goal of maximizing the weighted sum rate of all the users, the optimization problem is formulated as

$$\mathcal{P}1: \max_{\mathbf{\Phi},\mathbf{p},\mathbf{q}} \sum\nolimits_{k=1}^{K} w_k\hat{R}_k \tag{27a}$$

$$\text{s.t. } \hat{R}_k \geq 0, \ \forall k, \tag{27b}$$

$$\tau_c\left(p_k+q_k\right) \leq E, \ \forall k, \tag{27c}$$

$$\theta_n \in [0, 2\pi), \ \forall n, \tag{27d}$$

where w_k and E denote the priority of k-th user and the energy limitation at each user, respectively. Clearly, $\mathcal{P}1$ is non-convex and non-linear and the variables $\mathbf{\Phi}$, $\mathbf{p}$, and $\mathbf{q}$ are coupled. In addition, the rate constraints (27b), energy constraints (27c), and phase shifts constraints (27d) make the problem more complicated.

We propose an efficient algorithm based on the BCD method to tackle this problem [33]. The basic ideal of which is that the original problem $\mathcal{P}1$ is decoupled into two subproblems, which are solved alternately. Specifically, $\mathbf{\Phi}$ is optimized with given fixed $\mathbf{p}$ and $\mathbf{q}$, and in turn, suboptimal $\mathbf{p}$ and $\mathbf{q}$ are

$$\Gamma_k = \frac{M\tau_c p_k q_k \left(\xi_k+\gamma_k\right)^2}{\begin{array}{c}\tau_c p_k q_k\left(\Upsilon_k - M\xi_k^2\right)+\tau_c q_k\sum\limits_{i=1}^{K} p_i\left(\Omega_{ki}+\Psi_{ik}\right)+\sum\limits_{i=1}^{K} p_i^2\Upsilon_i+\frac{1}{\tau_c}\sum\limits_{i=1}^{K} p_i^2\left(M\tau_c\gamma_i^2+2M\tau_c\gamma_i\xi_i+2\gamma_i\xi_i+\gamma_i^2+\chi_i\right)\\ +\frac{1}{\tau_c}\sum\limits_{i=1}^{K}\sum\limits_{j=1}^{K} p_i p_j\Xi'_{ij}+\sum\limits_{i=1}^{K}\sum\limits_{j=1}^{K} p_i p_j\left(\Omega_{ij}+\Psi_{ij}\right)+\left(q_k\tau_c\left(\xi_k+\gamma_k\right)+\sum\limits_{i=1}^{K} p_i\left(\xi_i+\gamma_i\right)+1\right)\left(\sum\limits_{i=1}^{K} q_i\left(\xi_i+\gamma_i-\eta_i\right)+1\right)\end{array}}, \tag{21}$$

found when $\mathbf{\Phi}$ is fixed, where the suboptimal value of (27a) is updated iteratively until convergence. The above details are shown in the following.

B. Phase Shifts Design

With fixed power allocation $\mathbf{p}$ and $\mathbf{q}$, the subproblem to optimize $\mathbf{\Phi}$ can be formulated as

$$\mathcal{P}2: \max_{\mathbf{\Phi}} \sum\nolimits_{k=1}^{K} w_k\hat{R}_k \tag{28a}$$
$$\text{s.t. } \hat{R}_k \geq 0,\ \forall k, \tag{28b}$$
$$\theta_n \in [0, 2\pi),\ \forall n. \tag{28c}$$

Recalling (20), the objective function (28a) is such complex that traditional optimization methods, such as semidefinite relaxation (SDR) and minorization-maximization (MM) [34], are incapable for solving $\mathcal{P}2$. Fortunately, GA is an optimization algorithm based on natural selection and genetics principle [35], which is very suitable for this kind of intractable problem as $\mathcal{P}2$. Additionally, benefiting from the strong parallelism, GA can accelerate the speed of search through parallel computation. Hence, we solve $\mathcal{P}2$ by means of GA [36].

To initialize GA, a population with L individuals is generated. Each individual contains N chromosomes, composing matrix $\mathbf{\Phi}_t$, and the n-th chromosome corresponds to the RIS phase shift θ_n. The objective function (28a) is used as the fitness evaluation function to compute the fitness of each individual in the current population [38]. Those with high fitness are retained as elites to the next generation, those with low fitness experience mutation operation to generate offspring, and those with medium fitness are used to generate parents, and then cross the parents to generate offspring [37]. The detailed steps are summarized in Algorithm 1. It should be emphasized that we set those negative rates as zero when computing the raw fitness in each iteration such that the constraint (28b) is guaranteed.

Note that the complexity of Algorithm 1 mainly depends on the fitness calculation in step 3 and can be approximated as $\mathcal{O}\left(n_1^{iter} n_1^{cost} L\right)$ [39], where n_1^{iter} is the total number of iterations, n_1^{cost} is the complexity of calculating the fitness function, and L is the size of population.

Algorithm 1 Phase Shifts Design Based on GA

1: Initialize $L = L_e + L_m + L_p$ individuals in a population by generating a chromosome $\mathbf{\Phi}_t$ randomly for each one; iteration number $s = 1$.
2: **while** $s \leq 100 * L$ **do**
3: Use the objective function (28a) to compute the raw fitness of each individual and sort them in a descending order; scale the raw fitness based on [38, eq. (33)].
4: Select the top L_e individuals with higher fitness as elites from current population.
5: Select L_m individuals with lower fitness from the current population and produce L_m offspring based on uniform mutation [38] with probability p_m.
6: Select $2L_p$ parents from the remaining individuals based on stochastic universal sampling and then single-point crossover is used to create L_p offspring [38].
7: Generate the next generation population by combining the L_e elites, $L_m + L_p$ offspring; $s = s + 1$.
8: **end while**
9: **Output:** the highest fitness and the corresponding chromosome of individual in the current population.

C. Power Control Design

1) Proposed Algorithm: By fixing the phase shift matrix $\mathbf{\Phi}$, the subproblem to optimize $\mathbf{p}$ and $\mathbf{q}$ is written as

$$\mathcal{P}3: \max_{\mathbf{p},\mathbf{q}} \sum\nolimits_{k=1}^{K} w_k\hat{R}_k \tag{29a}$$
$$\text{s.t. } \hat{R}_k \geq 0,\ \forall k, \tag{29b}$$
$$\tau_c\left(p_k + q_k\right) \leq E,\ \forall k. \tag{29c}$$

Due to the complicated expression of $\hat{R}_k$, we simplify the objective function (29a) and constraint (29b) in the following. To this end, we introduce Lemma 2 as follows

Lemma 2*: $\hat{R}_k$ is a monotonically increasing function of SINR Γ_k in the region of $\hat{R}_k \geq 0$.*

Proof: Take the derivative w.r.t. Γ_k in (20), which yields $-\frac{1}{\Gamma_k^2 \ln 2} f'(\frac{1}{\Gamma_k})$. Recalling Theorem 1, the feasible region of $f\left(\Gamma_k^{-1}\right)$ is restricted in $0 < 1/\Gamma_k \leq g^{-1}\left(Q^{-1}(\epsilon)/\sqrt{\tau_c}\right)$ due to $f\left(\Gamma_k^{-1}\right) \geq 0$, where $g(x) = \frac{(x+1)\ln(1+1/x)}{\sqrt{2x+1}}$ [32]. Applying Lemma 1 in [32], we have $f'(\frac{1}{\Gamma_k}) \leq 0$. Hence, the derivative of Γ_k is equal and greater than zero.

Based on Lemma 2, the constraint (29b) can be equivalent to the SINR constraint as follows

$$\Gamma_k \geq 1/f^{-1}(0). \tag{30}$$

Next, we focus on the objective function (29a). First of all, by introducing auxiliary variables $\boldsymbol{\sigma}=[\sigma_1,\sigma_2\dots\sigma_K]$, $\mathcal{P}3$ can be relaxed as

$$\mathcal{P}4:\max_{\boldsymbol{\sigma},\mathbf{p},\mathbf{q}}\sum_{k=1}^{K}\frac{w_k}{\mathrm{ln}2}\left[\mathrm{ln}\left(1+\sigma_k\right)-\frac{Q^{-1}\left(\epsilon\right)}{\sqrt{\tau_c}}\sqrt{1-\frac{1}{\left(1+\sigma_k\right)^2}}\right] \tag{31a}$$

$$\text{s.t. } \Gamma_k\geq\sigma_k,\ \forall k, \tag{31b}$$

$$\sigma_k\geq 1/f^{-1}\left(0\right),\ \forall k, \tag{31c}$$

$$\tau_c\left(p_k+q_k\right)\leq E,\ \forall k. \tag{31d}$$

To prove that $\mathcal{P}3$ is equivalent to $\mathcal{P}4$, we utilize the contradiction method as follows. Assuming that $\{\sigma_k,p_k,q_k,\forall k\}$ is the optimal solution of $\mathcal{P}4$ and $\Gamma_k>\sigma_k$. Obviously, the value of objective function (31a) can be increased since it is a monotonically increasing function of σ_k according to Lemma 2. Hence, it is contradictory with the optimal solution and the optimal solution should satisfied $\Gamma_k=\sigma_k$. Then, the objective function (31a) is equivalent to (29a). $\mathcal{P}3$ and $\mathcal{P}4$ have the same power allocation solutions and the same objective function value. However, the objective function (31a) is still in an intractable form and needs to be further simplified. Then, we show the following lemmas [32]:

Lemma 3*: Let* $U\left(m\right)=\sqrt{1-\frac{1}{\left(1+m\right)^2}}$ *and* $\tilde{m}\geq\frac{\sqrt{17}-3}{4}$, $\forall m\geq\frac{\sqrt{17}-3}{4}$*, we have the following inequality*[4]*:*

$$J\left(m\right)=\vartheta\mathrm{ln}\left(m\right)+\nu\geq U\left(m\right), \tag{32}$$

where ϑ *and* ν *are respectively given by*

$$\vartheta=\frac{\tilde{m}}{\sqrt{\tilde{m}^2+2\tilde{m}}}-\frac{\tilde{m}\sqrt{\tilde{m}^2+2\tilde{m}}}{\left(1+\tilde{m}\right)^2}, \tag{33}$$

$$\nu=\sqrt{1-\frac{1}{\left(1+\tilde{m}\right)^2}}-\vartheta\mathrm{ln}\left(\tilde{m}\right). \tag{34}$$

In particular, we have $J'\left(\tilde{m}\right)=U'\left(\tilde{m}\right)$, $J\left(\tilde{m}\right)=U\left(\tilde{m}\right)$*, which means that the upper bound (32) is tight when* $m=\tilde{m}$*.*

Proof: Note that $\frac{\sqrt{17}-3}{4}$ comes from taking the derivation of the constructed function in the discussion of monotonicity. The details can be referred to Appendix C in [32].

Lemma 4*: Let* $\tilde{m}\geq 0$, $\forall m\geq 0$*, we have the following inequality:*

$$\hat{\vartheta}\mathrm{ln}\left(m\right)+\hat{\nu}\leq\mathrm{ln}\left(1+m\right), \tag{35}$$

where $\hat{\vartheta}$ *and* $\hat{\nu}$ *are respectively expressed as*

$$\hat{\vartheta}=\frac{\tilde{m}}{1+\tilde{m}},\ \hat{\nu}=\mathrm{ln}\left(1+\tilde{m}\right)-\frac{\tilde{m}}{1+\tilde{m}}\mathrm{ln}\left(\tilde{m}\right). \tag{36}$$

Particularly, (35) is tight at $m=\tilde{m}$*.*

Proof: The proof is similar to that of Lemma 3 and thus is omitted.

In the following, we aim to design an iterative algorithm, where the objective function (31a) is approximated via using Lemmas 3 and 4 in each iteration. Denote $\hat{\vartheta}_k^{(t)}$, $\vartheta_k^{(t)}$, $\hat{\nu}_k^{(t)}$, $\nu_k^{(t)}$, and $\sigma_k^{(t)}$ as the values of $\hat{\vartheta}_k$, ϑ_k, $\hat{\nu}_k$, ν_k, and σ_k in the t-th iteration, respectively. Then, the LB of (31a) in the $t+1$-th iteration is given by

$$\begin{aligned}&\sum_{k=1}^{K}\frac{w_k}{\mathrm{ln}2}\left[\mathrm{ln}\left(1+\sigma_k\right)-\frac{Q^{-1}\left(\epsilon\right)}{\sqrt{\tau_c}}U\left(\sigma_k\right)\right]\\&\geq\sum_{k=1}^{K}\frac{w_k}{\mathrm{ln}2}\left[\hat{\vartheta}_k^{(t)}\mathrm{ln}\sigma_k+\hat{\nu}_k^{(t)}-\frac{Q^{-1}\left(\epsilon\right)}{\sqrt{\tau_c}}\left(\vartheta_k^{(t)}\mathrm{ln}\sigma_k+\nu_k^{(t)}\right)\right],\end{aligned} \tag{37}$$

where $\vartheta_k^{(t)}$, $\nu_k^{(t)}$, $\hat{\vartheta}_k^{(t)}$, $\hat{\nu}_k^{(t)}$ are obtained by substituting $\tilde{m}=\sigma_k^{(t)}$ into (33), (34), and (36). Note that the equality holds when $\sigma_k=\sigma_k^{(t)}$ due to the tightness in Lemma 3 and Lemma 4. Then, by replacing (31a) with (37) and neglecting the constant terms, $\mathcal{P}4$ can be turned into the following problem:

$$\mathcal{P}5:\max_{\boldsymbol{\sigma},\mathbf{p},\mathbf{q}}\sum\nolimits_{k=1}^{K}\lambda_k^{(t)}\mathrm{ln}\sigma_k \tag{38a}$$

$$\text{s.t. } (31b)-(31d), \tag{38b}$$

where $\lambda_k^{(t)}=\frac{w_k}{\mathrm{ln}2}(\hat{\vartheta}_k^{(t)}-\frac{Q^{-1}(\varepsilon)}{\sqrt{\tau_c}}\vartheta_k^{(t)})$. In the following, we struggle to convert $\mathcal{P}5$ into a GP problem. By mathematic transformations, $\mathcal{P}5$ can be further written as

$$\mathcal{P}6:\max_{\boldsymbol{\sigma},\mathbf{p},\mathbf{q}}\prod\nolimits_{k=1}^{K}\sigma_k^{\lambda_k^{(t)}} \tag{39a}$$

$$\begin{aligned}&\left(\tau_c p_k q_k\Upsilon_k+\Sigma+\Theta_k\times\right.\\&\text{s.t. }\left.\left(\sum_{i=1}^{K}\frac{q_i}{\Theta_i}\left(\xi_i+\gamma_i\right)\left(\sum_{j=1}^{K}p_j\left(\xi_j+\gamma_j\right)+1\right)+1\right)\right)\sigma_k\\&\leq M\tau_c p_k q_k(\xi_k+\gamma_k)^2+M\tau_c p_k q_k\sigma_k\xi_k^2,\ \forall k,\end{aligned} \tag{39b}$$

$$(31c)-(31d), \tag{39c}$$

where

$$\begin{aligned}\Sigma=&\tau_c q_k\sum_{i=1}^{K}p_i\left(\Omega_{ki}+\Psi_{ik}\right)+\sum_{i=1}^{K}p_i^2\Upsilon_i\\&+\frac{1}{\tau_c}\sum_{i=1}^{K}p_i^2\left(M\tau_c\gamma_i^2+2M\tau_c\gamma_i\xi_i+2\gamma_i\xi_i+\gamma_i^2+\chi_i\right)\\&+\frac{1}{\tau_c}\sum_{i=1}^{K}\sum_{j=1}^{K}p_i p_j\Xi'_{ij}+\sum_{i=1}^{K}\sum_{j=1}^{K}p_i p_j\left(\Omega_{ij}+\Psi_{ij}\right),\end{aligned} \tag{40}$$

$$\Theta_k=q_k\tau_c\left(\xi_k+\gamma_k\right)+\sum\nolimits_{i=1}^{K}p_i\left(\xi_i+\gamma_i\right)+1. \tag{41}$$

In standard GP form, the left-hand side of a less-than inequality constraint should be a posynomial while the right-hand side should be a monomial [40], which is not satisfied in (39b). For ease of tractability, we resort to the SCA method by Lemma 5 to approximate (39b) [41].

Lemma 5*: The posynomial* $Z\left(\mathbf{x}\right)=\sum_i z_i\left(\mathbf{x}\right)$ *can be approximated by the following inequality:*

$$Z\left(\mathbf{x}\right)\geq\hat{Z}\left(\mathbf{x}\right)=\prod_i\left(\frac{z_i\left(\mathbf{x}\right)}{u_i}\right)^{u_i}, \tag{42}$$

where $u_i=z_i\left(\mathbf{x}_0\right)/Z\left(\mathbf{x}_0\right),\forall i$*,* $\mathbf{x}_0$ *represents the optimal solution from the last iteration. In particular, the approximation is tight for any fixed positive* $\mathbf{x}=\mathbf{x}_0$*.*

[4]Considering Lemma 2 and Lemma 3, we have the constraint $\Gamma_k\geq 1/g^{-1}\left(Q^{-1}\left(\epsilon\right)/\sqrt{\tau_c}\right)\geq\frac{\sqrt{17}-3}{4}$, which is readily satisfied in general URLLC scenario.

Proof: Please refer to Lemma 1 in [41].

Let $q_i^{(t)}$ and $p_i^{(t)}$ are the suboptimal power allocation of i-th user in the t-th iteration. Applying Lemma 5 for the illegal terms in (39b), we can obtain their approximations in the $t+1$-th iteration as follows

$$\begin{aligned}\Theta_i &= q_i\tau_c\left(\xi_i+\gamma_i\right)+\sum\nolimits_{j=1}^{K}p_j\left(\xi_j+\gamma_j\right)+1\\ &\geq\left(\frac{q_i\tau_c\left(\xi_i+\gamma_i\right)}{\upsilon_i^{(t)}}\right)^{\upsilon_i^{(t)}}\prod_{j=1}^{K}\left(\frac{p_j\left(\xi_j+\gamma_j\right)}{\mu_j^{(t)}}\right)^{\mu_j^{(t)}}\left(\frac{1}{\varsigma_i^{(t)}}\right)^{\varsigma_i^{(t)}}\\ &=\hat{\Theta}_i^{(t)},\ \forall i,\end{aligned}\tag{43}$$

where $\upsilon_i^{(t)}$, $\mu_i^{(t)}$, and $\varsigma_i^{(t)}$, $\forall i,j$ are respectively given by

$$\upsilon_i^{(t)}=\frac{q_i^{(t)}\tau_c\left(\xi_i+\gamma_i\right)}{\Theta_i^{(t)}},\ \mu_i^{(t)}=\frac{p_j^{(t)}\left(\xi_j+\gamma_j\right)}{\Theta_i^{(t)}},\ \varsigma_i^{(t)}=\frac{1}{\Theta_i^{(t)}},\tag{44}$$

with $\Theta_i^{(t)}=q_i^{(t)}\tau_c\left(\xi_i+\gamma_i\right)+\sum_{j=1}^{K}p_j^{(t)}\left(\xi_j+\gamma_j\right)+1$.

Similarly, in the $t+1$-th iteration, we have

$$\begin{aligned}\Lambda_k &= M\tau_c p_k q_k(\xi_k+\gamma_k)^2+M\tau_c p_k q_k\sigma_k\xi_k^2\\ &\geq\left(\frac{M\tau_c p_k q_k(\xi_k+\gamma_k)^2}{\iota_k^{(t)}}\right)^{\iota_k^{(t)}}\left(\frac{M\tau_c p_k q_k\sigma_k\xi_k^2}{\rho_k^{(t)}}\right)^{\rho_k^{(t)}}\\ &=\hat{\Lambda}_k^{(t)},\ \forall k,\end{aligned}\tag{45}$$

where $\iota_k^{(t)}$ and $\rho_k^{(t)}$, $\forall k$ are respectively expressed as

$$\iota_k^{(t)}=\frac{M\tau_c p_k^{(t)}q_k^{(t)}(\xi_k+\gamma_k)^2}{\Lambda_k^{(t)}},\ \rho_k^{(t)}=\frac{M\tau_c p_k^{(t)}q_k^{(t)}\sigma_k^{(t)}\xi_k^2}{\Lambda_k^{(t)}},\tag{46}$$

with $\Lambda_k^{(t)}=M\tau_c p_k^{(t)}q_k^{(t)}(\xi_k+\gamma_k)^2+M\tau_c p_k^{(t)}q_k^{(t)}\sigma_k^{(t)}\xi_k^2$. Then, in the $t+1$-th iteration, the constraint (39b) can be approximated as

$$\begin{aligned}&\left(\tau_c p_k q_k\Upsilon_k+\Sigma+\Theta_k\right.\\ &\left.\times\left(\sum_{i=1}^{K}\frac{q_i}{\hat{\Theta}_i^{(t)}}\left(\xi_i+\gamma_i\right)\left(\sum_{j=1}^{K}p_j\left(\xi_j+\gamma_j\right)+1\right)+1\right)\right)\sigma_k\\ &\leq\hat{\Lambda}_k^{(t)},\ \forall k,\end{aligned}\tag{47}$$

which conforms to the standard format of GP. Finally, $\mathcal{P}6$ can be rewritten a GP problem as

$$\mathcal{P}7:\max_{\boldsymbol{\sigma},\mathbf{p},\mathbf{q}}\prod\nolimits_{k=1}^{K}\sigma_k^{\lambda_k^{(t)}}\tag{48a}$$

$$\text{s.t. }(47),(31c)-(31d),\tag{48b}$$

which can be resolved efficiently by employing convex optimization package CVX [42]. Based on the above discussions, we summarize the details in Algorithm 2.

2) Algorithm Analysis: It should be noted that the constraint (47) is different in each iteration but the suboptimal solution $\{p_k^{(t)},q_k^{(t)},\sigma_k^{(t)},\forall k\}$ is a feasible solution to (47) in the next iteration, which is proved in the following.

Algorithm 2 Power Control Based on GP.

1: Initialize iteration number t = 1, the power allocation $\{p_k^{(1)},q_k^{(1)},\forall k\}$, error tolerance ζ.
2: Use (33), (36), (44), (46), and $\lambda_k^{(1)}=\frac{w_k}{\ln 2}(\hat{\vartheta}_k^{(1)}-\frac{Q^{-1}(\varepsilon)}{\sqrt{\tau_c}}\vartheta_k^{(1)})$ to compute $\{\vartheta_k^{(1)},\hat{\vartheta}_k^{(1)},\sigma_k^{(1)},\lambda_k^{(1)},\upsilon_k^{(1)},\mu_k^{(1)},\varsigma_k^{(1)},\iota_k^{(1)},\rho_k^{(1)},\forall k\}$ and compute the objective function of $\mathcal{P}4$, denoted as $OF^{(1)}$. Set $OF^{(0)}=OF^{(1)}\zeta$
3: **while** $\left|OF^{(t)}-OF^{(t-1)}\right|/OF^{(t)}\geq\zeta$ **do**
4: Given $\{\vartheta_k^{(t)},\hat{\vartheta}_k^{(t)},\sigma_k^{(t)},\lambda_k^{(t)},\upsilon_k^{(t)},\mu_k^{(t)},\varsigma_k^{(t)},\iota_k^{(t)},\rho_k^{(t)},\forall k\}$, Set $t=t+1$, solve $\mathcal{P}7$ by using CVX package, obtaining $\{p_k^{(t)},q_k^{(t)},\sigma_k^{(t)},\forall k\}$.
5: Update $\{\vartheta_k^{(t)},\hat{\vartheta}_k^{(t)},\lambda_k^{(t)},\upsilon_k^{(t)},\mu_k^{(t)},\varsigma_k^{(t)},\iota_k^{(t)},\rho_k^{(t)},\forall k\}$ and $OF^{(t)}$.
6: **end while**

Since $\{p_k^{(t)},q_k^{(t)},\sigma_k^{(t)},\forall k\}$ is the suboptimal solution in the t-th iteration, based on (47) we have

$$\begin{aligned}&\left(\tau_c p_k^{(t)}q_k^{(t)}\Upsilon_k+\Sigma+\Theta_k\times\right.\\ &\left.\left(\sum_{i=1}^{K}\frac{q_i^{(t)}\left(\xi_i+\gamma_i\right)}{\hat{\Theta}_i^{(t-1)}(q_i^{(t)},p_j^{(t)})}\left(\sum_{j=1}^{K}p_j^{(t)}\left(\xi_i+\gamma_i\right)+1\right)+1\right)\right)\\ &\times\sigma_k^{(t)}\leq\hat{\Lambda}_k^{(t-1)}(q_k^{(t)},p_k^{(t)}),\ \forall k.\end{aligned}\tag{49}$$

By using (43) and (45), in the t-th iteration, we have

$$\begin{aligned}\Theta_i(q_i^{(t)},p_j^{(t)}) &= q_i^{(t)}\tau_c\left(\xi_i+\gamma_i\right)+\sum\nolimits_{j=1}^{K}p_j^{(t)}\left(\xi_j+\gamma_j\right)+1\\ &\geq\left(\frac{q_i^{(t)}\tau_c\left(\xi_i+\gamma_i\right)}{\upsilon_i^{(t)}}\right)^{\upsilon_i^{(t)}}\prod_{j=1}^{K}\left(\frac{p_j^{(t)}\left(\xi_j+\gamma_j\right)}{\mu_j^{(t)}}\right)^{\mu_j^{(t)}}\left(\frac{1}{\varsigma_i^{(t)}}\right)^{\varsigma_i^{(t)}}\\ &=\hat{\Theta}_i^{(t-1)}(q_i^{(t)},p_j^{(t)}),\ \forall i,\end{aligned}\tag{50}$$

$$\begin{aligned}\Lambda_k(q_k^{(t)},p_k^{(t)}) &= M\tau_c p_k^{(t)}q_k^{(t)}(\xi_k+\gamma_k)^2+M\tau_c p_k^{(t)}q_k^{(t)}\sigma_k\xi_k^2\\ &\geq\left(\frac{M\tau_c p_k^{(t)}q_k^{(t)}(\xi_k+\gamma_k)^2}{\iota_k^{(t)}}\right)^{\iota_k^{(t)}}\left(\frac{M\tau_c p_k^{(t)}q_k^{(t)}\sigma_k\xi_k^2}{\rho_k^{(t)}}\right)^{\rho_k^{(t)}}\\ &=\hat{\Lambda}_k^{(t-1)}(q_k^{(t)},p_k^{(t)}),\ \forall k.\end{aligned}\tag{51}$$

Using the tightness of Lemma 5, we have

$$\hat{\Theta}_i^{(t)}(q_i^{(t)},p_j^{(t)})=\Theta_i(q_i^{(t)},p_j^{(t)})\geq\hat{\Theta}_i^{(t-1)}(q_i^{(t)},p_j^{(t)}),\tag{52}$$

$$\hat{\Lambda}_k^{(t)}(q_k^{(t)},p_k^{(t)})=\Lambda_k(q_k^{(t)},p_k^{(t)})\geq\hat{\Lambda}_k^{(t-1)}(q_k^{(t)},p_k^{(t)}).\tag{53}$$

Finally, by combing (49), (52), and (53), we have

$$\begin{aligned}&\left(\tau_c p_k^{(t)}q_k^{(t)}\Upsilon_k+\Sigma+\Theta_k\times\right.\\ &\left.\left(\sum_{i=1}^{K}\frac{q_i^{(t)}\left(\xi_i+\gamma_i\right)}{\hat{\Theta}_i^{(t)}(q_i^{(t)},p_j^{(t)})}\left(\sum_{j=1}^{K}p_j^{(t)}\left(\xi_i+\gamma_i\right)+1\right)+1\right)\right)\\ &\times\sigma_k^{(t)}\leq\hat{\Lambda}_k^{(t)}(q_k^{(t)},p_k^{(t)}),\ \forall k.\end{aligned}\tag{54}$$

Thus, $\{p_k^{(t)},q_k^{(t)},\sigma_k^{(t)},\forall k\}$ is a feasible solution in the $t+1$ iteration. Based on the above analysis, then we focus on the convergence of Algorithm 2. To begin with, we verify that

$OF^{(t)} \leq OF^{(t+1)}$. In the $t+1$ iteration, according to (37), we have

$$\begin{aligned}&\sum_{k=1}^{K} \frac{w_k}{\text{In}2}\left[\hat{\vartheta}_k^{(t)}\text{In}\sigma_k^{(t+1)} + \hat{\nu}_k^{(t)} - a\left(\vartheta_k^{(t)}\text{In}\sigma_k^{(t+1)} + \nu_k^{(t)}\right)\right] \\ &\geq \sum_{k=1}^{K} \frac{w_k}{\text{In}2}\left[\hat{\vartheta}_k^{(t)}\text{In}\sigma_k^{(t)} + \hat{\nu}_k^{(t)} - a\left(\vartheta_k^{(t)}\text{In}\sigma_k^{(t)} + \nu_k^{(t)}\right)\right] \\ &= \sum_{k=1}^{K} \frac{w_k}{\text{In}2}\left[\text{In}\left(1+\sigma_k^{(t)}\right) - aU\left(\sigma_k^{(t)}\right)\right] = OF^{(t)},\end{aligned} \tag{55}$$

where $a = Q^{-1}(\varepsilon)/\sqrt{\tau_c}$ and the equality holds because $\sigma_k = \sigma_k^{(t)}$ in (37). Similarly, the following equalities also hold when $\sigma_k = \sigma_k^{(t+1)}$:

$$\begin{aligned}OF^{(t+1)} &= \sum_{k=1}^{K} \frac{w_k}{\text{In}2}\left[\text{In}\left(1+\sigma_k^{(t+1)}\right) - aU\left(\sigma_k^{(t+1)}\right)\right] \\ &\geq \sum_{k=1}^{K} \frac{w_k}{\text{In}2}\left[\hat{\vartheta}_k^{(t)}\text{In}\sigma_k^{(t+1)} + \hat{\nu}_k^{(t)} - a\left(\vartheta_k^{(t)}\text{In}\sigma_k^{(t+1)} + \nu_k^{(t)}\right)\right].\end{aligned} \tag{56}$$

Obviously, $OF^{(t)} \leq OF^{(t+1)}$ can be obtained by combining (55) and (56) while $OF^{(t)}$ is upper bounded by the finite energy at users, which indicates that Algorithm 2 is convergent.

The complexity of Algorithm 2 of each iteration primarily comes from solving the GP $\mathcal{P}7$, which includes $3K$ constraints and $3K$ variables. Hence, the complexity of Algorithm 2 can be expressed as $\mathcal{O}(n_2^{iter} \times \max\{27K^3, n_2^{cost}\})$ [43], where n_2^{iter} and n_2^{cost} denote the total number of iterations and the computational complexity of calculating the first-order and second-order derivatives of the objective and constraint functions of $\mathcal{P}7$, respectively.

D. BCD Algorithm

1) Proposed Algorithm: Noticing that the initial solution in Algorithm 2 must be a feasible solution of $\mathcal{P}3$ while that of Algorithm 1 is generated randomly. To find such an initial solution for Algorithm 2, we firstly need to solve the following problem[5]:

$$\mathcal{P}8: \max_{r,\mathbf{p},\mathbf{q}} \ r \tag{57a}$$

$$\text{s.t. } \Gamma_k \geq r/f^{-1}(0), \ \forall k, \tag{57b}$$

$$\tau_c(p_k + q_k) \leq E, \ \forall k, \tag{57c}$$

where r is an auxiliary variable and $r \geq 1$ means that the solution is valid for initializing. In this paper, we guarantee $r \geq 1$ to keep the proposed algorithm running. Hence, the phase shifts design should be implemented before the power control design in the first iteration of BCD algorithm, which facilitates us to find the valid initial solution under optimized phase shifts $\mathbf{\Phi}$. Defining $R\left(\mathbf{\Phi}^{(l)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right)$ as the corresponding objective function value based on the solution in l-th iteration $\mathbf{p}^{(l)} = [p_1^{(l)}, p_2^{(l)}, \ldots, p_K^{(l)}]$, $\mathbf{q}^{(l)} = [q_1^{(l)}, q_2^{(l)}, \ldots, q_K^{(l)}]$, and $\mathbf{\Phi}^{(l)}$, the detail of BCD is provided in Algorithm 3.

[5]The problem can be also transformed into a GP problem and solved by using the same approximated approach as in $\mathcal{P}3$.

Algorithm 3 Joint Design Based on BCD.

1: Initialize iteration number $l = 1$, power allocation $\mathbf{p}^{(1)}$, $\mathbf{q}^{(1)}$, phase shift $\mathbf{\Phi}^{(1)}$, error tolerance Δ, and calculate $R\left(\mathbf{\Phi}^{(1)}, \mathbf{p}^{(1)}, \mathbf{q}^{(1)}\right)$ based on (20).
2: Given $\mathbf{p}^{(l)}, \mathbf{q}^{(l)}$, calculate the suboptimal phase shifts $\mathbf{\Phi}^{(l+1)}$ by solving $\mathcal{P}2$ with Algorithm 1.
3: Given $\mathbf{\Phi}^{(l+1)}$, calculate the parameters $\{f_k\left(\mathbf{\Phi}^{(l+1)}\right), \xi_k^{(l+1)}, \Upsilon_k^{(l+1)}, \chi_k^{(l+1)}, \Xi_{ki}^{(l+1)}, \Omega_{ki}^{(l+1)}, \Psi_{ij}^{(l+1)}, \forall i,j,k\}$ based on (12) and (22)−(26).
4: Given $\{\Upsilon_k^{(l+1)}, \chi_k^{(l+1)}, \Xi_{ki}^{(l+1)}, \Omega_{ki}^{(l+1)}, \Psi_{ij}^{(l+1)}, \forall i,j,k\}$, calculate the suboptimal power allocation $\mathbf{p}^{(l+1)}, \mathbf{q}^{(l+1)}$ by solving $\mathcal{P}7$ Algorithm 2.
5: Update $R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l+1)}, \mathbf{q}^{(l+1)}\right)$. If

$$\frac{\left|R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l+1)}, \mathbf{q}^{(l+1)}\right) - R\left(\mathbf{\Phi}^{(l)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right)\right|}{R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l+1)}, \mathbf{q}^{(l+1)}\right)} < \Delta,$$

terminate. Otherwise, set $l = l+1$, go to step 2.

2) Algorithm Analysis: Now let us analyze Algorithm 3 in terms of the convergence and complexity. Firstly, the suboptimal solution $\left(\mathbf{\Phi}^{(l)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right)$ obtained in step 5 is also a feasible solution to $\mathcal{P}2$ in step 2. With given $\mathbf{p}^{(l)}$, $\mathbf{q}^{(l)}$, the suboptimal solution for $\mathcal{P}2$ in step 2 is $\mathbf{\Phi}^{(l+1)}$ and thus we have $R\left(\mathbf{\Phi}^{(l)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right) \leq R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right)$. Similarly, $\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right)$ is also a feasible solution to $\mathcal{P}7$ in step 4 and we have $R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right) \leq R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l+1)}, \mathbf{q}^{(l+1)}\right)$ by using the optimality. Then, following that $R\left(\mathbf{\Phi}^{(l)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right) \leq R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l)}, \mathbf{q}^{(l)}\right) \leq R\left(\mathbf{\Phi}^{(l+1)}, \mathbf{p}^{(l+1)}, \mathbf{q}^{(l+1)}\right)$, the objective value is monotonically increasing. Also, the objective value is upper bounded by the energy constraint. Hence, the convergence of Algorithm 3 is guaranteed.

The complexity of Algorithm 3 is dominated by the step 2 and step 4. Based on the aforementioned complexity analysis, the overall complexity of Algorithm 3 is given by $\mathcal{O}\left(N_{iter}\left(n_1^{iter} n_1^{cost} L + n_2^{iter} \times \max\left\{27K^3, n_2^{cost}\right\}\right)\right)$, where N_{iter} denotes the number of iterations Algorithm 3.

V. Numerical Results

In this part, we demonstrate the superiority brought by integrating the SP and RIS into mMIMO URLLC systems. We consider a half circle with a center $(0,0,0)$ and a radius of 20 m, where the users with a height of 1.6 m are distributed evenly on the half circle [7]. We assume that the RIS and BS are located at $(0,0,30)$ and $(1000,0,25)$, respectively. All the AoA and AoD are fixed after generating randomly from $[0, 2\pi]$. For simplicity, the element spacing is set as $d = \frac{\lambda}{2}$. Unless otherwise stated, in the following simulations, we set number of users $K = 4$, number of reflecting elements of $N = 100$, number of antennas of $M = 100$, maximal available energy $E = 10$ dB, maximal available blocklength $\tau_c = 100$, decoding error rate $\epsilon = 10^{-5}$, noise power density of -174 dBm/Hz. Besides, large-scale path loss $\alpha_k = 10^{-3} d_{kUR}^{-\alpha_{kUR}}$, $\beta = 10^{-3} d_{RB}^{-\beta_{RB}}$, and $\gamma_k = 10^{-3} d_{kUB}^{-\gamma_{kUB}}, \forall k$, in which d_{kUR}, d_{RB}, and d_{kUB} denote the distance of user-k-RIS,

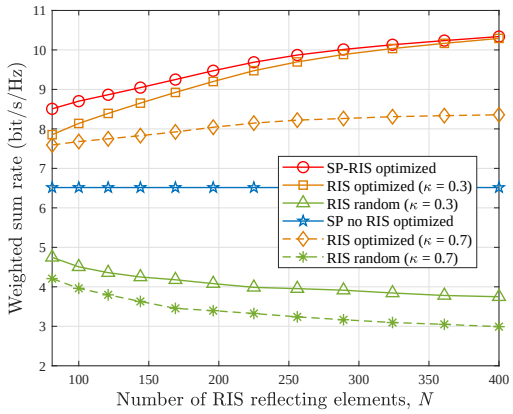

Fig. 4. Weighted sum rate versus RIS reflecting element number.

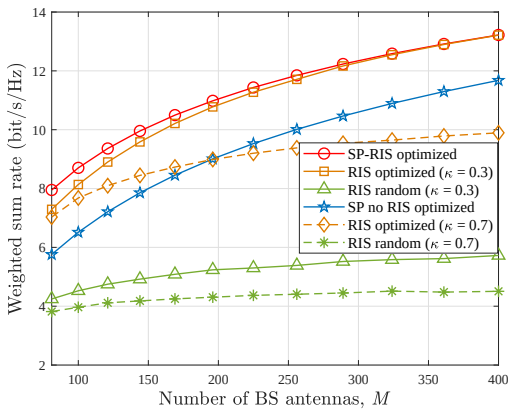

Fig. 5. Weighted sum rate versus antenna number.

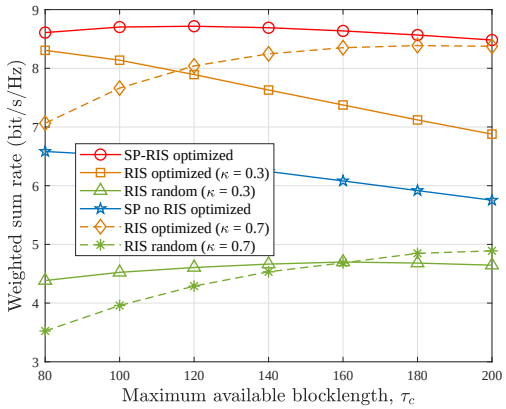

Fig. 6. Weighted sum rate versus maximum available blocklength.

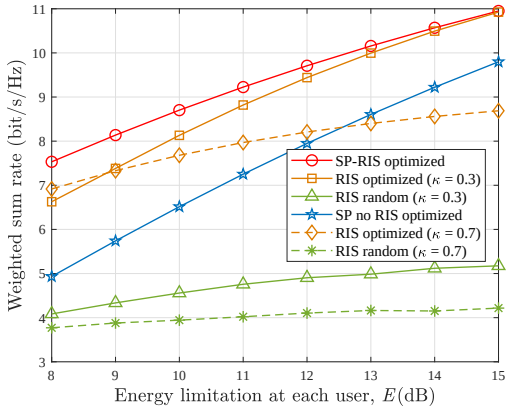

Fig. 7. Weighted sum rate versus energy limitation at each user.

RIS-BS, and user-k-BS, respectively, and path-loss exponents $\alpha_{kUR} = 2$, $\beta_{RB} = 2.5$, and $\gamma_{kUB} = 4, \forall k$. Rician factor $\delta = 1$, $\varepsilon_k = 10, \forall k$. The priority of users $w_k = 1, \forall k$. The GA parameters are set as $L = 200$, $L_e = 10$, $L_m = 38$, and $L_p = 152$.

To illustrate the superiority, we compare the weighted sum rate of following benchmark schemes with that of the proposed scheme where both phase shifts and power allocation are optimized (marked as '**SP-RIS optimized**' in the figures):

- **RIS optimized ($\kappa = 0.3/0.7$):** In this case, only the RIS phase shifts are optimized while the power of SP is fixed by power factor κ which equals to the ratio of data power to pilot power, and a low value of κ means a better channel estimation quality. In this paper, we have $p_k = \frac{\kappa E}{\tau_c}$, $q_k = \frac{(1-\kappa)E}{\tau_c}, \forall k$.
- **RIS random:** Neither the RIS phase shifts nor power control are optimized. Specifically, the RIS phase shifts $\mathbf{\Phi}$ are generated randomly for 5000 times and then averaged to obtain the result.
- **SP no RIS optimized:** Without RIS, the SP strategy is adopted to assist URLLC transmission under optimized power control.

Fig. 4 shows the weighted sum rate versus RIS reflecting element number N. We can observe that when the optimization of phase shifts is considered, the rate performance is improved as the increase of reflecting elements. This is because large N provides more degree of freedom to decrease the interference from other users and own pilot signal. The proposed scheme has the best performance and shows different gain compared with the schemes only optimizing phase shifts. Specifically, the gain from power control optimization becomes more and more obvious for the case of $\kappa = 0.7$ while that is only obvious in the small N region for the case of $\kappa = 0.3$. The reason of which is that full pilot power guarantees the channel estimation quality and then the element number begins to dominate the performance as N increases. Furthermore, It can be found that integrating RIS with SP will degrade the system performance if its phase shifts are random, since more interference is introduced by the cascaded channel and adding element number deteriorates this situation.

Fig. 5 illustrates the impact of antenna number M on the weighted sum rate. As expected, by increasing the antenna number, the system obtains a better performance due to multiplexing gain. Clearly, the performance gap between the proposed scheme and the scheme without RIS is significant, which indicates the advantage of integrating SP and RIS in mMIMO URLLC systems and the strength of RIS can not be reflected without optimization of phase shifts. Although only optimizing phase shifts with $\kappa = 0.7$ and $\kappa = 0.3$ can achieve a substantial improvement even a comparable performance to the proposed scheme, the $\kappa = 0.7$ scheme just outperforms the scheme without RIS when $M < 169$ and the $\kappa = 0.3$ scheme

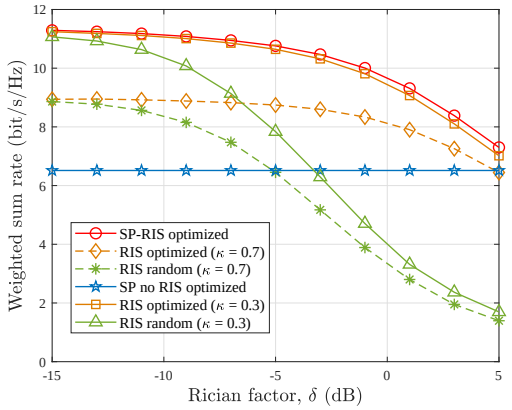

Fig. 8. Weighted sum rate versus the Rician factor δ.

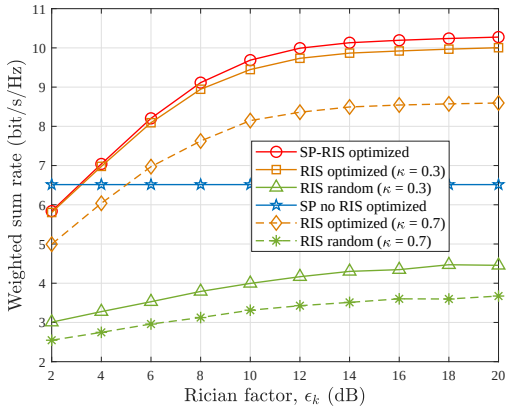

Fig. 9. Weighted sum rate versus the Rician factor ϵ_k.

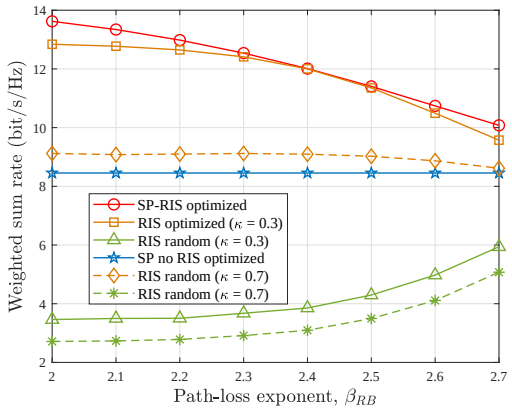

Fig. 10. Weighted sum rate versus the path-loss exponent β_{RB}.

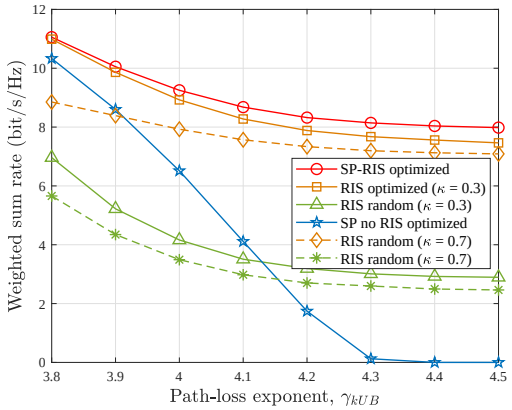

Fig. 11. Weighted sum rate versus the path-loss exponent γ_{kUB}.

requires a enormous antenna cost. Therefore, it is important to implement jointly optimizing power control and phase shifts in our system.

In Fig. 6, we study the impact of blocklength on the system performance. On the one hand, increasing blocklength means smaller rate loss and more accurate channel estimate in short packet transmission due to longer data and pilot sequences. On the other hand, the energy storage at users might be tight under SP with the growth of blocklength. As a result, a declining tendency can be observed from the benchmark schemes 'RIS optimized $\kappa = 0.3$' and 'SP no RIS optimized'. Noted that the rate of 'RIS optimized $\kappa = 0.3$' scheme is no longer closed to that of the proposed scheme as the above figures. In contrast, the performance of other benchmark schemes is enhanced by raising τ_c especially the schemes with $\kappa = 0.7$ which surpasses the counterpart with $\kappa = 0.3$. These manifest that the quality of channel estimation can be improved by lengthening the pilot when the pilot power is insufficient. Moreover, the proposed scheme is always over other benchmarks when τ_c rises, which reveals the superiority in URLLC transmission once again.

In Fig. 7, the influence of energy limitation is investigated. The figure illustrates that all the schemes obtain a boost on rate when more energy is available at each user. Clearly, in the low energy region, the proposed design exhibits a considerable gain since the pilot-data interference in SP and the multi-user interference can be reduced by taking advantage of the limited power. As energy growing, the rate of optimizing RIS with $\kappa = 0.3$ goes up more rapidly compared to that of $\kappa = 0.7$ and eventually approach that of the proposed design. In the large energy region, the performance gap between $\kappa = 0.3$ and $\kappa = 0.7$ is apparent no matter whether the phase shifts are optimized, which uncovers that assigning more power to pilot is necessary in SP.

Fig. 8 and Fig. 9 show the effect of Rician factor δ and $\epsilon_k, \forall k$ on the rate of different schemes, respectively, where N is set as 225. In Fig. 8, the rate achieved by the schemes merging RIS drops with δ in different degree. It is due to the fact that the LoS components start to play a dominant role in RIS-BS channel $\mathbf{H}_2$ while δ rises, resulting in a strong channel correlation among different users. Then, the spatial multiplexing gain is weakened and the inter-user interference is more severe. Specifically, the rate obtained by random phases falls sharply and below the rate of No-RIS scheme at $\delta = -3$, whereas the rate obtained by optimized phases has a moderate downward trend but always over the rate of No-RIS scheme. Hence, the optimization of phases is essential when the RIS is deployed in the surroundings with abundant scatters to support URLLC. Differently, Fig. 9 presents an upward trend on the rate of all the schemes, except for the No-RIS scheme. However, when $\epsilon_k = 2$, the proposed scheme has a worse performance compared to the scheme without RIS. This

is because the increase of ϵ_k enhances the dominance of LoS components and decreases the NLoS components of $\mathbf{h}_k$, which improves the SINR of users. Thus, when ϵ_k is large enough such that the user-k-RIS channel approaches a deterministic channel, the growth of rate tends to flatten out.

Fig. 10 illustrates the rate of various schemes versus the path-loss exponent β_{RB}, where $M = 169$ and $N = 225$. It is interesting to find that the performance of random RIS schemes become better with β_{RB} while that of optimized RIS schemes degrade at varying level. The reason is that the signal reflected by the RIS is weaker as β_{RB} increasing, giving rise to the worse SINR at users. Therefore, the gain from RIS link is eroded. To avoid this situation, the RIS should be deployed at a location with small β_{RB}.

Finally, we examine the impact of path-loss exponent γ_{kUB} in Fig. 11, where $M = 100$ and $N = 169$. As expected, all the weighted sum rate decreases with γ_{kUB}, since the signals from direct link become weaker. Obviously, the rate quickly slides into zero with γ_{kUB} when no RIS is deployed and other rate achieved by using RIS reaches a bottom. This unveils the benefit of integrating RIS with SP in URLLC scenario.

VI. Conclusion

In this paper, we combined the advantages of RIS and SP to enable URLLC in mMIMO systems. We derived the closed-form finite blocklength ergodic achievable rate LB while considering LMMSE channel estimation error and imperfect pilot interference removal. Then, the weighted sum rate maximization problem was studied by jointly designing the power control of SP at each user and the phase shifts at the RIS. To solve this intractable problem, we first decoupled it into the phase shift design subproblem and the power control design subproblem. A GA and an iterative algorithm based on GP were proposed to tackle these two subproblems, respectively. To optimize the variables in an alternating manner, the BCD algorithm was presented. Besides, the corresponding complexity and convergence analysis were also provided. Numerical results confirmed that the superiority of jointly designing RIS and SP in enhancing URLLC transmission rate.

Appendix A
Proof of Lemma 1

Before deriving the LMMSE estimator, we show the following results given in [38, Lemma 1] and [44]:

$$\mathbb{E}\left\{\left\|\mathbf{g}_k\right\|^2\right\} = M\xi_k, \tag{58}$$

$$\mathbb{E}\left\{\left\|\mathbf{f}_k\right\|^2\right\} = \mathbb{E}\left\{\left\|\mathbf{g}_k + \mathbf{d}_k\right\|^2\right\} = M\left(\xi_k + \gamma_k\right), \tag{59}$$

where

$$\xi_k = \frac{\beta\alpha_k}{\left(\delta+1\right)\left(\varepsilon_k+1\right)}\left(\delta\varepsilon_k\left|f_k\left(\mathbf{\Phi}\right)\right|^2 + \left(\delta+\varepsilon_k+1\right)N\right). \tag{60}$$

Given the received signal $\mathbf{y}_k$, using the Bayesian Gauss-Markov theorem [29, Theorem 12.1], the LMMSE estimator of $\mathbf{f}_k$ is given by

$$\begin{aligned}\hat{\mathbf{f}}_k &= \mathbb{E}\left\{\mathbf{f}_k\right\} + \mathrm{Cov}\left\{\mathbf{f}_k, \mathbf{y}_k\right\}\mathrm{Cov}^{-1}\left\{\mathbf{y}_k, \mathbf{y}_k\right\}\left(\mathbf{y}_k - \mathbb{E}\left\{\mathbf{y}_k\right\}\right)\\ &\overset{(a)}{=} \mathbb{E}\left\{\mathbf{f}_k\mathbf{y}_k^H\right\}\left\{\mathbb{E}\left\{\mathbf{y}_k\mathbf{y}_k^H\right\}\right\}^{-1}\mathbf{y}_k,\end{aligned} \tag{61}$$

where (a) is due to the fact that $\mathbb{E}\left\{\mathbf{f}_k\right\} = \mathbf{0}$ and $\mathbb{E}\left\{\mathbf{y}_k\right\} = \mathbf{0}$. Then, the expectations $\mathbb{E}\left\{\mathbf{f}_k\mathbf{y}_k^H\right\}$ and $\mathbb{E}\left\{\mathbf{y}_k\mathbf{y}_k^H\right\}$ are respectively calculated as

$$\mathbb{E}\left\{\mathbf{f}_k\mathbf{y}_k^H\right\} = \sqrt{q_k\tau_c}\mathbb{E}\left\{\mathbf{f}_k\mathbf{f}_k^H\right\} = \sqrt{q_k\tau_c}\left(\xi_k+\gamma_k\right)\mathbf{I}_M, \tag{62}$$

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{y}_k\right\|^2\right\}\\ &= q_k\tau_c\mathbb{E}\left\{\mathbf{f}_k\mathbf{f}_k^H\right\} + \frac{1}{\tau_c}\sum\nolimits_{i=1}^{K} p_i\mathbb{E}\left\{\mathbf{f}_i\mathbf{s}_i^H\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\right\}\\ &\quad + \frac{1}{\tau_c}\mathbb{E}\left\{\mathbf{W}\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\mathbf{W}^H\right\}\\ &= \left(q_k\tau_c\left(\xi_k+\gamma_k\right) + \sum\nolimits_{i=1}^{K} p_i\left(\xi_i+\gamma_i\right) + 1\right)\mathbf{I}_M.\end{aligned} \tag{63}$$

Substituting (62) and (63) into (61), we can obtain the LMMSE estimator.

Appendix B
Proof of Theorem 1

According to the appendix in [44], we have

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{f}_k\right\|^4\right\} = \mathbb{E}\left\{\left\|\mathbf{d}_k + \mathbf{g}_k\right\|^4\right\}\\ &= \mathbb{E}\left\{\left\|\mathbf{g}_k\right\|^4\right\} + 2M\gamma_k\xi_k + M\left(M+1\right)\gamma_k^2 + 2M^2\gamma_k\xi_k,\end{aligned} \tag{64}$$

$$\begin{aligned}\mathbb{E}\left\{\left|\mathbf{f}_k^H\mathbf{f}_i\right|^2\right\} &= \mathbb{E}\left\{\left|\left(\mathbf{d}_k^H + \mathbf{g}_k^H\right)\left(\mathbf{d}_i + \mathbf{g}_i\right)\right|^2\right\}\\ &= \mathbb{E}\left\{\left|\mathbf{g}_k^H\mathbf{g}_i\right|^2\right\} + M\gamma_k\xi_i + M\gamma_i\xi_k + M\gamma_i\gamma_k,\end{aligned} \tag{65}$$

where $\mathbb{E}\{\|\mathbf{g}_k\|^4\}$ and $\mathbb{E}\{|\mathbf{g}_k^H\mathbf{g}_i|^2\}$ has been given in [38, Lemma 1].

The detail of the above derivations is given by in [38] and [44]. Note that different from the RIS-free scenario, the user channel $\mathbf{f}_k$ is correlated to $\mathbf{f}_i$, $k \neq i$, since the existence of the common RIS-BS channel $\mathbf{H}_2$. Therefore, we need to calculate the following expectation:

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{f}_k\right\|^2\left\|\mathbf{f}_i\right\|^2\right\} = \mathbb{E}\left\{\left\|\mathbf{g}_k + \mathbf{d}_k\right\|^2\left\|\mathbf{g}_i + \mathbf{d}_i\right\|^2\right\}\\ &= \mathbb{E}\left\{\mathbf{g}_k^H\mathbf{g}_k\mathbf{g}_i^H\mathbf{g}_i\right\} + M\gamma_k\mathbb{E}\left\{\left\|\mathbf{g}_i\right\|^2\right\}\\ &\quad + M\gamma_i\mathbb{E}\left\{\left\|\mathbf{g}_k\right\|^2\right\} + M^2\gamma_k\gamma_i.\end{aligned} \tag{66}$$

Using the similar method to [38, Lemma 1], the first term in (66) is calculated as

$$\mathbb{E}\left\{\mathbf{g}_k^H\mathbf{g}_k\mathbf{g}_i^H\mathbf{g}_i\right\} = M\left(M\xi_k\xi_i + \Xi'_{ki}\right) = M\Xi_{ki}. \tag{67}$$

Then, combining (58) and (67) with (66), we can obtain the expectation $\mathbb{E}\{\|\mathbf{f}_k\|^2\|\mathbf{f}_i\|^2\}$. In the following, we derive the closed-form expression of SINR based on the above results.

To begin with, the numerator of (18) is calculated as

$$\left|\eta_k(\xi_k+\gamma_k)^{-1}\mathbb{E}\left\{\mathbf{f}_k^H\mathbf{f}_k\right\}\right|^2 = M^2\eta_k^2. \tag{68}$$

Next, to calculate the variance in the denominator of (18), we have

$$\begin{aligned}&\mathbb{E}\left\{\left|\eta_k(\xi_k+\gamma_k)^{-1}\mathbf{f}_k^H\mathbf{f}_k\right|^2\right\}=\frac{\eta_k^2}{(\xi_k+\gamma_k)^2}\mathbb{E}\left\{\|\mathbf{f}_k\|^4\right\}\\&=\frac{M\eta_k^2}{(\xi_k+\gamma_k)^2}\left(\Delta_k+2\gamma_k\xi_k+(M+1)\gamma_k^2+2M\gamma_k\xi_k\right).\end{aligned}\tag{69}$$

With (68) and (69), the variance is calculated as

$$\begin{aligned}&\mathrm{Var}\left(\eta_k(\xi_k+\gamma_k)^{-1}\mathbf{f}_k^H\mathbf{f}_k\right)\\&=M\eta_k^2\left(\frac{\Delta_k+2\gamma_k\xi_k+\gamma_k^2-M\xi_k^2}{(\xi_k+\gamma_k)^2}\right).\end{aligned}\tag{70}$$

Then, we derive the term $\mathbb{E}\{\|\mathbf{z}_k^H-\mathbb{E}\{\mathbf{z}_k^H\}\|^2\}$, which can be divided into the following parts:

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{z}_k^H-\mathbb{E}\left\{\mathbf{z}_k^H\right\}\right\|^2\right\}\\&=\mathbb{E}\left\{\left\|\mathbf{z}_k^H\right\|^2\right\}-\left\|\mathbb{E}\left\{\mathbf{z}_k^H\right\}\right\|^2\\&=\sum\nolimits_{i=1}^4\mathbb{E}\left\{\left\|\mathbf{u}_{ik}\right\|^2\right\}-\sum\nolimits_{i=1}^4\left\|\mathbb{E}\left\{\mathbf{u}_{ik}\right\}\right\|^2\\&\quad-2\mathrm{Re}\left\{\sum\nolimits_{i=1}^4\sum\nolimits_{j=i+1}^4\mathbb{E}\left\{\mathbf{u}_{ik}\right\}\mathbb{E}\left\{\mathbf{u}_{jk}^H\right\}\right\}\\&\quad+2\mathrm{Re}\left\{\mathbb{E}\left\{\sum\nolimits_{i=1}^4\sum\nolimits_{j=i+1}^4\mathbf{u}_{ik}\mathbf{u}_{jk}^H\right\}\right\},\end{aligned}\tag{71}$$

where $\mathbf{u}_{1k}=\sqrt{p_k}\bar{\mathbf{f}}_k^H\mathbf{f}_k\mathbf{s}_k^H$, $\mathbf{u}_{2k}=\sum_{i\neq k}^K\sqrt{p_i}\hat{\mathbf{f}}_k^H\mathbf{f}_i\mathbf{s}_i^H$, $\mathbf{u}_{3k}=\sum_{i=1}^K\sqrt{q_i}\hat{\mathbf{f}}_k^H\boldsymbol{\epsilon}_i\boldsymbol{\psi}_i^H$, and $\mathbf{u}_{4k}=\hat{\mathbf{f}}_k^H\mathbf{W}$. Then, we calculate the expectations in (71) one by one with the same method as in [45], [46].

First, we focus on $\mathbb{E}\{\|\mathbf{u}_{1k}\|^2\}$, which is decomposed as

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{u}_{1k}\right\|^2\right\}=p_k\mathbb{E}\left\{\left\|\bar{\mathbf{f}}_k^H\mathbf{f}_k\mathbf{s}_k^H\right\|^2\right\}\\&=p_k\mathbb{E}\left\{\left\|c_k\sum_{i=1}^K\sqrt{\frac{p_i}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{f}_k\mathbf{s}_k^H+c_k\boldsymbol{\psi}_k^H\frac{\mathbf{W}^H}{\sqrt{\tau_c}}\mathbf{f}_k\mathbf{s}_k^H\right\|^2\right\}\\&=\underbrace{p_k\,\mathbb{E}\left\{\left\|c_k\sum_{i=1}^K\sqrt{\frac{p_i}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{f}_k\mathbf{s}_k^H\right\|^2\right\}}_{A}\\&\quad+\underbrace{p_k\,\mathbb{E}\left\{\left\|c_k\boldsymbol{\psi}_k^H\frac{\mathbf{W}^H}{\sqrt{\tau_c}}\mathbf{f}_k\mathbf{s}_k^H\right\|^2\right\}}_{B}.\end{aligned}\tag{72}$$

The above terms in (72) are respectively given by

$$\begin{aligned}A&=\mathbb{E}\left\{\left\|c_k\sqrt{\frac{p_k}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_k\mathbf{f}_k^H\mathbf{f}_k\mathbf{s}_k^H\right\|^2\right\}\\&\quad+\mathbb{E}\left\{\left\|c_k\sum\nolimits_{i\neq k}^K\sqrt{\frac{p_i}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{f}_k\mathbf{s}_k^H\right\|^2\right\}\\&=Mc_k^2p_k\left((\tau_c+1)\Delta_k+M\tau_c\gamma_k^2\right.\\&\quad\left.+2\gamma_k\xi_k(M\tau_c+M+1)+(M+1)\gamma_k^2\right)\\&\quad+Mc_k^2\tau_c\sum\nolimits_{i=1}^K p_i\left(\xi_i\gamma_k+\gamma_i\xi_k+\gamma_i\gamma_k\right)\\&\quad+Mc_k^2\tau_c\sum\nolimits_{i\neq k}^K p_i\Omega_{ki},\end{aligned}\tag{73}$$

$$B=Mc_k^2\tau_c\left(\xi_k+\gamma_k\right).\tag{74}$$

Combining (72)−(74), we can obtain the expression of $\mathbb{E}\{\|\mathbf{u}_{1k}\|^2\}$. Similarly, the term $\mathbb{E}\{\|\mathbf{u}_{2k}\|^2\}$ can be expanded as follows

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{u}_{2k}\right\|^2\right\}=\mathbb{E}\left\{\left\|\sqrt{p_i}\sum\nolimits_{i\neq k}^K\hat{\mathbf{f}}_k^H\mathbf{f}_i\mathbf{s}_i^H\right\|^2\right\}\\&=\underbrace{\mathbb{E}\left\{\left\|c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^K\sqrt{q_k\tau_c}\mathbf{f}_k^H\mathbf{f}_i\mathbf{s}_i^H\right\|^2\right\}}_{C}\\&\quad+\underbrace{\mathbb{E}\left\{\left\|c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^K\sum\nolimits_{j=1}^K\sqrt{\frac{p_j}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_j\mathbf{f}_j^H\mathbf{f}_i\mathbf{s}_i^H\right\|^2\right\}}_{D}\\&\quad+\underbrace{\mathbb{E}\left\{\left\|c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^K\boldsymbol{\psi}_k^H\frac{\mathbf{W}^H}{\sqrt{\tau_c}}\mathbf{f}_i\mathbf{s}_i^H\right\|^2\right\}}_{E},\end{aligned}\tag{75}$$

where C in (75) is given by

$$\begin{aligned}C&=q_kc_k^2\tau_c^2p_i\sum\nolimits_{i\neq k}^K\mathbb{E}\left\{\left|\mathbf{f}_k^H\mathbf{f}_i\right|^2\right\}\\&=Mq_kc_k^2\tau_c^2\sum\nolimits_{i\neq k}^K p_i(\Omega_{ki}+\gamma_k\xi_i+\gamma_i\xi_k+\gamma_i\gamma_k),\end{aligned}\tag{76}$$

and D can be rewritten as

$$\begin{aligned}D&=\underbrace{\mathbb{E}\left\{\left\|c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^K\sqrt{\frac{p_i}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{f}_i\mathbf{s}_i^H\right\|^2\right\}}_{D_1}\\&\quad+\underbrace{\mathbb{E}\left\{\left\|c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^K\sum\nolimits_{j\neq i}^K\sqrt{\frac{p_j}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_j\mathbf{f}_j^H\mathbf{f}_i\mathbf{s}_i^H\right\|^2\right\}}_{D_2}.\end{aligned}\tag{77}$$

D_1 and D_2 are respectively derived as

$$\begin{aligned}D_1&=Mc_k^2\sum\nolimits_{i\neq k}^K p_i^2\left((\tau_c+1)\Delta_i+2(\tau_c+1)\gamma_i\xi_i\right.\\&\quad\left.+(M+1)\tau_c\gamma_i^2+2M\tau_c\gamma_i\xi_i+\gamma_i^2\right)\\&\quad+M^2c_k^2\sum\nolimits_{i\neq k}^K\sum\nolimits_{j\neq k}^K p_ip_j\left(\gamma_i\xi_j+\gamma_j\xi_i+\gamma_i\gamma_j\right)\\&\quad+Mc_k^2\sum\nolimits_{i\neq k}^K\sum\nolimits_{j\neq k,j\neq i}^K p_ip_j\Xi_{ij},\end{aligned}\tag{78}$$

$$D_2 = Mc_k^2\tau_c \sum\nolimits_{i\neq k}^{K} \sum\nolimits_{j\neq i}^{K} p_i p_j \left(\Omega_{ij} + \gamma_i\xi_j + \gamma_j\xi_i + \gamma_i\gamma_j\right). \tag{79}$$

Then, we have

$$\begin{aligned} E &= \mathbb{E}\left\{\left\| c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^{K} \boldsymbol{\psi}_k^H \frac{\mathbf{W}^H}{\sqrt{\tau_c}}\mathbf{f}_i\mathbf{s}_i^H \right\|^2\right\} \\ &= \frac{c_k^2}{\tau_c}\mathbb{E}\left\{\sqrt{p_i p_j}\sum\nolimits_{i\neq k}^{K}\sum\nolimits_{j\neq k}^{K} \boldsymbol{\psi}_k^H\mathbf{W}^H\mathbf{f}_i\mathbf{s}_i^H\mathbf{s}_j\mathbf{f}_j^H\mathbf{W}\boldsymbol{\psi}_k\right\} \\ &= \frac{c_k^2}{\tau_c}\mathbb{E}\left\{p_i\sum\nolimits_{i\neq k}^{K} \boldsymbol{\psi}_k^H\mathbf{W}^H\mathbf{f}_i\mathbf{s}_i^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{W}\boldsymbol{\psi}_k\right\} \\ &= M\tau_c c_k^2\sum\nolimits_{i\neq k}^{K} p_i\left(\xi_i+\gamma_i\right). \end{aligned} \tag{80}$$

Combining (75)−(80), we can obtain the expression of $\mathbb{E}\{\|\mathbf{u}_{2k}\|^2\}$. Next, since the channel estimation error is uncorrelated to the channel estimate in LMMSE, recalling (13) and (14), we have

$$\begin{aligned} \mathbb{E}\left\{\|\mathbf{u}_{3k}\|^2\right\} &= \mathbb{E}\left\{\left\|\sum\nolimits_{i=1}^{K}\sqrt{q_i}\hat{\mathbf{f}}_k^H\boldsymbol{\epsilon}_i\boldsymbol{\psi}_i^H\right\|^2\right\} \\ &= \sum\nolimits_{i=1}^{K} q_i\mathbb{E}\left\{\hat{\mathbf{f}}_k^H\boldsymbol{\epsilon}_i\boldsymbol{\psi}_i^H\boldsymbol{\psi}_i\boldsymbol{\epsilon}_i^H\hat{\mathbf{f}}_k\right\} \\ &= M\tau_c\eta_k\sum\nolimits_{i=1}^{K} q_i\left(\xi_i+\gamma_i-\eta_i\right). \end{aligned} \tag{81}$$

Then, we focus on $\mathbb{E}\{\|\mathbf{u}_{4k}\|^2\}$, which can be isolated three parts as

$$\begin{aligned} \mathbb{E}\left\{\|\mathbf{u}_{4k}\|^2\right\} &= \mathbb{E}\left\{\left\|\hat{\mathbf{f}}_k^H\mathbf{W}\right\|^2\right\} \\ &= \underbrace{\mathbb{E}\left\{\left\|c_k\sqrt{q_k\tau_c}\mathbf{f}_k^H\mathbf{W}\right\|^2\right\}}_{F} \\ &\quad + \underbrace{\mathbb{E}\left\{\left\|c_k\sum\nolimits_{i=1}^{K}\sqrt{\frac{p_i}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{W}\right\|^2\right\}}_{G} \\ &\quad + \underbrace{\mathbb{E}\left\{\left\|c_k\boldsymbol{\psi}_k^H\frac{\mathbf{W}^H\mathbf{W}}{\sqrt{\tau_c}}\right\|^2\right\}}_{H}, \end{aligned} \tag{82}$$

where F, G, and H in (82) can be respectively given as

$$F = Mc_k^2 q_k\tau_c^2(\xi_k+\gamma_k), \tag{83}$$

$$\begin{aligned} G &= \frac{c_k^2}{\tau_c}\sum\nolimits_{i=1}^{K} p_i\mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{W}\mathbf{W}^H\mathbf{f}_i\mathbf{s}_i^H\boldsymbol{\psi}_k\right\} \\ &= Mc_k^2\tau_c\sum\nolimits_{i=1}^{K} p_i(\xi_i+\gamma_i), \end{aligned} \tag{84}$$

$$H = c_k^2 M\left(M+\tau_c\right). \tag{85}$$

Combining (82)−(85), we can obtain the expression of $\mathbb{E}\{\|\mathbf{u}_{4k}\|^2\}$. Next, we focus on $\|\mathbb{E}\{\mathbf{u}_{ik}\}\|^2$, $1\le i\le 4$. Here we omit their detailed process since they are straightforward.

$$\begin{aligned} \|\mathbb{E}\{\mathbf{u}_{1k}\}\|^2 &= c_k^2 p_k\left\|\mathbb{E}\left\{\sqrt{\frac{p_k}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_k\mathbf{f}_k^H\mathbf{f}_k\mathbf{s}_k^H\right\}\right\|^2 \\ &= M^2 p_k^2 c_k^2(\xi_k+\gamma_k)^2, \end{aligned} \tag{86}$$

$$\begin{aligned} \|\mathbb{E}\{\mathbf{u}_{2k}\}\|^2 &= \left\|\mathbb{E}\left\{c_k\sqrt{p_i}\sum\nolimits_{i\neq k}^{K}\sqrt{\frac{p_i}{\tau_c}}\boldsymbol{\psi}_k^H\mathbf{s}_i\mathbf{f}_i^H\mathbf{f}_i\mathbf{s}_i^H\right\}\right\|^2 \\ &= M^2c_k^2\left(\sum\nolimits_{i\neq k}^{K} p_i(\xi_i+\gamma_i)\right)^2, \end{aligned} \tag{87}$$

$$\|\mathbb{E}\{\mathbf{u}_{3k}\}\|^2 = 0, \tag{88}$$

$$\|\mathbb{E}\{\mathbf{u}_{4k}\}\|^2 = M^2c_k^2. \tag{89}$$

Similar to the above derivations, the diploid terms in (71) are derived as follows

$$\begin{aligned} &2\mathrm{Re}\left\{\sum\nolimits_{i=1}^{4}\sum\nolimits_{j=i+1}^{4}\mathbb{E}\{\mathbf{u}_{ik}\}\mathbb{E}\{\mathbf{u}_{jk}^H\}\right\} \\ &= 2\mathrm{Re}\left\{\mathbb{E}\{\mathbf{u}_{1k}\}\mathbb{E}\{\mathbf{u}_{2k}^H\} + \mathbb{E}\{\mathbf{u}_{1k}\}\mathbb{E}\{\mathbf{u}_{4k}^H\}\right. \\ &\quad \left.+\mathbb{E}\{\mathbf{u}_{2k}\}\mathbb{E}\{\mathbf{u}_{4k}^H\}\right\} \\ &= 2M^2c_k^2p_k(\xi_k+\gamma_k)\sum\nolimits_{i\neq k}^{K} p_i(\xi_i+\gamma_i) \\ &\quad + 2M^2c_k^2p_k(\xi_k+\gamma_k) + 2M^2c_k^2\sum\nolimits_{i\neq k}^{K} p_i(\xi_i+\gamma_i), \end{aligned} \tag{90}$$

$$\begin{aligned} &2\mathrm{Re}\left\{\mathbb{E}\left\{\sum\nolimits_{i=1}^{4}\sum\nolimits_{j=i+1}^{4}\mathbf{u}_{ik}\mathbf{u}_{jk}^H\right\}\right\} \\ &= 2\mathrm{Re}\left\{\mathbb{E}\{\mathbf{u}_{1k}\mathbf{u}_{2k}^H\} + \mathbb{E}\{\mathbf{u}_{1k}\mathbf{u}_{4k}^H\} + \mathbb{E}\{\mathbf{u}_{2k}\mathbf{u}_{4k}^H\}\right\} \\ &= 2Mc_k^2p_k\sum\nolimits_{i\neq k}^{K} p_i\left(\Xi_{ik} + M\gamma_k\xi_i + M\gamma_i\xi_k + M\gamma_k\gamma_i\right) \\ &\quad + 2M^2c_k^2p_k\left(\xi_k+\gamma_k\right) + 2M^2c_k^2\sum\nolimits_{i\neq k}^{K} p_i\left(\xi_i+\gamma_i\right). \end{aligned} \tag{91}$$

Finally, substituting $\mathbb{E}\{\|\mathbf{u}_{ik}\|^2\}$, $1\le i\le 4$, and (86)−(91) into (18) and performing some algebra simplifications, we can complete the derivation of SINR expression.

REFERENCES

[1] Hamidi-Sepehr *et al.*, "5G URLLC: Evolution of high-performance wireless networking for industrial automation," *IEEE Commun. Stand. Mag.*, vol. 5, no. 2, pp. 132–140, Jun. 2021.

[2] D. Feng, C. She, K. Ying, L. Lai, Z. Hou, T. Q. S. Quek, Y. Li, and B. Vucetic, "Toward ultrareliable low-latency communications: Typical scenarios, possible solutions, and open issues," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 94–102, Jun. 2019.

[3] G. J. Sutton *et al.*, "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, 3rd Quart. 2019.

[4] P. Schulz *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[5] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[6] D. Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[7] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[8] X. Yuan *et al.*, "Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 136–143, Apr. 2021.

[9] Q. Peng, H. Ren, C. Pan, M. Elkashlan, A. G. Armada, and P. Popovski, "Two-timescale design for reconfigurable intelligent surface-aided URLLC," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 13 664–13 677, Oct. 2024.

[10] R. Hashemi, S. Ali, N. H. Mahmood, and M. Latva-Aho, "Joint sum rate and blocklength optimization in RIS-Aided short packet URLLC systems," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1838–1842, Aug. 2022.

[11] Z. Li, H. Shen, W. Xu, P. Zhu, and C. Zhao, "Resource allocation for IRS-assisted uplink URLLC systems," *IEEE Commun. Lett.*, vol. 27, no. 6, pp. 1540–1544, Jun. 2023.

[12] J. Cheng, C. Shen, Z. Chen, and N. Pappas, "Robust beamforming design for IRS-aided URLLC in D2D networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6035–6049, Sept. 2022.

[13] L. Zhou, J. Dai, and W. Xu, "Joint multi-user channel estimation for RIS-assisted massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 13 993–14 006, Oct. 2024.

[14] X. Wei, D. Shen, and L. Dai, "Channel estimation for RIS assisted wireless communications-part II: An improved solution based on double-structured sparsity," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1403–1407, May 2021.

[15] C. She, C. Sun, Z. Gu, Y. Li, C. Yang *et al.*, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proc. IEEE*, vol. 109, no. 3, pp. 204–246, Mar. 2021.

[16] D. Verenzuela, E. Björnson, and L. Sanguinetti, "Spectral and energy efficiency of superimposed pilots in uplink massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7099–7115, Nov. 2018.

[17] R. Shafin and L. Liu, "Superimposed pilot for multi-cell multi-user massive FD-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3591–3606, May 2020.

[18] Q. Sun, X. Ji, Z. Wang, X. Chen, Y. Yang, J. Zhang, and K.-K. Wong, "Uplink performance of hardware-impaired cell-free massive MIMO with multi-antenna users and superimposed pilots," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6711–6726, Nov. 2023.

[19] D. Verenzuela, E. Björnson, X. Wang, M. Arnold, and S. ten Brink, "Massive-MIMO iterative channel estimation and decoding (MICED) in the uplink," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 854–870, Feb. 2020.

[20] L. Xu, J. Jiao, Y. Wang, S. Wu, R. Lu, and Q. Zhang, "Low-correlation superimposed pilot grant-free massive access for satellite internet of things," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7087–7101, Dec. 2023.

[21] X. Zhou, W. Xia, Q. Zhang, J. Zhang, and H. Zhu, "Power allocation of superimposed pilots for URLLC with short-packet transmission in IIoT," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2365–2369, Nov. 2022.

[22] N. Garg, H. Ge, and T. Ratnarajah, "Generalized superimposed training scheme in IRS-assisted cell-free massive MIMO systems," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 1157–1171, Aug. 2022.

[23] P. Popovski, Č. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[24] J. Ding, M. Nemati, S. R. Pokhrel, O.-S. Park, J. Choi, and F. Adachi, "Enabling grant-free URLLC: An overview of principle and enhancements by massive MIMO," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 384–400, Jan. 2022.

[25] K. Zhi, C. Pan, G. Zhou, H. Ren, M. Elkashlan, and R. Schober, "Is RIS-aided massive MIMO promising with ZF detectors and imperfect CSI?" *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 3010–3026, Oct. 2022.

[26] Q.-U.-A. Nadeem, A. Zappone, and A. Chaaban, "Achievable rate analysis and max-min SINR optimization in intelligent reflecting surface assisted cell-free MIMO uplink," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1295–1322, 2022.

[27] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. Di Renzo, and B. Ottersten, "Reconfigurable intelligent surface-assisted cell-free massive MIMO systems over spatially-correlated channels," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5106–5128, Jul. 2022.

[28] Y. Zhang, W. Xia, H. Zhao, G. Zheng, S. Lambotharan, and L. Yang, "Performance analysis of RIS-assisted cell-free massive MIMO systems with transceiver hardware impairments," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7258–7272, Dec. 2023.

[29] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[30] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[31] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[32] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[33] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, Jan. 2013.

[34] Z. Peng, R. Weng, C. Pan, G. Zhou, M. D. Renzo, and A. L. Swindlehurst, "Robust transmission design for RIS-Assisted secure multiuser communication systems in the presence of hardware impairments," *IEEE Trans. on Wireless Commun.*, vol. 22, no. 11, pp. 7506–7521, 2023.

[35] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1998.

[36] K. Zhi, C. Pan, H. Ren, and K. Wang, "Power scaling law analysis and phase shift optimization of RIS-Aided massive MIMO systems with statistical CSI," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3558–3574, May 2022.

[37] J. Dai, F. Zhu, C. Pan, H. Ren, and K. Wang, "Statistical CSI-based transmission design for reconfigurable intelligent surface-aided massive MIMO systems with hardware impairments," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 38–42, Jan. 2022.

[38] K. Zhi, C. Pan, H. Ren, and K. Wang, "Power scaling law analysis and phase shift optimization of RIS-aided massive MIMO systems with statistical CSI," *2020, arXiv: 2010.13525*.

[39] Z. Ye, C. Pan, H. Zhu, and J. Wang, "Tradeoff caching strategy of the outage probability and fronthaul usage in a Cloud-RAN," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6383–6397, Jul. 2018.

[40] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optim. Eng.*, vol. 8, no. 1, pp. 67–127, May 2007.

[41] M. Chiang, C. W. Tan, D. P. Palomar, D. O'neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, Jul. 2007.

[42] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. (2016). [Online]. Available: http://cvxr.com/cvx

[43] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint pilot design and uplink power allocation in multi-cell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2000–2015, Mar. 2018.

[44] K. Zhi, C. Pan, H. Ren, and K. Wang, "Statistical CSI-based design for reconfigurable intelligent surface-aided massive MIMO systems with direct links," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 1128–1132, May 2021.

[45] X. Zhou, W. Xia, J. Zhang, W. Wen, and H. Zhu, "Joint optimization of frame structure and power allocation for URLLC in short blocklength regime," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7333–7346, Dec. 2023.

[46] X. Zhou, Y. Zhu, W. Xia, J. Zhang, and K.-K. Wong, "Optimized payload length and power allocation for generalized superimposed pilot in URLLC transmissions," *IEEE Trans. Commun.*, vol. 72, no. 10, pp. 6073–6086, Oct. 2024.