Optimized Payload Length and Power Allocation for Generalized Superimposed Pilot in URLLC Transmissions

Xingguang Zhou, *Graduate Student Member, IEEE*, Yongxu Zhu, *Senior Member, IEEE*, Wenchao Xia, *Member, IEEE*, Jun Zhang, *Senior Member, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

***Abstract*—Ultra-reliable and low-latency communication (URLLC) is recognized as the most challenging use case for the next generation of wireless networks. Existing research on URLLC is based on the regular pilot (RP) scheme, which is tough to ensure a high transmission rate with stringent latency and reliability requirements due to the impact of finite blocklength, especially in massive connectivity scenarios. In this paper, we propose to use generalized superimposed pilot (GSP) scheme for URLLC transmission in massive multi-input multi-output (mMIMO) systems. Distinguishing from the conventional superimposed pilot (SP) scheme, the GSP scheme eliminates mutual interference between the pilot and data, where the data length is optimized, and the data symbols are precoded to spread over the whole transmission block. With the GSP scheme, we first formulate a weighted sum rate maximization problem by jointly optimizing the data length, pilot power, and data power and then derive closed-form results, including suboptimal data length and achievable rate lower bounds with maximum-ratio combining (MRC) and zero-forcing (ZF) detectors, respectively. Based on the closed-form results, we provide the corresponding iterative algorithms for the MRC and ZF cases where the problems are transformed into geometry program format by using log-function and successive convex approximation methods. Finally, the performance of the RP, SP, and GSP schemes are compared through simulation results, which reflect the superiority and robustness of the GSP scheme in URLLC scenarios.**

***Index Terms*—URLLC, massive MIMO, superimposed pilot, finite blocklength**

I. Introduction

With the large-scale commercialization of 5G, the global industry has embarked on research and exploration of the next-generation mobile communication technology (6G). Looking towards the year 2030 and beyond, human society is entering an era of ubiquitous connectivity, where numerous emerging applications impose increasingly stringent requirements on networks in terms of latency and reliability. [2]. Hence, ultra-reliable and low-latency communication (URLLC) is a critical use case of the emerging 6G systems, which will enable various applications such as intelligent traffic management, metaverse [3], and tactile Internet [4]. Typical key performance indicator (KPI) for URLLC of 5G refers to 1-millisecond end-to-end latency and 10^{-5} decoding error rate for a packet with 32 bytes. Practically, the KPI of URLLC varies for different applications. For instance, industrial automation has reliability requirement of $1-10^{-9}$ at latency 0.25$\sim$10 ms with data size of 80$\sim$1000 bytes when the communication range is between 10 to 1000 m, while smart grid has more relaxed latency (3$\sim$20 ms) and reliability ($1-10^{-5}$) requirements [5]. These KPIs will be improved by one order of magnitude in 6G. However, the relevant research is still in its infancy since the targets above are conflicting and challenging to satisfy at the same time.

Low latency implies a finite number of channel uses (blocklength). In other words, the packet size or the codeword length is very short, which is different from traditional communications where the number of channel uses goes to infinity [6]. Hence, in the regime of short packet transmission, the classic Shannon theorem is not applicable anymore. Fortunately, the finite blocklength information theorem has made great progress in recent years, which reveals that short packet incurs a degradation of transmission rate and non-negligible decoding error. This phenomenon can be mathematically characterized by the URLLC achievable rate, which manifests that the degradation is inversely proportional to the blocklength [7]. Therefore, how to reduce the influence of finite blocklength such that the throughput of URLLC can be improved is a crucial issue in future research of 6G.

A. Related Works

Channel state information (CSI) plays a critical role in mandating URLLC. The channel estimation accuracy determines the level of signal-to-interference-plus-noise ratio (SINR), influencing the reliability and transmission rate. Many existing works of URLLC with short packet transmission assume that perfect CSI can be available by using ignorable pilot overhead. For example, in [8], [9], He and Nasir *et al.* investigated the resource allocation and beamforming design in the finite blocklength under different scenarios. Later, [10] discussed the rate and energy efficiency of URLLC in cell-free massive multi-input multi-output (mMIMO) by using a

This work was supported in part by the National Natural Science Foundation of China under Grant 92367302, Grant 62201285, and Grant 62071247; in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2020084-1; in part by the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grant BK20212001; in part by the China Postdoctoral Science Foundation under Grant 2023M744104 and Grant 2022M722669, and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX22_0943. (Corresponding author: Wenchao Xia)

Parts of this work have been published at the IEEE Global Communications Conference (GLOBECOM) Workshops 2023 [1].

Xingguang Zhou, Wenchao Xia, and Jun Zhang 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, zhangjun@njupt.edu.cn).

Y. Zhu is with the National Communications Research Laboratory, Southeast University, Nanjing 210096, China, (e-mail: yongxu.zhu@seu.edu.cn).

Kai-Kit Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: kai-kit.wong@ucl.ac.uk).

low complexity conjugate beamforming scheme. Intelligent reflecting surface was also applied to assist URLLC and various cases were considered in [11]. However, the above assumption is impractical since the impact of pilot overhead will be more significant when the packet length is shortened. More practically, some works have studied pilot-based short packet transmission for URLLC. In [12], joint power control for uplink URLLC in a cell-free mMIMO system was investigated. A low-complexity algorithm was developed to allocate the channel uses for the pilot and data in [13]. Moreover, [14] claimed that the imperfection of channel reciprocity has an influence on downlink URLLC when channel inversion power control is adopted. The performance of two different short frame structures in the downlink URLLC transmission was compared in [15].

In the aforementioned literature, the conventional regular pilot (RP) scheme is exploited to send short packets where pilot and data are separately transmitted. Then, the data blocklength becomes smaller, which results in more serious rate degradation, especially when the pilot length linearly increases with the number of users [16]. Although pilot-free scheme has been studied to replace the pilot-based scheme in short packet transmission [17], [18], the reliability of message decoding is still questionable due to the absence of channel knowledge. Hence, novel transmission strategy should be proposed to replace the conventional RP scheme.

Superimposed transmission schemes such as sparse vector coding-based sparse superimposed transmission and superimposed pilot (SP) have been studied to enable URLLC in terms of channel coding [19] and channel estimation [20, 21], whose similarity lies in guaranteeing low latency by superimposing multiple independently transmitted signals. In this paper, we focus on applying superimposed transmission in the channel estimation of URLLC. The primary feature of SP is that the pilot and data occupy the same transmission block, which implicitly indicates a larger number of channel uses for data and more available pilot sequences for connected users [22]. Our previous work [20] has verified that the influence of finite blocklength can be reduced by using the SP scheme, especially in the massive connectivity setting. However, the SP scheme suffers from mutual interference (MI) between the pilot and payload, impairing channel estimation and signal detection quality. To reduce MI, some works have proposed effective solutions in mMIMO systems. In [21], we suppressed the MI by jointly optimizing the superimposed length, pilot power and data power. In [23], the authors used an iterative channel estimation approach to mitigate the MI of the SP scheme in mMIMO systems. Also, [24] and [25] proposed combining the advantages of the RP and SP schemes to obtain better performance.

Although the above research has shown effectiveness of the solutions, the impact of MI can only be alleviated not eliminated. The generalized superimposed pilot (GSP) in [26] scheme has been recently proposed to enhance the performance further. The main idea of the GSP scheme is that instead of sending a data sequence of the same length as the pilot sequence, the data sequence is shortened, and the data symbols are precoded, to reduce the correlation between the pilot and data. Thus, the MI can be removed by optimizing the data length. In [27], the performance of the GSP scheme in cell-free mMIMO systems with centralized and distributed modes are respectively analyzed, and the work was extended to the more complicated system with intelligent reflective surface in [28]. However, these researches are based on the infinite blocklength regime and the optimal payload length has not been given in a closed form. Although the performance of the GSP scheme with maximum-ratio combining (MRC) receiver has been studied in the finite blocklength regime [1], the merit of the GSP scheme for URLLC is not well understood.

B. Contributions

Motivated by the aforementioned facts, this paper investigates the short packet transmission performance of the GSP scheme in a multi-user uplink mMIMO system, with given latency and decoding error probability and then conduct performance comparisons of the RP, SP, and GSP schemes. Specifically, the contributions are summarized as follows:

- To the best of our knowledge, we are the first to study the GSP scheme for uplink mMIMO to meet the KPIs of URLLC. We avoid MI by ensuring the orthogonality between the precoding matrix and pilot. The imperfect channel estimation and two detectors, i.e., MRC and zero-forcing (ZF), are considered in this paper. We propose a weighted sum rate maximization problem under given delay and reliability targets, where the variables including data length, pilot power, and data power are jointly optimized subject to the minimum rate and energy constraints. To simplify the problem, the closed-form achievable rate lower bounds (LBs) for the MRC and ZF cases with finite blocklength are derived.
- For the MRC detector, the optimization problem is a mixed-integer nonlinear programming problem and hard to obtain a globally optimal solution. We first derive the suboptimal data length in a closed form, which shifts our focus to power allocation. Next, we utilize log-function to iteratively approximate the achievable rate LB, which facilitates that the problem can be converted into a sequence of geometric program (GP) problems. Finally, we propose an iterative algorithm to find a locally optimal solution, and the convergence analysis is also provided.
- For the ZF detector, we prove that the suboptimal data length has the same closed-form solution as that in the MRC case. The problem can not be transformed into the GP format directly due to the complicated SINR expression. To address the issue, we adopt successive convex approximation (SCA) to construct monomial functions to approximate the posynomial functions in each iteration. Then, the optimization problem can also be transformed into a sequence of GP problems, in which the corresponding algorithm and convergence analysis are provided.
- In our simulations, the conventional RP [15] and SP [22] schemes are used as the benchmarks for comparison in the same setup . The results suggest that the GSP scheme

is superior to the RP and SP schemes in the MRC case and outperforms the SP scheme in the ZF case. Specifically, with fixed delay and error probability, the GSP scheme shows that more gain can be obtained in the low energy, small channel uses, and large user number regions, respectively, which emphasizes the robustness of the GSP scheme in URLLC.

The remainder of this paper is organized as follows. In Section II, the system model and problem formulation are provided. In Section III and Section IV, we derive the suboptimal data length and design the corresponding iterative algorithms to optimize power allocation for the MRC and ZF cases, respectively. The simulation results and analysis are given in Section V. Finally, Section VI draws conclusions.

Notation: We use lower-case bold-face letters to denote vectors (e.g., $\mathbf{x}$), while bold-face capital letters are used for matrices (e.g., $\mathbf{X}$). $\mathbf{x}^H$, $\|\mathbf{x}\|$, $\mathbb{E}\{\mathbf{x}\}$, $tr\{\mathbf{x}\}$, and $\mathbf{R_e}\{\mathbf{x}\}$ represent the Hermitian transpose, Euclidean norm, expectation, trace, and real part operation of vector $\mathbf{x}$, respectively. $\mathcal{D}(\mathbf{X})$ and $\mathbf{X}^{-1}$ denote a diagonal matrix with the diagonal components and the inverse matrix of $\mathbf{X}$, respectively. $\mathcal{CN}(\cdot,\cdot)$ and $\mathbb{C}$ represent a circularly symmetric complex Gaussian distribution and the set of complex numbers, respectively.

II. System Model and Problem Formulation

A. System Model

This paper considers the uplink short packet transmission in a single-cell system consisting of one M-antenna base station (BS) and K single-antenna users. For simplicity, we denote the set of users as $\mathcal{K} = \{1, 2, \cdots, K\}$. The short packets are transmitted simultaneously by the users utilizing the same bandwidth B Hz. To achieve URLLC, we stipulate that each transmission block contains no more than T channel uses for the given decoding error probability ε. Note that our system model can be applied to typical local or wide area URLLC communication scenarios [29].

B. Channel Model

Generally, the short packet transmission time is greatly small, in comparison with the channel coherence time. Hence, the channel is modelled as a quasi-static fading channel in which the channel $\mathbf{h}_i \in \mathbb{C}^{M\times 1}$ is constant during the same transmission block but varies in different blocks. We assume that the channel fading follows Rayleigh distribution, i.e., $\mathbf{h}_i \sim \mathcal{CN}(\mathbf{0}, \alpha_i \mathbf{I}_M)$, where α_i denotes the statistic CSI between the i-th user and the BS.

C. GSP Scheme

To guarantee latency and reliability, the GSP training scheme is employed instead of the conventional RP scheme. The frame structure of the GSP scheme is shown in Fig. 1. As presented in Fig. 1, the transmitted signal $\mathbf{z}_i \in \mathbb{C}^{T\times 1}$ is a superimposition of the pilot and precoded data. In this paper, we consider the signal in a period of T channel uses. Let τ and $\mathbf{W}_i \in \mathbb{C}^{T\times\tau}$ denote the data length and the orthogonal precoder matrix, respectively. The transmitted signal $\mathbf{z}_i$ can be written as

$$\mathbf{z}_i = \sqrt{\rho_i}\boldsymbol{\varphi}_i + \sqrt{\eta_i}\mathbf{W}_i\mathbf{s}_i, \tag{1}$$

where ρ_i and η_i are the normalized transmitting power on the pilot and data for the i-th user, respectively, $\boldsymbol{\varphi}_i \in \mathbb{C}^{T\times 1}$ denotes the orthogonal pilot vector, and $\mathbf{s}_i \in \mathbb{C}^{\tau\times 1}$ is the data vector following the distribution of $\mathcal{CN}\left(\mathbf{0}, \frac{1}{\tau}\mathbf{I}_\tau\right)$. Then, the received signal at the BS is given by

$$\mathbf{Y} = \sum\nolimits_{i\in\mathcal{K}} \mathbf{h}_i\mathbf{z}_i^H + \mathbf{N}, \tag{2}$$

where $\mathbf{N} \in \mathbb{C}^{M\times T}$ is the additive white Gaussian noise (AWGN) matrix and its each entry follows the distribution of $\mathcal{CN}(0,1)$.

Under the superimposed scheme, the data and pilot will interfere with each other in channel estimation and data detection [22]. To ensure transmission reliability, in this paper, we eliminate the MI by designing the precoding matrix carefully.

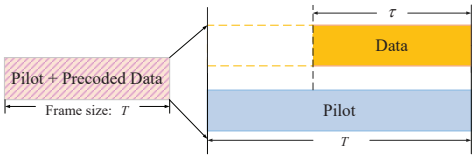


Fig. 1. Frame structure of generalized superimposed pilot.

1) Precoding Matrix Design: We assume $K < T$ such that the BS can choose K columns as pilots for the users from the $T \times T$ orthogonal matrix. For the i-th user, the precoding matrix $\mathbf{W}_i$ is chosen to satisfy the orthogonality with pilot $\boldsymbol{\varphi}_j$, i.e., $\mathbf{W}_i^H\boldsymbol{\varphi}_j = \mathbf{0}$. Specifically, $\mathbf{W}_i$ is obtained by randomly selecting τ columns from the remaining $T-K$ columns. In this case, there is no MI between the data and pilot. In particular, when $\tau = T-K$, all the users have the same precoding matrix.

Using the orthogonality of pilot, we have the following conclusion:

$$\mathbf{W}_i^H\mathbf{W}_j = \begin{cases} T\boldsymbol{\Phi}_{ij}, & i \neq j \\ T\mathbf{I}_\tau, & i = j \end{cases}, \tag{3}$$

where $\boldsymbol{\Phi}_{ij}$ is a permutation matrix with rank r_{ij}. The rank r_{ij} of the matrix $\boldsymbol{\Phi}_{ij}$ implies $\mathbf{W}_i$ and $\mathbf{W}_j$ having r_{ij} identical columns. For each user, the transmit power has the following constraint:

$$\begin{aligned} \mathbb{E}\left\{\|\mathbf{z}_i\|^2\right\} &= \mathbb{E}\left\{\|\sqrt{\rho_i}\boldsymbol{\varphi}_i + \mathbf{W}_i\sqrt{\eta_i}\mathbf{s}_i\|^2\right\} \\ &= \rho_i\|\boldsymbol{\varphi}_i\|^2 + \eta_i\mathbb{E}\left\{\|\mathbf{W}_i\mathbf{s}_i\|^2\right\} \\ &= T(\rho_i + \eta_i). \end{aligned} \tag{4}$$

D. Channel Estimation

At the BS, the signal of the k-th user is extracted by multiplying $\mathbf{Y}$ with $\boldsymbol{\varphi}_k \Big/ \sqrt{T}$, which yields

$$\begin{aligned} \mathbf{y}_k &= \mathbf{Y}\frac{\boldsymbol{\varphi}_k}{\sqrt{T}} \\ &= \sum_{i\in\mathcal{K}} \sqrt{\rho_i}\mathbf{h}_i\boldsymbol{\varphi}_i^H\frac{\boldsymbol{\varphi}_k}{\sqrt{T}} + \sum_{i\in\mathcal{K}}\sqrt{\eta_i}\mathbf{h}_i\mathbf{s}_i^H\mathbf{W}_i^H\frac{\boldsymbol{\varphi}_k}{\sqrt{T}} + \mathbf{N}\frac{\boldsymbol{\varphi}_k}{\sqrt{T}} \\ &= \sqrt{\rho_k T}\mathbf{h}_k + \mathbf{n}_k, \end{aligned} \tag{5}$$

where $\mathbf{n}_k = \mathbf{N}\frac{\boldsymbol{\varphi}_k}{\sqrt{T}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. From (5), it can be seen that there is no interference from data, and $\mathbf{y}_k$ is still a Gaussian signal. Then, we use the minimum mean squared error (MMSE) estimator to obtain the corresponding channel estimation

$$\hat{\mathbf{h}}_k = \lambda_k \mathbf{y}_k, \tag{6}$$

in which

$$\lambda_k = \frac{\sqrt{\rho_k T}\alpha_k}{\rho_k T \alpha_k + 1}. \tag{7}$$

It should be noted that $\hat{\mathbf{h}}_k$ and estimation error are independent of each other in MMSE estimation. The covariance matrices of $\hat{\mathbf{h}}_k$ and estimation error $\boldsymbol{\epsilon}_k = \mathbf{h}_k - \hat{\mathbf{h}}_k$ are, respectively, given by

$$\mathbb{E}\left\{\hat{\mathbf{h}}_k \hat{\mathbf{h}}_k^H\right\} = \beta_k \mathbf{I}_M, \tag{8}$$

$$\mathbb{E}\left\{\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^H\right\} = (\alpha_k - \beta_k)\mathbf{I}_M, \tag{9}$$

where $\beta_k = \sqrt{\rho_k T}\lambda_k \alpha_k$.

E. Data Detection

The BS performs data detection with conventional linear detector after channel estimation. The cases of MRC and ZF detectors are discussed, respectively, in the following.

Denote $\hat{\mathbf{H}} = \left[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \cdots \hat{\mathbf{h}}_K\right]$ as the channel estimation matrix. Based on $\hat{\mathbf{H}}$, the received matrices of the MRC and ZF detectors are given by

$$\boldsymbol{\Psi} = \begin{cases} \hat{\mathbf{H}}, & \text{for MRC,} \\ \hat{\mathbf{H}}\left(\hat{\mathbf{H}}^H \hat{\mathbf{H}}\right)^{-1}, & \text{for ZF.} \end{cases} \tag{10}$$

Then, using the matrix to process the received signal, we obtain

$$\hat{\mathbf{S}}^H = \boldsymbol{\Psi}^H \mathbf{Y}. \tag{11}$$

Finally, the estimated data of the k-th user is obtained by multiplying the precoded matrix $\mathbf{W}_k$.

$$\begin{aligned} \hat{\mathbf{s}}_k^H &= \boldsymbol{\psi}_k^H \sum\nolimits_{i\in\mathcal{K}} \sqrt{\rho_i}\mathbf{h}_i \boldsymbol{\varphi}_i^H \mathbf{W}_k \\ &\quad + \boldsymbol{\psi}_k^H \sum\nolimits_{i\in\mathcal{K}} \sqrt{\eta_i}\mathbf{h}_i \mathbf{s}_i^H \mathbf{W}_i^H \mathbf{W}_k + \boldsymbol{\psi}_k^H \mathbf{N}\mathbf{W}_k, \end{aligned} \tag{12}$$

where $\boldsymbol{\psi}_k$ denotes the k-th column of $\boldsymbol{\Psi}$. To derive the SINR of mMIMO systems, the effective channel gain can be approximated by its mean value, which is referred to as the use-and-then-forget (UatF) technique and very accurate with the channel hardening effect [30]. Thus, we use the technique to rewrite (12) as

$$\begin{aligned} \hat{\mathbf{s}}_k^H &= \sqrt{\eta_k}\mathbb{E}\left\{\mathbf{a}_k^H \mathbf{h}_k\right\} \mathbf{s}_k^H \mathbf{W}_k^H \mathbf{W}_k \\ &\quad + \sqrt{\eta_k}\left(\mathbf{a}_k^H \mathbf{h}_k - \mathbb{E}\left\{\mathbf{a}_k^H \mathbf{h}_k\right\}\right) \mathbf{s}_k^H \mathbf{W}_k^H \mathbf{W}_k + \boldsymbol{\omega}_k, \end{aligned} \tag{13}$$

where the combining vector $\mathbf{a}_k$ and effective noise $\boldsymbol{\omega}_k$ are respectively defined as

$$\mathbf{a}_k = \begin{cases} \beta_k \alpha_k^{-1} \mathbf{h}_k, & \text{for MRC,} \\ \boldsymbol{\psi}_k, & \text{for ZF,} \end{cases} \tag{14}$$

$$\boldsymbol{\omega}_k = \begin{cases} \sqrt{\eta_k}\bar{\mathbf{h}}_k^H \mathbf{h}_k \mathbf{s}_k^H \mathbf{W}_k^H \mathbf{W}_k \\ \quad + \sqrt{\eta_i}\sum_{i\in\mathcal{K}\setminus k} \hat{\mathbf{h}}_k^H \mathbf{h}_i \mathbf{s}_i^H \mathbf{W}_i^H \mathbf{W}_k \\ \quad + \hat{\mathbf{h}}_k^H \mathbf{N}\mathbf{W}_k, & \text{for MRC,} \\ \sqrt{\eta_i}\sum_{i\in\mathcal{K}\setminus k} \mathbf{a}_k^H \mathbf{h}_i \mathbf{s}_i^H \mathbf{W}_i^H \mathbf{W}_k \\ \quad + \mathbf{a}_k^H \mathbf{N}\mathbf{W}_k, & \text{for ZF,} \end{cases} \tag{15}$$

with $\bar{\mathbf{h}}_k = \hat{\mathbf{h}}_k - \lambda_k\sqrt{\rho_k T}\mathbf{h}_k$. From (15), it can be observed that the interference term related to pilot in (12) is removed. Thus, the effective SINR of the k-th user can be expressed as

$$\gamma_k = \frac{\eta_k T^2 \left|\mathbb{E}\left\{\mathbf{a}_k^H \mathbf{h}_k\right\}\right|^2}{\eta_k T^2 \mathrm{Var}\left(\mathbf{a}_k^H \mathbf{h}_k\right) + \mathbb{E}\left\{\left\|\boldsymbol{\omega}_k{}^H - \mathbb{E}\left\{\boldsymbol{\omega}_k{}^H\right\}\right\|^2\right\}}. \tag{16}$$

F. Problem Formulation

In the short packet transmission with the GSP scheme, the ergodic achievable rate depends on not only the SINR but also the number of transmitted data symbols, the number of channel uses, and decoding error probability. The relationship among them can be characterized as follows in the unit of bits/channel use [7], i.e.,

$$R_k = \frac{\tau}{T}\mathbb{E}\left[\log_2(1+\Gamma_k) - \frac{Q^{-1}(\varepsilon)}{\ln 2\sqrt{T}}\sqrt{V(\Gamma_k)}\right], \tag{17}$$

where $V(\Gamma_k) = 1 - \frac{1}{(1+\Gamma_k)^2}$, Γ_k is the instantaneous SINR for the k-th user, $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian Q-function. Here, the number of channel uses is equal to T because the τ data symbols are spread over the whole transmission block under the GSP scheme[1].

It should be note that the channel uses means a latency constraint. Specifically, the channel uses is equal to the product of transmission duration (latency) and system bandwidth. Hence, given bandwidth B Hz and channel uses T, the latency is guaranteed [6]. In this paper, we consider the weighted sum rate maximization problem where the data length, pilot power, and data power are jointly optimized. Based on (17), the problem can be formulated as

$$\mathbf{P1}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\tau} \sum\nolimits_{k\in\mathcal{K}} \mu_k R_k \tag{18a}$$

$$\text{s.t. } R_k \geqslant R_{\min},\ \forall k, \tag{18b}$$

$$T(\rho_k + \eta_k) \leqslant E,\ \forall k, \tag{18c}$$

$$1 \leqslant \tau \leqslant T - K, \tau \in \mathbf{N}^+, \tag{18d}$$

where $\boldsymbol{\rho} = \{\rho_k, \forall k\}$, $\boldsymbol{\eta} = \{\eta_k, \forall k\}$, μ_k is the weight of the k-th user. Obviously, the rate expression (17) can take a negative value which is not practically possible. Thus, it is general to set a minimum rate $R_{\min}$ to impose a constraint (18b) on each user. Besides, (18c) is the energy constraint, and (18d) means that data length τ is an integer and can not be more than $T-K$. Otherwise, the SINR will deteriorate sharply since the MI can not be cancelled completely.

Finding a globally optimal solution for a mixed-integer non-linear programming problem such as $\mathbf{P1}$ is very challengeable. Instead, we seek the locally optimal solution to $\mathbf{P1}$ by means of efficient algorithms in an iterative manner.

[1]In the RP scheme, the number of channel uses is equal to the data length τ [16], which is less than that in the GSP scheme. Accordingly, the impact of finite blocklength is more significant and causes a larger rate loss in (17).

III. Joint Optimization for MRC

In this section, we focus on the optimization problem in (18) with the MRC detector. Unfortunately, deriving the closed-form expression of the ergodic achievable rate is extremely difficult. In contrast, its LB is readily available [31] and can be expressed as

$$R_k \geqslant \hat{R}_k \triangleq \frac{\tau}{T\ln 2}\Phi\left(\mathbb{E}\left\{\left(\Gamma_k\right)^{-1}\right\}\right), \tag{19}$$

where $\Phi(x) = \ln(1+1/x) - \frac{Q^{-1}(\varepsilon)}{\sqrt{T}}\sqrt{V(1/x)}$. Based on (16) and (19), we introduce the following theorem:

Theorem 1: For the ergodic achievable rate of the k-th user in URLLC transmission with the GSP scheme and MRC detector, its LB can be expressed as

$$\hat{R}_k \triangleq \frac{\tau}{T\ln 2}\Phi\left(\gamma_k^{-1}\right), \tag{20}$$

where the effective SINR (16) is further written as

$$\gamma_k = \frac{M\rho_k\eta_k\alpha_k^2}{\left(\rho_k\alpha_k + \frac{1}{T}\right)\left(\eta_k\alpha_k + \frac{1}{\tau}\sum_{i\in\mathcal{K}\setminus k}\eta_i\alpha_i r_{ik} + \frac{\tau}{T}\right)}, \tag{21}$$

where r_{ik} denotes the number of identical columns between the precoded matrix of the k-th user and the precoded matrix of the i-th user.

Proof: Please refer to Appendix A.

Therefore, we use $\hat{R}_k$ to replace R_k in $\mathbf{P1}$ in the following sections.

A. Data Length Optimization

Theorem 1 implies a tradeoff in the data length of the GSP scheme. On the one hand, increasing τ will increase the payload capability of frames. On the other hand, overladen data will deteriorate the SINR. In the following, we struggle to find a suboptimal data length based on the approximated LB (20).

Theorem 2: Without MI in the GSP scheme, for the MRC detector, the suboptimal data length τ for each user in URLLC transmission is $T-K$ when $R_{\min} \geqslant \frac{\tau}{T\ln 2}$.

Proof: Please refer to Appendix B.

To further illustrate the optimality, we let κ denote the power allocation factor such that $p_k = \kappa P$ and $q_k = (1-\kappa)P$, where P is transmitting total power. As shown in Fig. 2, we present the sum rate versus data length with various power allocations. As expected, the sum rate increases with data length when $\tau \leqslant T-K$ and peaks at $T-K$. Besides, as indicated by the dashed lines in Fig. 2, excessive data causes significant performance degradation since the MI is present.

Using Theorem 2, the variables τ can be substituted by constant $T-K$ in $\mathbf{P1}$ and the SINR in (21) is rewritten as

$$\gamma_k = \frac{M\rho_k\eta_k\alpha_k^2}{\left(\rho_k\alpha_k + \frac{1}{T}\right)\left(\sum_{i\in\mathcal{K}}\eta_i\alpha_i + \frac{T-K}{T}\right)}. \tag{22}$$

Based on the results, in the following, we design an iterative algorithm to find a locally optimal power solution.

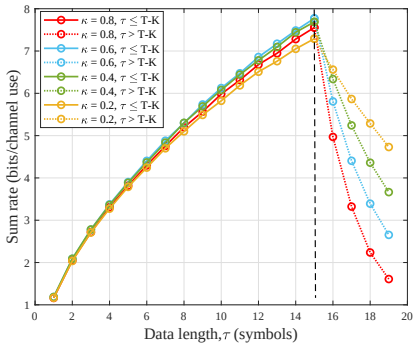


Fig. 2. Sum rate versus data length with different power allocation factor κ for MRC detector, where $K=5$, $T=20$, $\varepsilon=10^{-5}$, and $M=100$.

B. Power Allocation Optimization

An essential aspect of the SP is how to distribute power between the pilot and data symbols. The typical idea is to allocate a fraction of data symbol power to the pilot symbol so that the overall power budget remains the same, as illustrated in III.A. In this paper, we jointly optimize the power allocation between pilot and data symbols of each user while guaranteeing the energy constraint (18c). For ease of tractability, we introduce the following lemma.

Lemma 2: Given $\hat{R}_k \geqslant 0$, function $\hat{R}_k$ is monotonically increasing with respect to (w.r.t.) γ_k.

Proof: The first derivative w.r.t. γ_k is given by $\hat{R}_k' = -\frac{\tau}{T\ln 2}\gamma_k^{-2}\Phi'\left(\gamma_k^{-1}\right)$. $\hat{R}_k$ is always non-negative under the minimum rate constraint. According to [31], the feasible region of γ_k is $\Theta = \left\{x \,\middle|\, 0 < 1/\gamma_k \leqslant g^{-1}(\delta)\right\}$, where $\delta = \frac{Q^{-1}(\varepsilon)}{\sqrt{T}}$. Therefore, Lemma 1 in [31] holds and we have $\Phi'\left(\gamma_k^{-1}\right) \leqslant 0$. Then, we have $\hat{R}_k' \geqslant 0$.

Based on the above results, we reformulate $\mathbf{P1}$ as the following optimization problem.

$$\mathbf{P2}: \max_{\boldsymbol{\rho},\boldsymbol{\eta}} \sum\nolimits_{k\in\mathcal{K}} \mu_k\hat{R}_k \tag{23a}$$

$$\text{s.t. } \gamma_k \geqslant 1\Big/\Phi^{-1}\left(\ln 2\frac{R_{\min}T}{\tau}\right),\ \forall k, \tag{23b}$$

$$T(\rho_k+\eta_k) \leqslant E,\ \forall k, \tag{23c}$$

where (23b) is obtained by applying Lemma 2. However, the objective function (23a) is very complex, which hinders the goal of obtaining a solution. Hence, it is necessary to simplify the objective function (23a). First of all, we introduce auxiliary variables $\boldsymbol{\upsilon} = \{\upsilon_k, \forall k\}$ to transform $\mathbf{P2}$ into the following equivalent problem:

$$\mathbf{P3}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}} \sum\nolimits_{k\in\mathcal{K}} \varpi_k\left[\ln(1+\upsilon_k) - \delta P(\upsilon_k)\right] \tag{24a}$$

$$\text{s.t. } \gamma_k \geqslant \upsilon_k,\ \forall k, \tag{24b}$$

$$\upsilon_k \geqslant 1\Big/\Phi^{-1}\left(\ln 2\frac{R_{\min}T}{\tau}\right),\ (23c),\ \forall k, \tag{24c}$$

where $\varpi_k = \frac{\tau\mu_k}{T\ln 2}$ and $P(\upsilon_k) = \sqrt{V(\upsilon_k)}$. Note that $\mathbf{P2}$ and $\mathbf{P3}$ have the same solutions and optimal value, which can be proved by exploiting the contradiction method. Obviously, the objective function (24a) is still a complicated function. To turn (24a) into a tractable form, we present the following lemmas.

Lemma 3: Given $t \geqslant \frac{\sqrt{17}-3}{4}$, $\forall y \geqslant \frac{\sqrt{17}-3}{4}$, the function $P(y)$ is upper bounded by [31][2]

$$P(y) \leqslant \sigma \mathrm{In}(y) + \theta \triangleq H(y), \tag{25}$$

where σ and θ are defined as, respectively,

$$\sigma = \frac{t}{\sqrt{t^2+2t}} - \frac{t\sqrt{t^2+2t}}{(1+t)^2}, \tag{26}$$

and

$$\theta = \sqrt{1 - \frac{1}{(1+t)^2}} - \sigma \mathrm{In}(t). \tag{27}$$

Additionally, when $y = t$, the upper bound is tight and we have $P(t) = H(t)$ and $P'(t) = H'(t)$.

Proof: Please refer to Appendix C in [31].

Lemma 4: Given $t \geqslant 0$, $\forall y \geqslant 0$, the LB of function $\mathrm{In}(1+y)$ is given by [31]

$$\mathrm{In}(1+y) \geqslant \hat{\sigma}\mathrm{In}(y) + \hat{\theta}, \tag{28}$$

where $\hat{\sigma}$ and $\hat{\theta}$ are defined as, respectively,

$$\hat{\sigma} = \frac{t}{1+t},\ \hat{\theta} = \mathrm{In}(1+t) - \frac{t}{1+t}\mathrm{In}(t). \tag{29}$$

Similarly, the bound is tight at $y = t$.

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

Resort to Lemma 3 and Lemma 4, we can obtain the LB of the objective function (24a), which enables us to solve $\mathbf{P3}$. The main idea is to develop an iterative algorithm where the LB is updated to approximate (24a) in each iteration. Specifically, we denote variables ρ_k, η_k, and υ_k, $\forall k$ in the n-th iteration as $\rho_k^{(n)}$, $\eta_k^{(n)}$, and $\upsilon_k^{(n)}$, $\forall k$. Then, based on (26), (27), and (29), the objective function (24a) is approximated by computing $\sigma_k^{(n)}$, $\hat{\sigma}_k^{(n)}$, $\theta_k^{(n)}$, and $\hat{\theta}_k^{(n)}$ with $t = \upsilon_k^{(n)}$ in the $n+1$-th iteration. Then, substituting $\sigma_k^{(n)}$, $\hat{\sigma}_k^{(n)}$, $\theta_k^{(n)}$, and $\hat{\theta}_k^{(n)}$ into (25) and (28) , we acquire the LB of (24a) in the $n+1$-th iteration as

$$\begin{aligned}&\sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\mathrm{In}(1+\upsilon_k) - \delta P(\upsilon_k)\right] \\ &\geqslant \sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\hat{\sigma}_k^{(n)}\mathrm{In}\upsilon_k + \hat{\theta}_k^{(n)} - \delta\sigma_k^{(n)}\mathrm{In}\upsilon_k - \delta\theta_k^{(n)}\right].\end{aligned} \tag{30}$$

Besides, the LB is tight at $\upsilon_k = \upsilon_k^{(n)}$. Thus, the objective function (24a) is replaced by its LB and $\mathbf{P3}$ is transformed as

$$\mathbf{P4}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}} \sum\nolimits_{k\in\mathcal{K}} \chi_k^{(n)} \mathrm{In}\upsilon_k \tag{31a}$$

$$\text{s.t. } (23c), (24b), (24c), \tag{31b}$$

where $\chi_k^{(n)} = \varpi_k\hat{\sigma}_k^{(n)} - \delta\varpi_k\sigma_k^{(n)}$ and the constant $\varpi_k\hat{\theta}_k^{(n)} - \delta\varpi_k\theta_k^{(n)}$ is omitted. Then, substituting (22) into (24b) and performing simple mathematic transformations, $\mathbf{P4}$ can be further turned into the following GP problem [32]:

$$\mathbf{P5}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}} \prod\nolimits_{k\in\mathcal{K}} \upsilon_k^{\chi_k^{(n)}} \tag{32a}$$

$$\text{s.t. } \left(\rho_k\alpha_k + \frac{1}{T}\right)\left(\sum\nolimits_{i\in\mathcal{K}} \eta_i\alpha_i + 1 - \frac{K}{T}\right)\upsilon_k \tag{32b}$$

$$\leqslant M\rho_k\eta_k\alpha_k^2, \forall k, \tag{32c}$$

$$\upsilon_k \geqslant 1 \Big/ \Phi^{-1}\left(\mathrm{In}2\frac{R_{\min}T}{\tau}\right),\ \forall k, \tag{32d}$$

$$T(\rho_k + \eta_k) \leqslant E,\ \forall k. \tag{32e}$$

In general, to resolve the GP problem, the powerful CVX tool with MOSEK solver is employed, which can convert the GP problem into a convex form via logarithmic change [33]. The detailed steps to resolve $\mathbf{P5}$ is shown in Algorithm 1 [31].

Algorithm 1 Algorithm to solve $\mathbf{P5}$.

1: Initialize iteration number n = 1, error tolerance ξ, and a feasible power allocation $\{\rho_k^{(0)}, \eta_k^{(0)}, \forall k\}$.
2: Compute $\{\upsilon_k^{(0)}, \sigma_k^{(0)}, \hat{\sigma}_k^{(0)}, \chi_k^{(0)}, \forall k\}$ with (26), (29), and $\chi_k^{(0)} = \varpi_k\hat{\sigma}_k^{(0)} - \delta\varpi_k\sigma_k^{(0)}$. Compute the objective function of $\mathbf{P3}$, denoted as $OF^{(0)}$.
3: Given $\{\upsilon_k^{(n-1)}, \sigma_k^{(n-1)}, \hat{\sigma}_k^{(n-1)}, \chi_k^{(n-1)}, \forall k\}$, use the CVX tool to solve $\mathbf{P5}$, obtaining $\{\rho_k^{(n)}, \eta_k^{(n)}, \upsilon_k^{(n)}, \forall k\}$.
4: Update $\{\hat{\sigma}_k^{(n)}, \sigma_k^{(n)}, \chi_k^{(n)}, \forall k\}$.
5: Compute new objective function $OF^{(n)}$, when $\left|OF^{(n)} - OF^{(n-1)}\right|/OF^{(n)} < \xi$, stop iteration. Otherwise, set $n = n+1$, go to step 3.

To trigger the iteration process, we need to provide an initial solution for the algorithm. To this end, we solve the following optimization problem:

$$\mathbf{P6}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\nu} \nu \tag{33a}$$

$$\text{s.t. } \gamma_k \geqslant \nu \Big/ \Phi^{-1}\left(\mathrm{In}2\frac{R_{\min}T}{\tau}\right),\ (23c)\ \forall k, \tag{33b}$$

where ν is an auxiliary variable that decides the availability of the initial feasible solution. Specifically, if $\nu \geqslant 1$, the power solution of $\mathbf{P6}$ can be used to initialize Algorithm 1. Otherwise, the SINR constraint can not be satisfied and we set the objective function to zero in this iteration. The process of solving $\mathbf{P6}$ is omitted here since $\mathbf{P6}$ can also be turned into a GP problem.

C. Convergence and Complexity Analysis

In the subsection, we prove that Algorithm 1 is convergent [31]. Firstly, we validate $OF^{(n)} \leqslant OF^{(n+1)}$. Let $\left\{\upsilon_k^{(n+1)}, \forall k\right\}$ represent the optimal solution in the $n+1$-th iteration. Then, we have

$$\begin{aligned}&\sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\hat{\sigma}_k^{(n)}\mathrm{In}\upsilon_k^{(n+1)} + \hat{\theta}_k^{(n)} - \delta\sigma_k^{(n)}\mathrm{In}\upsilon_k^{(n+1)} - \delta\theta_k^{(n)}\right] \\ &\geqslant \sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\hat{\sigma}_k^{(n)}\mathrm{In}\upsilon_k^{(n)} + \hat{\theta}_k^{(n)} - \delta\sigma_k^{(n)}\mathrm{In}\upsilon_k^{(n)} - \delta\theta_k^{(n)}\right] \\ &= \sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\mathrm{In}\left(1+\upsilon_k^{(n)}\right) - \delta P\left(\upsilon_k^{(n)}\right)\right] = OF^{(n)},\end{aligned} \tag{34}$$

[2]Recalling Lemma 2, we have $\gamma_k \geqslant 1/g^{-1}(\delta)$. In this paper, the inequality $1/g^{-1}(\delta) \geqslant \frac{\sqrt{17}-3}{4}$ is satisfied such that Lemma 3 can be applied.

where the equality holds due to the tightness of (30) at $\upsilon_k = \upsilon_k^{(n)}$. Similarly, for $\upsilon_k = \upsilon_k^{(n+1)}$, we have

$$\begin{aligned} &OF^{(n+1)} \\ &= \sum_{k\in\mathcal{K}} \varpi_k \left[\ln\left(1+\upsilon_k^{(n+1)}\right) - \delta P\left(\upsilon_k^{(n+1)}\right)\right] \\ &= \sum_{k\in\mathcal{K}} \varpi_k \left[\hat{\sigma}_k^{(n)} \ln \upsilon_k^{(n+1)} + \hat{\theta}_k^{(n)} - \delta\sigma_k^{(n)} \ln \upsilon_k^{(n+1)} - \delta\theta_k^{(n)}\right]. \end{aligned} \tag{35}$$

Combining (34) and (35), we readily obtain $OF^{(n)} \leqslant OF^{(n+1)}$. Note that the objective function can not be infinite owing to the energy constraint, which suggests that Algorithm 1 is convergent.

According to [34], the computational complexity of Algorithm 1 is $\mathcal{O}\left(L_1 \times \max\left\{(3K)^3, L_2\right\}\right)$ with $3K$ variables and $3K$ constraints in $\mathbf{P5}$, where L_1 is the number of iterations and L_2 is the computational complexity of calculating the first-order and second-order derivatives of the objective and constraint functions of $\mathbf{P5}$.

IV. Joint optimization for ZF

When the ZF detector is considered, $\mathbf{P1}$ becomes more intractable due to the complicated SINR expression. To tackle the issue, we first approximate the expression by the SCA method such that the derivations in the MRC case can be applied, and then an iterative algorithm is developed.

Theorem 3: For the ergodic achievable rate of the k-th user in URLLC transmission with the GSP scheme and ZF detector, its LB can be expressed as (20) with

$$\gamma_k = \frac{(M-K)\,\rho_k\eta_k\alpha_k^2}{\frac{1}{T}\eta_k\alpha_k + \left(\rho_k\alpha_k + \frac{1}{T}\right)\left(\frac{1}{\tau}\sum_{i\in\mathcal{K}\setminus k}\frac{\eta_i\alpha_i r_{ik}}{\rho_i T\alpha_i+1} + \frac{\tau}{T}\right)}. \tag{36}$$

Proof: Please refer to Appendix C.

Remark 1: In fact, the closed-form SINR expression of ZF detection in the SP scheme is hard to obtain. For the RP scheme, the expectation $\mathbb{E}\{[(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}]_{k,k}\}$ can be calculated by using the random matrix theorem with the Gaussian nature of estimated channels when we derive the closed-form SINR with ZF receiver [35]. However, this conclusion can not be applied in the SP scheme since the estimated channel is not Gaussian [22]. Unlike the SP scheme, the GSP scheme removes the MI such that the estimated channel is still Gaussian and (36) can be derived.

A. Data Length Optimization

Similarly, we show the suboptimal data length in a closed form, which is summarized as follows.

Theorem 4: Without MI in the GSP scheme, for the ZF detector, the suboptimal data length τ for each user in URLLC transmission is $T-K$ when $R_{\min} \geqslant \frac{\tau}{T\ln 2}$.

Proof: Please refer to Appendix B.

As illustrated in Fig. 3, to avoid MI and maximize the sum rate in the GSP scheme, the data length should be set to $T-K$, which validates the accuracy of Theorem 4.

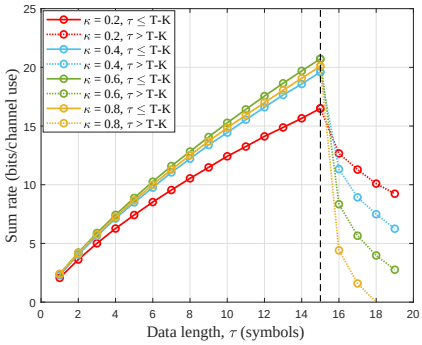


Fig. 3. Sum rate versus data length with different power allocation factor κ for ZF detector, where $K=5$, $T=20$, $\varepsilon=10^{-5}$, and $M=100$.

Theorem 4 implies that $\mathbf{\Phi}_{ik}$ is a full rank matrix, and then the SINR in (36) is rewritten as

$$\gamma_k = \frac{(M-K)\,\rho_k\eta_k\alpha_k^2}{\left(\rho_k\alpha_k + \frac{1}{T}\right)\left(\sum_{i\in\mathcal{K}}\frac{\eta_i\alpha_i}{\rho_i T\alpha_i+1} + \frac{\tau}{T}\right)}. \tag{37}$$

B. Power Allocation Optimization

In this part, we manage to convert $\mathbf{P2}$ for the ZF detector into a GP problem. Unfortunately, the SINR constraint (23b) with (37) can not be written as a standard GP constraint because the denominator of (37) is not a posynomial function. To overcome this obstacle, we approximate the denominator by means of the following lemma [36].

Lemma 5: Denote a posynomial as $T(\mathbf{x}) = \sum_i t_i(\mathbf{x})$, in the l-th iteration, we have

$$T(\mathbf{x}) \geqslant \hat{T}(\mathbf{x}) = \prod_i \left(\frac{t_i(\mathbf{x})}{\vartheta_i}\right)^{\vartheta_i}, \tag{38}$$

where $\vartheta_i = t_i(\mathbf{x}_{l-1})/T(\mathbf{x}_{l-1})$ and $\mathbf{x}_{l-1}$ is the optimal solution obtained in the $(l-1)$-th iteration. Additionally, the bound is tight at $\mathbf{x}=\mathbf{x}_{l-1}$.

Proof: Please refer to Lemma 1 in [36].

From (38), it can be seen that the denominator of the term $\Upsilon_i = \frac{\eta_i\alpha_i}{\rho_i T\alpha_i+1}$ is a posynomial. In the $l+1$-th iteration, we apply Lemma 5 to the denominator of $\Upsilon_i, \forall i\in\mathcal{K}$, constructing K corresponding monomial functions as their approximations, which is given by

$$\rho_i T\alpha_i + 1 \geqslant \left(\frac{\rho_i T\alpha_i}{\vartheta_{i,1}^{(l)}}\right)^{\vartheta_{i,1}^{(l)}} \left(\frac{1}{\vartheta_{i,2}^{(l)}}\right)^{\vartheta_{i,2}^{(l)}} = F_l(\rho_i),\ \forall i\in\mathcal{K}, \tag{39}$$

with

$$\vartheta_{i,1}^{(l)} = \frac{\rho_i^{(l)} T\alpha_i}{\rho_i^{(l)} T\alpha_i + 1},\ \vartheta_{i,2}^{(l)} = \frac{1}{\rho_i^{(l)} T\alpha_i + 1},\ \forall i\in\mathcal{K}, \tag{40}$$

where $\rho_i^{(l)}$ is the pilot power solution of the i-th user in the l-th iteration. Using the approximation, the denominator of (37) becomes a monomial, and then the SINR constraint (23b) in

the $(l+1)$-th iteration is rewritten as[3]

$$\left(\rho_k\alpha_k+\frac{1}{T}\right)\left(\sum_{i\in\mathcal{K}}\frac{\eta_i\beta_i}{F_l\left(\rho_i\right)}+\frac{\tau}{T}\right)\upsilon_k\leqslant\left(M-K\right)\rho_k\eta_k\alpha_k^2. \tag{41}$$

Next, we perform the same transformation for $\mathbf{P2}$ as in the case of MRC. Thus, the GP format of $\mathbf{P2}$ in the ZF case is given by

$$\mathbf{P7}:\max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}}\prod\nolimits_{k\in\mathcal{K}}\upsilon_k^{\chi_k^{(l)}} \tag{42a}$$

$$\text{s.t. }(23c),(41), \tag{42b}$$

$$\upsilon_k\geqslant 1\Big/\Phi^{-1}\left(\text{In2}\frac{R_{\min}T}{\tau}\right),\ \forall k. \tag{42c}$$

where $\chi_k^{(l)}=\varpi_k\hat{\sigma}_k^{(l)}-\delta\varpi_k\sigma_k^{(l)}$. Similarly, an iterative algorithm is developed to resolve $\mathbf{P7}$, as presented in Algorithm 2 [31].

Algorithm 2 Algorithm to solve $\mathbf{P7}$.

1: Initialize iteration number l = 1, error tolerance ξ, and a feasible power allocation $\{\rho_k^{(0)},\eta_k^{(0)},\forall k\}$.
2: Compute $\{\upsilon_k^{(0)},\sigma_k^{(0)},\hat{\sigma}_k^{(0)},\chi_k^{(0)},\vartheta_{i,1}^{(0)},\vartheta_{i,2}^{(0)},\forall k\}$ with (26), (29), (40), and $\chi_k^{(0)}=\varpi_k\hat{\sigma}_k^{(0)}-\delta\varpi_k\sigma_k^{(0)}$. Compute the objective function of $\mathbf{P3}$, denoted as $OF^{(0)}$.
3: Given $\{\upsilon_k^{(l-1)},\sigma_k^{(l-1)},\hat{\sigma}_k^{(l-1)},\chi_k^{(l-1)},\vartheta_{i,1}^{(l-1)},\vartheta_{i,2}^{(l-1)},\forall k\}$, use the CVX tool to solve $\mathbf{P7}$, obtaining $\{\rho_k^{(l)},\eta_k^{(l)},\upsilon_k^{(l)},\forall k\}$.
4: Update $\{\hat{\sigma}_k^{(l)},\sigma_k^{(l)},\chi_k^{(l)},\vartheta_{i,1}^{(l)},\vartheta_{i,2}^{(l)},\forall k\}$.
5: Compute new objective function $OF^{(n)}$, when $\left|OF^{(n)}-OF^{(n-1)}\right|/OF^{(n)}<\xi$, stop iteration. Otherwise, set $n=n+1$, go to step 3.

C. Convergence and Complexity Analysis

In each iteration, Algorithm 2 has different SINR constraints. Therefore, we only need to prove that the optimal solution from the last iteration is feasible for the next iteration. Then, the method to analyze the convergence of Algorithm 1 can be utilized for Algorithm 2. We only need to analyze the constraint (41) since it is different in each iteration.

Let $\left\{\rho_k^{(l)},\eta_k^{(l)},\upsilon_k^{(l)},\forall k\right\}$ be the optimal solution in the l-th iteration. According to (41), we have

$$\left(\rho_k^{(l)}\alpha_k+\frac{1}{T}\right)\left(\sum\nolimits_{i\in\mathcal{K}}\frac{\eta_i^{(l)}\alpha_i}{F_{l-1}(\rho_i^{(l)})}+\frac{\tau}{T}\right)\upsilon_k^{(l)} \leqslant\left(M-K\right)\rho_k^{(l)}\eta_k^{(l)}\alpha_k^2. \tag{43}$$

By using (40), we have

$$\rho_i^{(l)}T\alpha_i+1\geqslant\left(\frac{\rho_i^{(l)}T\alpha_i}{\vartheta_{i,1}^{(l-1)}}\right)^{\vartheta_{i,1}^{(l-1)}}\left(\frac{1}{\vartheta_{i,2}^{(l-1)}}\right)^{\vartheta_{i,2}^{(l-1)}}=F_{l-1}(\rho_i^{(l)}). \tag{44}$$

Applying the tightness of Lemma 5, we have

$$\rho_i^{(l)}T\alpha_i+1=\left(\frac{\rho_i^{(l)}T\alpha_i}{\vartheta_{i,1}^{(l)}}\right)^{\vartheta_{i,1}^{(l)}}\left(\frac{1}{\vartheta_{i,2}^{(l)}}\right)^{\vartheta_{i,2}^{(l)}}=F_l\left(\rho_i^{(l)}\right). \tag{45}$$

In the end, combining (43)-(45), we have

$$\left(\rho_k^{(l)}\alpha_k+\frac{1}{T}\right)\left(\sum\nolimits_{i\in\mathcal{K}}\frac{\eta_i^{(l)}\alpha_i}{F_l(\rho_i^{(l)})}+\frac{\tau}{T}\right)\upsilon_k^{(l)} \leqslant\left(M-K\right)\rho_k^{(l)}\eta_k^{(l)}\alpha_k^2, \tag{46}$$

which indicates that $\left\{\rho_k^{(l)},\eta_k^{(l)},\upsilon_k^{(l)},\forall k\right\}$ is a feasible solution in the $(l+1)$ iteration.

Similarly, $\mathbf{P7}$ has $3K$ variables and $3K$ constraints and thus the computational complexity of Algorithm 2 is on the order of $\mathcal{O}\left(L_3\times\max\left\{(3K)^3,L_4\right\}\right)$ [34], where L_3 is the number of iterations and L_4 is the computational complexity of calculating the first-order and second-order derivatives of the objective and constraint functions of $\mathbf{P7}$.

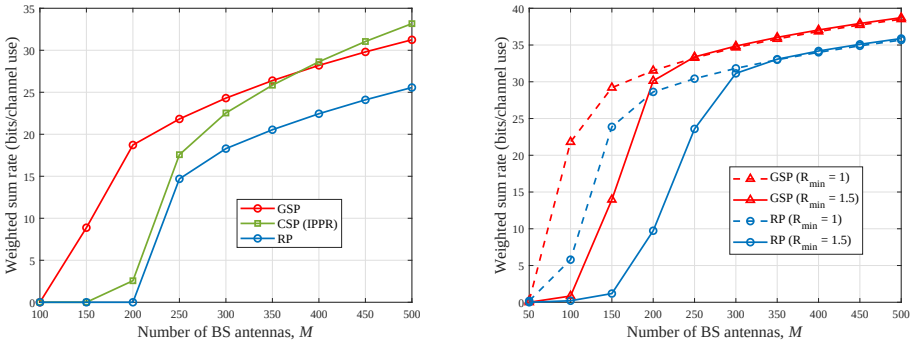
V. Simulation Results

In this section, we analyze the performance of the GSP scheme in short packet transmission from the perspective of simulation experiments. We consider a rectangle simulation area, where a BS is deployed in the centre and the locations of users follow a uniform distribution. Without loss of generality, the pathloss model is chosen as $\text{PL}_k=35.3+37.6\log_{10}d_k\,(\text{dB})$ [37], where d_k is the distance from the k-th user to the BS. The noise power spectral density and the decoding error rate ε are set as -174 dBm/Hz and 10^{-9}, respectively. Note that the energy limit E is normalized in the following, i.e., $E=\frac{E_0}{\sigma_N}$, where σ_N is noise power. For a given channel uses (latency) T, we compare the weighted sum rate of various schemes in the following. Note that the benchmark schemes RP and SP are optimized in the same setup [21], which means that the pilot length of the RP scheme and the power allocation for the SP and RP schemes are optimized under the same blocklength T and energy limit E as the GSP scheme[4]. For the SP scheme, we consider the practical situation of imperfect pilot interference removal (IPPR) [20]. We average 100 trails to obtain the Monte-Carlo results, in which the locations of users are randomly generated in each trail.

Fig. 4 illustrates the impact of the number of BS antennas on the weighted sum rate. It can be seen that the system performance is improved with the increase in the number of antennas at the BS. In the MRC case, the GSP scheme displays superior performance to the traditional RP scheme when antenna number is below 400. This is because larger channel uses in the GSP scheme reduces the rate loss caused by finite blocklength. Compared with the SP scheme, the

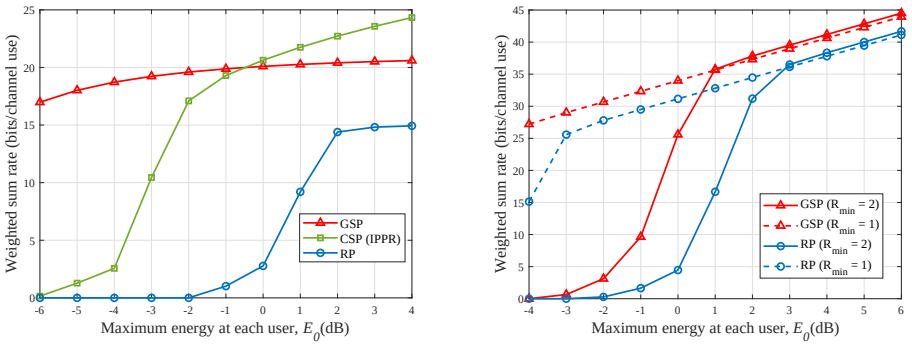
[3]In [31], the product of K posynomials is approximated by using Theorem 3, which easily generate a overflowing result in simulation especially when the number of users K is large, due to the best local monomial approximation is a continued product of power functions. Here, we approximate each posynomial independently such that the simulation results can be obtained even when K is large.

[4]The optimization problems for the RP and SP schemes are similar to [21] except for the SINRs and minimum rate constraints, and the optimal pilot length of the RP scheme is equal to the number of users, which can be proved by using Lemma 1. In the case of ZF, we only compare the GSP and RP schemes with different rate requirements since the closed-form SINR expression of SP is hard to obtain. The performance comparison between the GSP and SP schemes will be further studied in future work.



(a) MRC case (b) ZF case

Fig. 4. Impact of the number of BS antennas on the weighted sum rate: (a) MRC case: $K = 20$, $T = 40$, $E_0 = -4$ dB, and $R_{\min} = \frac{T-K}{T\ln 2}$; (b) ZF case: $K = 20$, $T = 40$, and $E_0 = -5$ dB.



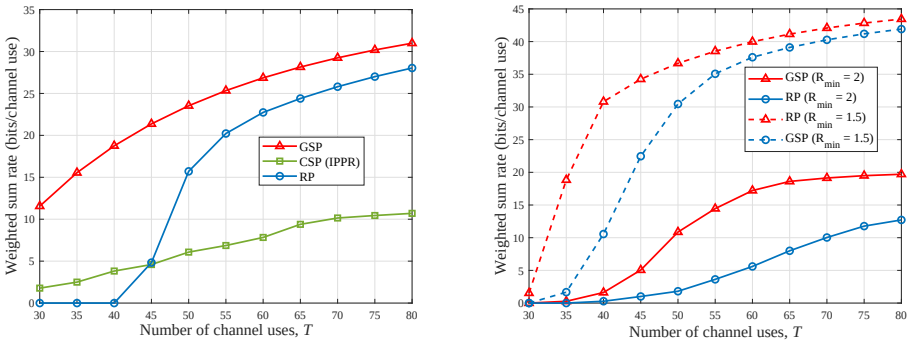
(a) MRC case (b) ZF case

Fig. 5. Impact of the maximum energy limit of each user on the weighted sum rate: (a) MRC case: $K = 20$, $T = 40$, $M = 200$, and $R_{\min} = \frac{T-K}{T\ln 2}$; (b) ZF case: $K = 20$, $T = 40$, and $M = 100$.

GSP scheme achieves higher reliability by removing the MI between the pilot and data. However, the SP scheme outperforms the GSP scheme when $M = 400$, which is due to the number of antennas starting to dominate the performance of the SP scheme. This brings an enormous expense and is unrealistic in actual deployment [38]. Besides, in the ZF case, we observe a significant performance gap between the GSP and RP schemes at medium antenna numbers with a more strict rate requirement, which indicates that the GSP scheme has the potential to carry out the target of low latency, high rate, and ultra-reliability. Besides, under a more relaxed rate requirement, the gap becomes smaller but is still evident.

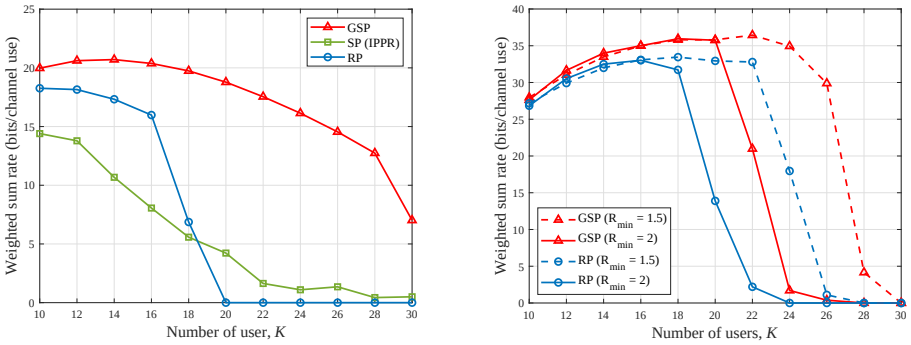
Fig. 5 demonstrates the impact of the maximum energy limit at each user on the weighted sum rate. As shown in the figures, increasing the energy budget of users can provide more performance gain in the low energy region, which is owing to the SINR being improved. For the MRC case, the GSP scheme has the best performance even in a shortage of energy, while that of the RP scheme is the worst. However, in the high energy regime, we observe that the SP scheme outperforms the GSP scheme when $E_0 = 0$ dB. The reason for which is that the joint power allocation decreases the MI in the SP scheme as energy increases, and more superposed data symbols brings high transmission efficiency. In fact, the energy reserve of users, such as sensors, is very limited in practical scenarios [39]. Hence, the GSP scheme is more suitable for uplink URLLC transmission than the SP scheme in reality. For the ZF case, we observe that the rate can be improved by more than 10 bits/channel use under different fairness requirements, which states the superiority of the GSP scheme again.

Fig. 6 shows the impact of the number of channel uses on the weighted sum rate. As expected, the rate increases with the growth of channel uses number because of the availability of more resources. The dependence on energy leads to an obvious performance gap between the SP scheme and the other two schemes. It should be noted that the GSP scheme is superior to the SP and RP schemes for the MRC and ZF cases, especially in the small channel uses region, which justifies that the GSP scheme is appropriate for finite blocklength transmission. Moreover, it can be seen that the gap between the GSP and RP schemes gradually becomes small with the increase in channel uses. This is because the rate loss caused by finite blocklength transmission is reduced in the RP scheme. Moreover, the impact of pilot overhead reduces in the RP scheme while energy budget becomes tight in the SP scheme with the increase of channel uses number. Hence, we can find the RP scheme outperforms the SP scheme at $T = 45$. Additionally, we find that increasing the minimum rate requirement leads to a worse sum rate performance in the ZF case. This is because the increase in channel uses and minimum rate make energy become the dominant factor in system performance.



(a) MRC case (b) ZF case

Fig. 6. Impact of the number of channel uses on the weighted sum rate: (a) MRC case: $K = 20$, $E_0 = -4$ dB, $M = 200$, and $R_{\min} = \frac{T-K}{T\ln 2}$; (b) ZF case: $K = 20$, $E_0 = -2$ dB, and $M = 100$.



(a) MRC case (b) ZF case

Fig. 7. Impact of the number of users on the weighted sum rate: (a) MRC case: $T = 40$, $E_0 = -4$ dB, $M = 200$, and $R_{\min} = \frac{T-K}{T\ln 2}$; (b) ZF case: $T = 40$, $E_0 = 1$ dB, and $M = 100$.

Fig. 7 depicts the impact of the number of users on the weighted sum rate. For the two cases, owing to multi-user diversity, it is clear to see that the GSP scheme increases with user number in a small K. Then, for the MRC case, the rate decreases with K since the multi-user interference becomes severe. However, the decrease of the GSP scheme is slight while the others fall sharply, which reveals strong robustness of the GSP scheme. Moreover, it can be observed that the performance of the RP scheme is worse than that of the SP scheme when K larger than 20 due to substantial pilot overhead, which indicates the RP scheme is not suitable for massive URLLC transmission. Similarly, the GSP scheme also shows its robustness in the ZF case. In other words, the GSP scheme can support more users at a higher rate than the RP scheme.

VI. Conclusion

In this paper, we investigated the GSP scheme supporting URLLC in mMIMO systems under the MRC and ZF detection. The key idea behind the GSP scheme is to eliminate the MI in the SP scheme by shrinking the data length and precoding for the data symbols. We formulate a mixed-integer nonlinear programming problem where the data length, pilot and data power allocation are jointly optimized to maximize the weighted sum rate. The suboptimal data lengths and achievable rate LBs for the MRC and ZF cases are derived in a closed form, respectively, which shows that the suboptimal data length of the GSP scheme is equal to $T - K$ no matter which detector is utilized. Based on the results, the optimization problems can be transformed into a sequence of GP problems, and we then develop two iterative algorithms to obtain locally optimal solutions. Simulation results demonstrate that the GSP scheme is superior to the conventional RP and SP schemes in the MRC case in terms of supporting massive connectivity with low resource configuration. In the ZF case, the GSP scheme still outperforms the RP scheme and the comparison to the SP scheme is left in future work.

Appendix A
Proof of Theorem 1

The effective SINR for the MRC detector is given by

$$\gamma_k = \frac{\eta_k T^2 \left|\beta_k \alpha_k^{-1} \mathbb{E}\left\{\mathbf{h}_k^H \mathbf{h}_k\right\}\right|^2}{\eta_k T^2 \mathrm{Var}\left(\beta_k \alpha_k^{-1} \mathbf{h}_k^H \mathbf{h}_k\right) + \mathbb{E}\left\{\left\|\boldsymbol{\omega}_k^H - \mathbb{E}\left\{\boldsymbol{\omega}_k^H\right\}\right\|^2\right\}}. \tag{47}$$

To calculate the expectations and variances in (47), we have

$$\left|\beta_k \alpha_k^{-1} \mathbb{E}\left\{\mathbf{h}_k^H \mathbf{h}_k\right\}\right|^2 = M^2 \alpha_k^2, \tag{48}$$

$$\mathbb{E}\left\{\left|\beta_k \alpha_k^{-1} \mathbf{h}_k^H \mathbf{h}_k\right|^2\right\} = \beta_k^2 \alpha_k^{-2} \mathbb{E}\left\{\left\|\mathbf{h}_k\right\|^4\right\} = M\left(M+1\right)\alpha_k^2. \tag{49}$$

Using the results of (48) and (49), we have

$$\begin{aligned}&\mathrm{Var}\left(\beta_k\alpha_k^{-1}\mathbf{h}_k^H\mathbf{h}_k\right)\\&=\mathbb{E}\left\{\left|\beta_k\alpha_k^{-1}\mathbf{h}_k^H\mathbf{h}_k\right|^2\right\}-\left|\mathbb{E}\left\{\beta_k\alpha_k^{-1}\mathbf{h}_k^H\mathbf{h}_k\right\}\right|^2=M\alpha_k^2.\end{aligned}\tag{50}$$

For the sake of derivation, we calculate the variance of effective noise by decomposing it, i.e.,

$$\begin{aligned}&\mathbb{E}\left\{\left\|\boldsymbol{\omega}_k^H-\mathbb{E}\left\{\boldsymbol{\omega}_k^H\right\}\right\|^2\right\}\\&=\mathbb{E}\left\{\left\|\boldsymbol{\omega}_k^H\right\|^2\right\}-\left\|\mathbb{E}\left\{\boldsymbol{\omega}_k^H\right\}\right\|^2\\&=\sum_{i=1}^{3}\mathbb{E}\left\{\left\|\mathbf{u}_{ik}\right\|^2\right\}-\left\|\mathbb{E}\left\{\mathbf{u}_{ik}\right\}\right\|^2\\&+2\mathbf{R_e}\left\{\sum_{i=1}^{3}\sum_{j=i+1}^{3}\mathbb{E}\left\{\mathbf{u}_{ik}\mathbf{u}_{jk}^H\right\}-\mathbb{E}\left\{\mathbf{u}_{ik}\right\}\mathbb{E}\left\{\mathbf{u}_{jk}^H\right\}\right\},\end{aligned}\tag{51}$$

where $\mathbf{u}_{1k}=\sqrt{\eta_k}T\bar{\mathbf{h}}_k^H\mathbf{h}_k\mathbf{s}_k^H$, $\mathbf{u}_{2k}=\sqrt{\eta_i}T\sum_{i\in\mathcal{K}\setminus k}\hat{\mathbf{h}}_k^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}$, and $\mathbf{u}_{3k}=\hat{\mathbf{h}}_k^H\mathbf{N}\mathbf{W}_k$. In the following, we show the derivation of each expectation in the expansion (51). The main idea is to isolate the correlated and uncorrelated parts and then compute them respectively.

$$\begin{aligned}\mathbb{E}\left\{\left\|\mathbf{u}_{1k}\right\|^2\right\}&=\eta_k\mathbb{E}\left\{\left\|\lambda_k T\mathbf{n}_k^H\mathbf{h}_k\mathbf{s}_k^H\right\|^2\right\}\\&=\eta_k\lambda_k^2T\mathbb{E}\left\{\left\|\boldsymbol{\varphi}_k^H\mathbf{N}^H\mathbf{h}_k\mathbf{s}_k^H\right\|^2\right\}\\&=\eta_k\lambda_k^2T\mathbb{E}\left\{\boldsymbol{\varphi}_k^H\mathbf{N}^H\mathbf{h}_k\mathbf{s}_k^H\mathbf{s}_k\mathbf{h}_k^H\mathbf{N}\boldsymbol{\varphi}_k\right\}\\&=M\lambda_k^2T^2\eta_k\alpha_k\left(\rho_kT\alpha_k+1-\rho_kT\alpha_k\right)\\&=M\eta_k\alpha_k\beta_kT^2-M\eta_k\beta_k^2T^2\\&=MT^2\eta_k\beta_k\left(\alpha_k-\beta_k\right).\end{aligned}\tag{52}$$

After expansion, $\mathbb{E}\left\{\left\|\mathbf{u}_{2k}\right\|^2\right\}$ can be expressed as

$$\begin{aligned}\mathbb{E}\left\{\left\|\mathbf{u}_{2k}\right\|^2\right\}=&\underbrace{\mathbb{E}\left\{\left\|\sqrt{\eta_i}T\sum_{i\in\mathcal{K}\setminus k}\lambda_k\sqrt{\rho_kT}\mathbf{h}_k^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\right\|^2\right\}}_{A}\\&+\underbrace{\mathbb{E}\left\{\left\|\sqrt{\eta_i}T\sum_{i\in\mathcal{K}\setminus k}\lambda_k\mathbf{n}_k^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\right\|^2\right\}}_{B}.\end{aligned}\tag{53}$$

The derivations of A and B are given by

$$\begin{aligned}A&=\rho_k\lambda_k^2T^3\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\mathbb{E}\left\{\left\|\mathbf{h}_k^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\right\|^2\right\}\\&=\rho_k\lambda_k^2T^3\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\mathbb{E}\left\{\mathbf{h}_k^H\mathbf{h}_i\mathbf{h}_i^H\mathbf{h}_k\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\boldsymbol{\Phi}_{[ik]}^H\mathbf{s}_i\right\}\\&=M\frac{T^3}{\tau}\rho_k\lambda_k^2\alpha_k\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\alpha_ir_{ik},\end{aligned}\tag{54}$$

$$\begin{aligned}B&=\lambda_k^2\eta_iT\sum\nolimits_{i\in\mathcal{K}\setminus k}\mathbb{E}\left\{\left\|\boldsymbol{\varphi}_k^H\mathbf{N}^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\right\|^2\right\}\\&=\lambda_k^2\eta_iT\sum\nolimits_{i\in\mathcal{K}\setminus k}\mathbb{E}\left\{\boldsymbol{\varphi}_k^H\mathbf{N}^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\boldsymbol{\Phi}_{[ik]}^H\mathbf{s}_i\mathbf{h}_i^H\mathbf{N}\boldsymbol{\varphi}_k\right\}\\&=M\lambda_k^2\frac{T^2}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\alpha_ir_{ik}.\end{aligned}\tag{55}$$

Combining (53)-(55), we have

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{u}_{2k}\right\|^2\right\}\\&=M\frac{T^3}{\tau}\rho_k\lambda_k^2\beta_k\sum_{i\in\mathcal{K}\setminus k}\eta_i\alpha_ir_{ik}+M\lambda_k^2\frac{T^2}{\tau}\sum_{i\in\mathcal{K}\setminus k}\eta_i\alpha_ir_{ik}\\&=M\lambda_k^2\frac{T^2}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\alpha_ir_{ik}\left(\rho_k\tau_c\alpha_k+1\right)\\&=M\beta_k\frac{T^2}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\alpha_ir_{ik}\end{aligned}\tag{56}$$

Similarly, $\mathbb{E}\left\{\left\|\mathbf{u}_{3k}\right\|^2\right\}$ is expanded as

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{u}_{3k}\right\|^2\right\}\\&=\underbrace{\mathbb{E}\left\{\left\|\lambda_k\sqrt{\rho_kT}\mathbf{h}_k^H\mathbf{N}\mathbf{W}_k\right\|^2\right\}}_{C}+\frac{\lambda_k^2}{T}\underbrace{\mathbb{E}\left\{\left\|\boldsymbol{\varphi}_k^H\mathbf{N}^H\mathbf{N}\mathbf{W}_k\right\|^2\right\}}_{D},\end{aligned}\tag{57}$$

where the first term in (57) is simplified as

$$\begin{aligned}C&=\lambda_k^2\rho_kTtr\left\{\mathbf{W}_k^H\mathbb{E}\left\{\mathbf{N}^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{N}\right\}\mathbf{W}_k\right\}\\&=\lambda_k^2\rho_kT\alpha_ktr\left\{\mathbf{W}_k^H\mathbb{E}\left\{\mathbf{N}^H\mathbf{N}\right\}\mathbf{W}_k\right\}\\&=M\lambda_k^2\rho_kT\alpha_ktr\left\{\mathbf{W}_k^H\mathbf{W}_k\right\}\\&=M\lambda_k^2T^2\tau\rho_k\alpha_k.\end{aligned}\tag{58}$$

For ease of derivation, we rewrite the noise matrix $\mathbf{N}$ in the form of a block matrix, i.e., $\mathbf{N}_*=[\bar{\mathbf{n}}_1,\ldots,\bar{\mathbf{n}}_T]$, where each entry is a column vector of $\mathbf{N}$. Then, we have $\left[\mathbf{N}_*^H\mathbf{N}_*\right]_{i,j}=\bar{\mathbf{n}}_i^H\bar{\mathbf{n}}_j$ and the second term in (57) is further computed as

$$\begin{aligned}D&=tr\left\{\mathbb{E}\left\{\mathbf{N}^H\mathbf{N}\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\mathbf{N}^H\mathbf{N}\mathbf{W}_k\mathbf{W}_k^H\right\}\right\}\\&=\sum_{i,j,l,r}\mathbb{E}\left\{\left[\mathbf{N}_*^H\mathbf{N}_*\right]_{i,j}\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{j,l}\left[\mathbf{N}_*^H\mathbf{N}_*\right]_{l,r}\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{r,i}\right\}\\&=\sum_{i,j,l,r}\mathbb{E}\left\{\bar{\mathbf{n}}_i^H\bar{\mathbf{n}}_j\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{j,l}\bar{\mathbf{n}}_l^H\bar{\mathbf{n}}_r\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{r,i}\right\}\\&=\sum_{i=j,l=r,i\neq l}\mathbb{E}\left\{\bar{\mathbf{n}}_i^H\bar{\mathbf{n}}_i\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{i,l}\bar{\mathbf{n}}_l^H\bar{\mathbf{n}}_l\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{l,i}\right\}\\&\quad+\sum_{l=j,r=i,i\neq l}\mathbb{E}\left\{\bar{\mathbf{n}}_i^H\bar{\mathbf{n}}_j\bar{\mathbf{n}}_j^H\bar{\mathbf{n}}_i\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{j,j}\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{i,i}\right\}\\&\quad+\sum_{i=j=l=r}\mathbb{E}\left\{\bar{\mathbf{n}}_i^H\bar{\mathbf{n}}_i\bar{\mathbf{n}}_i^H\bar{\mathbf{n}}_i\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{i,i}\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{i,i}\right\}\\&=M^2\sum_{i,l,i\neq l}\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{i,l}\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{l,i}\\&\quad+M\sum_{i,j,i\neq j}\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{j,j}\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{i,i}\\&\quad+\left(M^2+M\right)\sum\left[\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right]_{i,i}\left[\mathbf{W}_k\mathbf{W}_k^H\right]_{i,i}\\&=M^2\left(tr\left\{\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\mathbf{W}_k\mathbf{W}_k^H\right\}-tr\left\{\mathbf{W}_k\mathbf{W}_k^H\boldsymbol{\Lambda}\right\}\right)\\&\quad+M\left(tr\left\{\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H\right\}tr\left\{\mathbf{W}_k\mathbf{W}_k^H\right\}-tr\left\{\mathbf{W}_k\mathbf{W}_k^H\boldsymbol{\Lambda}\right\}\right)\\&\quad+\left(M^2+M\right)tr\left\{\mathbf{W}_k\mathbf{W}_k^H\boldsymbol{\Lambda}\right\}\\&=MT^2\tau,\end{aligned}\tag{59}$$

where $\mathbf{\Lambda} = \mathcal{D}\left(\boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^H\right)$. Then, combining (57)-(59), we have

$$\mathbb{E}\left\{\left\|\mathbf{u}_{3k}\right\|^2\right\} = M\beta_k T\tau. \tag{60}$$

The calculative procedure of the remaining terms in (51) is omitted since they are equal to zero. Finally, we can acquire (21) by substituting (48), (50), (52), (56), and (60) into (47).

Appendix B
Proof of Theorem 2

Let $x = \frac{\tau}{T} \in \left[\frac{1}{T}, \frac{T-K}{T}\right]$, $\delta = Q^{-1}(\varepsilon)\Big/\sqrt{T}$. Next, we define function $f(x) = x\log_2\left(1+\frac{a}{bx+c}\right) - \frac{\delta}{\ln 2}x\sqrt{1-\frac{1}{\left(1+\frac{a}{bx+c}\right)^2}}$, and provide the following Lemma.

Lemma 1: When $f(x) \geqslant \frac{\tau}{T\ln 2}$, for positive a, b, and c, $f(x)$ is a strictly monotonic increasing function w.r.t. x.

Proof: According to the known conditions $f(x) \geqslant \frac{x}{\ln 2}$, we have $\ln\left(1+\frac{a}{bx+c}\right) - \delta\sqrt{1-\frac{1}{\left(1+\frac{a}{bx+c}\right)^2}} \geqslant 1$. Then, taking the first derivative w.r.t. x, we obtain

$$f'(x) = \frac{1}{\ln 2}\left\{\ln\left(1+\frac{a}{bx+c}\right) - \frac{abx}{(bx+c)(bx+c+a)} - \delta\sqrt{1-\frac{1}{\left(1+\frac{a}{bx+c}\right)^2}} + \delta\frac{abx}{\sqrt{\left(\frac{a}{bx+c}\right)^2+2\left(\frac{a}{bx+c}\right)}(bx+c+a)^2}\right\} > 0. \tag{61}$$

The inequality in (61) is due to the fact that $\frac{bx}{bx+c}\cdot\frac{a}{bx+c+a} < 1$. For the MRC case, we apply Lemma 1 with $a = M\rho_k\eta_k\alpha_k^2$, $b = \rho_k\alpha_k + \frac{1}{T}$, and $c = \left(\rho_k\alpha_k+\frac{1}{T}\right)\eta_k\alpha_k + \left(\rho_k\alpha_k+\frac{1}{T}\right)\frac{1}{\tau}\sum_{i\in\mathcal{K}\setminus k}\eta_i\alpha_i r_{ik}$. To maximize the achievable rate of users, i.e. $f(x)$, the value of x should be taken $\frac{T-K}{T}$, which proves Theorem 2. Similarly, for the ZF case, we let $a = M\rho_k\eta_k\alpha_k^2$, $b = \rho_k\alpha_k + \frac{1}{T}$, $c = \left(\rho_k\alpha_k+\frac{1}{T}\right)\left(\frac{1}{\tau}\sum_{i\in\mathcal{K}\setminus k}\frac{\eta_i\alpha_i r_{ik}}{\rho_i T\alpha_i+1}\right) + \frac{1}{T}\eta_k\alpha_k$, which proves Theorem 4.

Appendix C
Proof of Theorem 3

The effective SINR for the ZF detector is rewritten as

$$\gamma_k = \frac{\eta_k T^2\left|\mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbf{h}_k\right\}\right|^2}{\eta_k T^2\mathrm{Var}\left(\boldsymbol{\psi}_k^H\mathbf{h}_k\right) + \mathbb{E}\left\{\left\|\boldsymbol{\omega}_k{}^H - \mathbb{E}\left\{\boldsymbol{\omega}_k{}^H\right\}\right\|^2\right\}}. \tag{62}$$

Under the ZF detector, we have $\boldsymbol{\psi}_k^H\hat{\mathbf{h}}_k = 1$ and $\boldsymbol{\psi}_k^H\hat{\mathbf{h}}_i = 0$ if $i \neq k$. Using the independence between $\hat{\mathbf{h}}_k$ and $\boldsymbol{\epsilon}_k$ in MMSE estimation, the term $\left|\mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbf{h}_k\right\}\right|^2$ is derived as

$$\left|\mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbf{h}_k\right\}\right|^2 = \left|\mathbb{E}\left\{\boldsymbol{\psi}_k^H\left(\hat{\mathbf{h}}_k+\boldsymbol{\epsilon}_k\right)\right\}\right|^2 = 1. \tag{63}$$

Based on the above results, we have

$$\begin{aligned}
&\mathrm{Var}\left(\boldsymbol{\psi}_k^H\mathbf{h}_k\right)\\
&= \mathbb{E}\left\{\left|\boldsymbol{\psi}_k^H\mathbf{h}_k - \mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbf{h}_k\right\}\right|^2\right\}\\
&= \mathbb{E}\left\{\left|\boldsymbol{\psi}_k^H\mathbf{h}_k\right|^2\right\} - \left|\mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbf{h}_k\right\}\right|^2\\
&= \mathbb{E}\left\{\left|\boldsymbol{\psi}_k^H\hat{\mathbf{h}}_k\right|^2\right\} + \mathbb{E}\left\{\left|\boldsymbol{\psi}_k^H\boldsymbol{\varepsilon}_k\right|^2\right\} - 1\\
&= 1 + \mathbb{E}\left\{\boldsymbol{\psi}_k^H\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_k^H\boldsymbol{\psi}_k\right\} - 1\\
&= \mathbb{E}\left\{\boldsymbol{\psi}_k^H\left(\alpha_k-\beta_k\right)\mathbf{I}_M\boldsymbol{\psi}_k\right\}\\
&= \left(\alpha_k-\beta_k\right)\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k\right\|^2\right\},
\end{aligned} \tag{64}$$

where the fifth equality holds owing to (9). To calculate the term $\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k\right\|^2\right\}$, define $\mathbf{\Omega} = diag\left\{\sqrt{\beta_1},\cdots,\sqrt{\beta_K}\right\}$ and $\breve{\mathbf{H}} = \left[\frac{\hat{\mathbf{h}}_1}{\sqrt{\beta_1}},\cdots,\frac{\hat{\mathbf{h}}_K}{\sqrt{\beta_K}}\right]$ with the independent and identical distributed (i.i.d.) columns distributed as $\hat{\mathbf{h}}_i\Big/\sqrt{\beta_i} \sim \mathcal{CN}(\mathbf{0},\mathbf{I}), \forall i \in \mathcal{K}$. Then, the term $\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k\right\|^2\right\}$ can be derived as

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k\right\|^2\right\}\\
&= \mathbb{E}\left\{\left[\mathbf{\Psi}^H\mathbf{\Psi}\right]_{k,k}\right\}\\
&= \mathbb{E}\left\{\left[(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\right]_{k,k}\right\} = \mathbb{E}\left\{\left[(\mathbf{\Omega}\breve{\mathbf{H}}^H\breve{\mathbf{H}}\mathbf{\Omega})^{-1}\right]_{k,k}\right\}\\
&= \frac{1}{\beta_k}\mathbb{E}\left\{\left[(\breve{\mathbf{H}}^H\breve{\mathbf{H}})^{-1}\right]_{k,k}\right\} = \frac{1}{K\beta_k}\mathbb{E}\left\{tr[(\breve{\mathbf{H}}^H\breve{\mathbf{H}})^{-1}]\right\}\\
&= \frac{1}{(M-K)\beta_k},
\end{aligned} \tag{65}$$

where the last equality holds by using the identity , $\mathbb{E}\left\{tr\left(\mathbf{\Xi}^{-1}\right)\right\} = \frac{m}{n-m}$ [35], where $\mathbf{\Xi} \sim \mathcal{W}_m(n,\mathbf{I}_n)$ is an $m\times m$ central complex Wishart matrix with n $(n > m)$ degrees of freedom. As a result, (64) is given by

$$\mathrm{Var}\left(\boldsymbol{\psi}_k^H\mathbf{h}_k\right) = \frac{\alpha_k-\beta_k}{(M-K)\beta_k}. \tag{66}$$

For the variance of effective noise, we adopt the same method as the MRC case to compute it. For simplicity, we only show the nonzero terms in the expansion, which is expressed as

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{\omega}_k^H - \mathbb{E}\left\{\boldsymbol{\omega}_k^H\right\}\right\|^2\right\}\\
&= \mathbb{E}\left\{\left\|\mathbf{v}_{1k}\right\|^2\right\} + \mathbb{E}\left\{\left\|\mathbf{v}_{2k}\right\|^2\right\} - \left\|\mathbb{E}\left\{\mathbf{v}_{2k}\right\}\right\|^2,
\end{aligned} \tag{67}$$

where $\mathbf{v}_{1k} = \sum_{i\in\mathcal{K}\setminus k}\sqrt{\eta_i}T\mathbf{a}_k^H\mathbf{h}_i\mathbf{s}_i^H\mathbf{\Phi}_{[ik]}$, $\mathbf{v}_{2k} = \mathbf{a}_k^H\mathbf{N}\mathbf{W}_k$.

The first term in (68) is calculated as

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\mathbf{v}_{1k}\right\|^2\right\} \\
&= \mathbb{E}\left\{\left\|\sum\nolimits_{i\in\mathcal{K}\setminus k}\sqrt{\eta_i}T\boldsymbol{\psi}_k^H\mathbf{h}_i\mathbf{s}_i^H\boldsymbol{\Phi}_{[ik]}\right\|^2\right\} \\
&= \eta_i T^2\frac{r_{ik}}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k^H\mathbf{h}_i\right\|^2\right\} \\
&= \eta_i T^2\frac{r_{ik}}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k^H\left(\hat{\mathbf{h}}_i+\boldsymbol{\varepsilon}_i\right)\right\|^2\right\} \\
&= \eta_i T^2\frac{r_{ik}}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\left(\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k^H\hat{\mathbf{h}}_i\right\|^2\right\}+\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k^H\boldsymbol{\varepsilon}_i\right\|^2\right\}\right) \\
&= \eta_i T^2\frac{r_{ik}}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\mathbb{E}\left\{\boldsymbol{\psi}_k^H\mathbb{E}\left\{\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^H\right\}\boldsymbol{\psi}_k\right\} \\
&= \eta_i T^2\frac{r_{ik}}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\mathbb{E}\left\{\left\|\boldsymbol{\psi}_k\right\|^2\right\}(\alpha_i-\beta_i) \\
&= \frac{\sum_{i\in\mathcal{K}\setminus k}\eta_i T^2\frac{r_{ik}}{\tau}(\alpha_i-\beta_i)}{(M-K)\beta_k}.
\end{aligned} \tag{68}$$

The second term in (67) can be derived as

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\mathbf{v}_{2k}\right\|^2\right\} \\
&= \mathbb{E}\left\{\left|\boldsymbol{\psi}_k^H\mathbf{N}\mathbf{W}_k\right|^2\right\} \\
&= tr\mathbb{E}\left\{\mathbf{W}_k^H\mathbf{N}^H\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\mathbf{N}\mathbf{W}_k\right\} \\
&= \sum_{i,j,l,r,t}\mathbb{E}\left\{\left[\mathbf{W}_k^H\right]_{i,j}\left[\mathbf{N}^H\right]_{j,l}\left[\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\right]_{l,r}[\mathbf{N}]_{r,t}[\mathbf{W}_k]_{t,i}\right\} \\
&= \sum_{t=j,r=l}\mathbb{E}\left\{\left[\mathbf{W}_k^H\right]_{i,j}\left[\mathbf{N}^H\right]_{j,l}\left[\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\right]_{l,l}[\mathbf{N}]_{l,j}[\mathbf{W}_k]_{j,i}\right\} \\
&= \sum_{i,j,l}\mathbb{E}\left\{\left[\mathbf{W}_k^H\right]_{i,j}[\mathbf{W}_k]_{j,i}\left[\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\right]_{l,l}\right\} \\
&= \sum_{i,j}\left[\mathbf{W}_k^H\right]_{i,j}[\mathbf{W}_k]_{j,i}\sum_{l}\mathbb{E}\left\{\left[\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\right]_{l,l}\right\} \\
&= tr\left\{\mathbf{W}_k^H\mathbf{W}_k\right\}\mathbb{E}\left\{tr\left\{\boldsymbol{\psi}_k\boldsymbol{\psi}_k^H\right\}\right\} \\
&= \frac{T\tau}{(M-K)\beta_k}.
\end{aligned} \tag{69}$$

The value of $\left\|\mathbb{E}\left\{\mathbf{v}_{2k}\right\}\right\|^2$ in (67) is very small compared with the other two terms, and thus we set it to zero. Finally, by combining (63), (64), (67)-(69), we can obtain (36).

REFERENCES

[1] X. Zhou, Y. Zhu, W. Xia, J. Zhang, and K.-K. Wong, "Generalized superimposed pilot enabled URLLC in the finite blocklength regime," in *2023 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2023, pp. 1499–1504.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, May 2020.

[3] Z. Meng, C. She, G. Zhao, and D. De Martini, "Sampling, communication, and prediction co-design for synchronizing the real-world device and digital model in metaverse," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 288–300, Jan. 2023.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] A. A. Nasir, H. D. Tuan, H. H. Nguyen, M. Debbah, and H. V. Poor, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[10] A. A. Nasir, H. D. Tuan, H. Q. Ngo, T. Q. Duong, and H. V. Poor, "Cell-free massive MIMO in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5861–5871, Sept. 2021.

[11] H. Ren, K. Wang, and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 707–723, Jan. 2022.

[12] Q. Peng, H. Ren, C. Pan, N. Liu, and M. Elkashlan, "Resource allocation for uplink cell-free massive MIMO enabled URLLC in a smart factory," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 553–568, Jan. 2023.

[13] Y. Lin, C. Shen, Y. Hu, B. Ai, and Z. Zhong, "Joint design of channel training and data transmission for MISO-URLLC systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8646–8659, Oct. 2022.

[14] C. Li, S. Yan, N. Yang, and X. Zhou, "Truncated channel inversion power control to enable one-way URLLC with imperfect channel reciprocity," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2313–2327, Apr. 2022.

[15] J. Cao, X. Zhu, Y. Jiang, Y. Liu, Z. Wei, S. Sun, and F.-C. Zheng, "Independent pilots versus shared pilots: Short frame structure optimization for heterogeneous-traffic URLLC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5755–5769, Aug. 2022.

[16] M. Mousaei and B. Smida, "Optimizing pilot overhead for ultra-reliable short-packet transmission," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–5.

[17] H. Ji, W. Kim, and B. Shim, "Pilot-less sparse vector coding for short packet transmission," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1036–1039, Aug. 2019.

[18] J. Choi and N. Lee, "Generalized differential index modulation for pilot-free communications," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7973–7984, May 2021.

[19] X. Zhang, D. Zhang, B. Shim, G. Han, D. Zhang, and T. Sato, "Sparse superimposed coding for short-packet URLLC," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5275–5289, Apr. 2022.

[20] 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.

[21] 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.

[22] 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.

[23] X. Jing, M. Li, H. Liu, S. Li, and G. Pan, "Superimposed pilot optimization design and channel estimation for multiuser massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11 818–11 832, Dec. 2018.

[24] J. Li, C. Yuen, D. Li, X. Wu, and H. Zhang, "On hybrid pilot for channel estimation in massive MIMO uplink," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6670–6685, Jul. 2019.

[25] L. A. Lago, Y. Zhang, N. Akbar, Z. Fei, N. Yang, and Z. He, "Pilot decontamination based on superimposed pilots assisted by time-multiplexed pilots in massive MIMO networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 405–417, Jan. 2020.

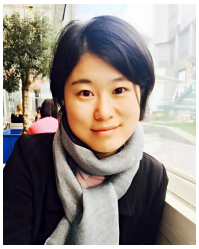
[26] N. Garg, A. Jain, and G. Sharma, "Partially loaded superimposed training scheme for large MIMO uplink systems," *Wireless Pers. Commun.*, vol. 100, no. 4, pp. 1313–1338, Jun. 2018.

[27] N. Garg and T. Ratnarajah, "Generalized superimposed training scheme in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7668–7681, Sep. 2022.

[28] 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.

[29] 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.
[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] 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.
[32] 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.
[33] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. (2016). [Online]. Available: http://cvxr.com/cvx
[34] 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.
[35] A. Tulino and S. Verdú, *Random Matrix Theory and Wireless Communications*. Found. Trends Commun. Inf. Theory, 2004, vol. 1, no. 1.
[36] 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.
[37] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, Jun. 2017.
[38] H. Yan, A. Ashikhmin, and H. Yang, "Can massive MIMO support URLLC?" in *in Proc. IEEE 93rd Veh. Technol. Conf.*, Oct Apr. 2021, pp. 1–5.
[39] J. Chen, L. Zhang, Y.-C. Liang, X. Kang, and R. Zhang, "Resource allocation for wireless-powered IoT networks with short packet communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1447–1461, Feb. 2019.

Xingguang Zhou received the B.S. degree from the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020. He is currently pursuing the Ph.D. degree in Information and Communication Engineering with the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests lie in ultra-reliable and low-latency communication, short packet communication, massive MIMO, and channel estimation.



Yongxu Zhu earned her Ph.D. degree from University College London, U.K., in 2017. She then pursued a Postdoctoral Fellowship at Loughborough University, U.K. Following this, she joined the Department of Engineering at London South Bank University, U.K., where she held the positions of Lecturer and Senior Lecturer consecutively from 2019 to 2022. Following this, she was an Assistant Professor at the University of Warwick from 2022 to 2023. Currently, she is a Professor at Southeast University. Her research interests include B5G/6G, heterogeneous networks, non-terrestrial networks, and physical-layer security. She also serves as an Editor for the IEEE Wireless Communications Letters and the IEEE Transactions on Wireless Communications.

Wenchao Xia (S'16-M'19) received his B.S. degree in communication engineering and Ph.D. degree in communication and information systems from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014 and 2019, respectively. From 2019 to 2020, he was a Postdoctoral Research Fellow with Singapore University of Technology and Design, Singapore. He is currently with the faculty of the Jiangsu Key Laboratory of Wireless Communications, College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. His research interests include edge intelligence and multi-antenna communications.

He was a recipient of the IEEE Globecom Best Paper Award in 2016 and the IEEE JC&S Best Paper Award in 2022. He serves as an Associate Editor for the IET Electronics Letters.

Jun Zhang (S'10–M'14–SM'21) received the M.S. degree in Statistics with Department of Mathematics from Southeast University, Nanjing, China, in 2009, and the Ph.D. degree in Communications Information System with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China, in 2013. From 2013 to 2015, he was a Postdoctoral Research Fellow with Singapore University of Technology and Design, Singapore. Since 2015, he is with the faculty of the Jiangsu Key Laboratory of Wireless Communications, College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, where he is currently a Professor. His research interests include massive MIMO communications, RIS-assisted wireless communications, UAV-assisted wireless communications, physical layer security, and large dimensional random matrix theory. Dr. Zhang was a recipient of the Globcom Best Paper Award in 2016, the IEEE APCC Best Paper Award in 2017, the IEEE JC&S Best Paper Award in 2022, and the IEEE/CIC ICCC Best Paper Award in 2023. He has served as an Associate Editor for the IEEE COMMUNICATIONS LETTERS.

Kai-Kit Wong (M'01-SM'08-F'16) received the BEng, the MPhil, and the PhD degrees, all in Electrical and Electronic Engineering, from the Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions at the University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group of Stanford University, and the University of Hull, UK. He is Chair in Wireless Communications at the Department of Electronic and Electrical Engineering, University College London, UK. His current research centers around 6G and beyond mobile communications. He is Fellow of IEEE and IET. He served as the Editor-in-Chief for IEEE Wireless Communications Letters between 2020 and 2023.