Generalized Superimposed Pilot Enabled URLLC in the Finite Blocklength Regime

Xingguang Zhou[1], Yongxu Zhu[2], Wenchao Xia[1], Jun Zhang[1], Kai-Kit Wong[3]
[1] Jiangsu Key Laboratory of Wireless Communications,
Nanjing University of Posts and Telecommunications, Nanjing, China
[2] School of Engineering, University of Warwick, London, U.K.
[3] Department of Electronic and Electrical Engineering, University College London, London, U.K.
Email: {2020010304, xiawenchao, zhangjun}@njupt.edu.cn,
Yongxu.Zhu@warwick.ac.uk, kai-kit.wong@ucl.ac.uk

***Abstract*—Conventional regular pilot (RP) scheme is not applicable for ultra-reliable and low-latency communication (URLLC) due to the impact of finite blocklength. 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 existing 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 optimal data length and achievable rate lower bound with maximum-ratio combining (MRC) detector. Based on the closed-form results, we provide the corresponding iterative algorithm where the problem is transformed into the geometry program format by using log-function approximation method. 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.**

I. Introduction

The sixth generation (6G) wireless system is anticipated to have more stringent requirements in terms of ultra-high reliability, capacity, energy efficiency, and low latency compared to 5G. 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 autonomous vehicles, virtual reality, and tactile Internet [1]. Typical key performance indicators (KPIs) for URLLC of 5G refer to 1-millisecond end-to-end latency and 10^{-5} decoding error rate for a packet with 32 bytes [2], which will be improved by one order of magnitude in 6G. Low latency implies a finite blocklength or number of channel uses. In other words, the packet size or the codeword length is very short. Hence, short packet transmission is the primary feature of URLLC scenarios. However, the relevant research is still in its infancy since the targets above are conflicting and challenging to satisfy at the same time.

Channel estimation plays a critical role in mandating URLLC. The accuracy of channel estimation determines the level of signal-to-interference-plus-noise ratio (SINR), influencing the reliability and transmission rate. Some works have studied pilot-based short packet transmission for URLLC. In [3], joint power control for uplink URLLC in a cell-free massive multi-input multi-output (mMIMO) system was investigated. Moreover, [4] claimed that the imperfection of channel reciprocity has an influence on one-way URLLC with channel inversion power control. The performance of the independent pilot and shared pilot in the downlink URLLC transmission was compared in [5]. In the aforementioned literature, the short packet transmission is based on the conventional regular pilot (RP) scheme where the data blocklength becomes smaller, which results in a larger rate degradation due to more significant impact of finite blocklength.

To reduce the influence of finite blocklength, we proposed exploiting superimposed pilot (SP) scheme in URLLC transmission [6]. 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. However, the SP scheme suffers from mutual interference (MI) between the pilot and data, impairing the accuracy of channel and data estimation. The generalized superimposed pilot (GSP) scheme in [7] has been recently proposed to enhance the performance of the SP scheme 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. However, existing studies of the GSP scheme [8, 9] are based on the assumption of infinite blocklength and the optimal payload length has not been given in a closed form. The merit of the GSP scheme for URLLC is not well

The work is supported in part by the National Natural Science Foundation of China under Grant 92067201, Grant 62201285, Grant 62071247; in part by Royal Societys International Exchanges Scheme under Grant IEC\NSFC\211011; in part by 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 2022M722669; in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001, and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX22_0943.

understood.

Motivated by the aforementioned facts, this paper investigates the short packet transmission performance of the GSP scheme in a multi-user uplink mMIMO system. Specifically, the contributions are summarized as follows:

- We avoid the generation of MI by ensuring the orthogonality of the precoding matrix. Considering imperfect channel estimation and maximum-ratio combining (MR-C) detection, a weighted sum rate maximization problem under given delay and reliability targets is formulated. To simplify the problem, the closed-form achievable rate lower bound (LB) with finite blocklength is derived.
- The optimization problem is a mixed-integer nonlinear programming problem and hard to obtain a globally optimal solution. We first derive the optimal data length in a closed form. Then, we use log-functions 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.
- The simulation results suggest that the GSP scheme is superior to the RP and SP schemes, which emphasizes the robustness of the GSP scheme in URLLC.

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. To achieve URLLC, we stipulate that each transmission block contains no more than T channel uses for the given decoding error probability ε. We consider the quasi-static Rayleigh fading channel, i.e., $\mathbf{h}_i \sim \mathcal{CN}\left(\mathbf{0}, \alpha_i \mathbf{I}_M\right)$, where α_i denotes the large-scale fading between the i-th user and the BS.

B. 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 $\mathcal{CN}\left(0, 1\right)$.

Under the superimposed scheme, the data and pilot will interfere with each other in channel estimation and data detection [10]. To ensure transmission reliability, in this paper, we eliminate the MI by designing the precoding matrix carefully.

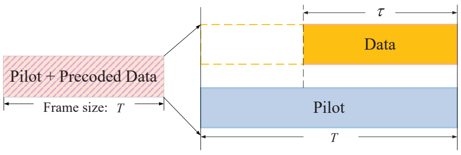

Fig. 1. Frame structure of generalized superimposed pilot.

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.

Based on the above discussion, 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 the i-th user, the transmit power has the following constraint:

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

C. Channel Estimation

At the BS, the received signal is de-spread by multiplying $\mathbf{Y}$ with $\boldsymbol{\varphi}_k \Big/ \sqrt{T}$, which yields

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

where $\mathbf{n}_k = \mathbf{N}\frac{\boldsymbol{\varphi}_k}{\sqrt{T}} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}_M\right)$. 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 $\lambda_k = \frac{\sqrt{\rho_k T\alpha_k}}{\rho_k T\alpha_k + 1}$. It should be noted that the MMSE estimate and estimation error are independent of each other. 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

$$\mathbb{E}\left\{\hat{\mathbf{h}}_k\hat{\mathbf{h}}_k^H\right\} = \beta_k\mathbf{I}_M, \quad \mathbb{E}\left\{\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^H\right\} = \left(\alpha_k - \beta_k\right)\mathbf{I}_M, \tag{7}$$

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

D. Data Detection

After the channel estimation, the conventional MRC detector is employed to perform data detection. Using the channel estimate $\hat{\mathbf{h}}_k$ and known precoding matrix $\mathbf{W}_k$ to process the received signal, the data of the k-th user is recovered as

$$\begin{aligned}\hat{\mathbf{s}}_k^H &= \hat{\mathbf{h}}_k^H \sum\nolimits_{i\in\mathcal{K}} \sqrt{\rho_i}\mathbf{h}_i \boldsymbol{\varphi}_i^H \mathbf{W}_k \\ &+ \hat{\mathbf{h}}_k^H \sum\nolimits_{i\in\mathcal{K}} \sqrt{\eta_i}\mathbf{h}_i \mathbf{s}_i^H \mathbf{W}_i^H \mathbf{W}_k + \hat{\mathbf{h}}_k^H \mathbf{N}\mathbf{W}_k,\end{aligned} \tag{8}$$

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 [11]. Thus, we use the technique to rewrite (8) as

$$\begin{aligned}\hat{\mathbf{s}}_k^H &= \sqrt{\eta_k}\beta_k\alpha_k^{-1}\mathbb{E}\left\{\mathbf{h}_k^H\mathbf{h}_k\right\}\mathbf{s}_k^H\mathbf{W}_k^H\mathbf{W}_k \\ &+ \sqrt{\eta_k}\beta_k\alpha_k^{-1}\left(\mathbf{h}_k^H\mathbf{h}_k - \mathbb{E}\left\{\mathbf{h}_k^H\mathbf{h}_k\right\}\right)\mathbf{s}_k^H\mathbf{W}_k^H\mathbf{W}_k + \boldsymbol{\omega}_k,\end{aligned} \tag{9}$$

where the effective noise $\boldsymbol{\omega}_k$ are defined as

$$\begin{aligned}\boldsymbol{\omega}_k &= \sqrt{\eta_k}\bar{\mathbf{h}}_k^H\mathbf{h}_k\mathbf{s}_k^H\mathbf{W}_k^H\mathbf{W}_k \\ &+ \sqrt{\eta_i}\sum\nolimits_{i\in\mathcal{K}\backslash k}\hat{\mathbf{h}}_k^H\mathbf{h}_i\mathbf{s}_i^H\mathbf{W}_i^H\mathbf{W}_k + \hat{\mathbf{h}}_k^H\mathbf{N}\mathbf{W}_k,\end{aligned} \tag{10}$$

with $\bar{\mathbf{h}}_k = \hat{\mathbf{h}}_k - \lambda_k\sqrt{\rho_k T}\mathbf{h}_k$. From (10), it can be observed that the interference term related to pilot in (8) is removed. Thus, the effective SINR of the k-th user can be expressed as

$$\gamma_k = \frac{\eta_k T^2 \left|\mathbb{E}\left\{\beta_k\alpha_k^{-1}\mathbf{h}_k^H\mathbf{h}_k\right\}\right|^2}{\eta_k T^2 \mathrm{Var}\left(\beta_k\alpha_k^{-1}\mathbf{h}_k^H\mathbf{h}_k\right) + \mathbb{E}\left\{\left\|\boldsymbol{\omega}_k{}^H - \mathbb{E}\left\{\boldsymbol{\omega}_k{}^H\right\}\right\|^2\right\}}. \tag{11}$$

E. 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 [12], i.e.,

$$R_k = \frac{\tau}{T}\mathbb{E}\left[\log_2\left(1+\Gamma_k\right) - \frac{Q^{-1}\left(\varepsilon\right)}{\mathrm{In}2\sqrt{T}}\sqrt{V\left(\Gamma_k\right)}\right], \tag{12}$$

where $V\left(\Gamma_k\right) = 1 - \frac{1}{(1+\Gamma_k)^2}$, Γ_k is the instantaneous SINR for the k-th user, $Q^{-1}\left(\cdot\right)$ 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.

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 (12), the problem can be formulated as

$$\mathbf{P1}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\tau} \sum\nolimits_{k\in\mathcal{K}} \mu_k R_k \tag{13a}$$

$$\text{s.t. } R_k \geqslant R_{\min},\ \forall k, \tag{13b}$$

$$T\left(\rho_k + \eta_k\right) \leqslant E,\ \forall k, \tag{13c}$$

$$1 \leqslant \tau \leqslant T-K, \tau \in \mathbf{N}^+, \tag{13d}$$

where $\boldsymbol{\rho} = \{\rho_k, \forall k\}$, $\boldsymbol{\eta} = \{\eta_k, \forall k\}$, μ_k is the weight of the k-th user, (13b) is the minimum rate constraint imposed for each user, (13c) is the energy constraint, and (13d) means that the 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 nonlinear programming problem such as $\mathbf{P1}$ is very challenging. Instead, we seek the locally optimal solution to $\mathbf{P1}$ by means of an efficient algorithm in an iterative manner.

III. Optimizing Payload Length and Power Allocation for the GSP Scheme

In this section, we aim to solve the optimization problem in (13). Unfortunately, deriving the closed-form expression of the ergodic achievable rate is extremely difficult. In contrast, its LB is readily available [13] and can be expressed as

$$R_k \geqslant \hat{R}_k \triangleq \frac{\tau}{T\mathrm{In}2}\Phi\left(\mathbb{E}\left\{\left(\Gamma_k\right)^{-1}\right\}\right), \tag{14}$$

where $\Phi\left(x\right) = \mathrm{In}\left(1+1/x\right) - \frac{Q^{-1}(\varepsilon)}{\sqrt{T}}\sqrt{V\left(1/x\right)}$. Based on (11) and (14), 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

$$\hat{R}_k \triangleq \frac{\tau}{T\mathrm{In}2}\Phi\left(\gamma_k^{-1}\right), \tag{15}$$

where the effective SINR (11) is further written as

$$\gamma_k = \frac{M\rho_k\eta_k\alpha_k^2}{\left(\rho_k\alpha_k + \frac{1}{T}\right)\left(\eta_k\alpha_k + \vartheta_k + \frac{\tau}{T}\right)}, \tag{16}$$

with $\vartheta_k = \frac{1}{\tau}\sum_{i\in\mathcal{K}\backslash k}\eta_i\alpha_i r_{ik}$.

Proof: Please refer to Appendix A.

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

A. Optimal Data Length

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 addition, (16) is a complicated function with respect to (w.r.t.) the data length τ. To derive the optimal data length in a closed form, we have the result in the following theorem.

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

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 $q_k = \left(1-\kappa\right)P$, 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.

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

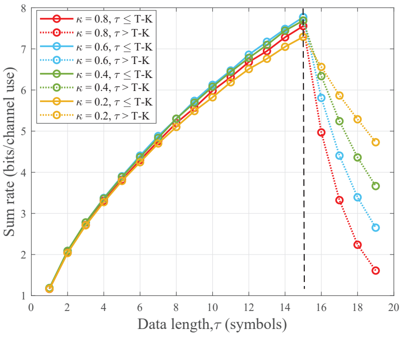

Fig. 2. Sum rate versus data length with different power allocation factor κ for the MRC detector, where $K=5$, $T=20$, $\varepsilon=10^{-5}$, and $M=100$.

B. Optimal Power Allocation

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 (13c). For ease of tractability, we introduce the following lemma.

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

Proof: The first derivative w.r.t. γ_k is given by $\hat{R}'_k = -\frac{\tau}{T\text{In}2}\gamma_k^{-2}\Phi'\left(\gamma_k^{-1}\right)$. $\hat{R}_k$ is always non-negative under the minimum rate constraint. According to [13], the feasible region of γ_k is $\Theta = \left\{x\left|0<1/\gamma_k \leqslant g^{-1}\left(\delta\right)\right.\right\}$, where $\delta = \frac{Q^{-1}(\varepsilon)}{\sqrt{T}}$. Therefore, Lemma 1 in [13] 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 **P1** as the following optimization problem.

$$\mathbf{P2}: \max_{\boldsymbol{\rho},\boldsymbol{\eta}} \sum\nolimits_{k\in\mathcal{K}} \mu_k \hat{R}_k \tag{17a}$$

$$\text{s.t. } \gamma_k \geqslant 1\Big/\Phi^{-1}\left(\text{In}2\frac{R_{\min}T}{\tau}\right),\ \forall k, \tag{17b}$$

$$T\left(\rho_k+\eta_k\right) \leqslant E,\ \forall k, \tag{17c}$$

where (17b) is obtained by applying Lemma 2. However, the objective function (17a) is very complex, which hinders the goal of obtaining a solution. Hence, it is necessary to simplify the objective function (17a). First of all, we introduce auxiliary variables $\boldsymbol{\upsilon} = \{\upsilon_k, \forall k\}$ to transform **P2** into the following equivalent problem:

$$\mathbf{P3}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}} \sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\text{In}\left(1+\upsilon_k\right) - \delta P\left(\upsilon_k\right)\right] \tag{18a}$$

$$\text{s.t. } \gamma_k \geqslant \upsilon_k,\ \forall k, \tag{18b}$$

$$\upsilon_k \geqslant 1\Big/\Phi^{-1}\left(\text{In}2\frac{R_{\min}T}{\tau}\right),\ (17c), \forall k, \tag{18c}$$

where $\varpi_k = \frac{\tau\mu_k}{T\text{In}2}$ and $P\left(\upsilon_k\right) = \sqrt{V\left(\upsilon_k\right)}$. Note that **P2** and **P3** have the same solutions and optimal value, which can be proved by exploiting the contradiction method. Obviously, the objective function (18a) is still a complicated function. To turn (18a) 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 $P\left(y\right)$ is upper bounded by [13][1]

$$P\left(y\right) \leqslant \sigma\text{In}\left(y\right) + \theta \triangleq H\left(y\right), \tag{19}$$

where σ and θ are defined as, respectively,

$$\sigma = \frac{t}{\sqrt{t^2+2t}} - \frac{t\sqrt{t^2+2t}}{\left(1+t\right)^2},\ \theta = \sqrt{1-\frac{1}{\left(1+t\right)^2}} - \sigma\text{In}\left(t\right). \tag{20}$$

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

Proof: Please refer to Appendix C in [13].

Lemma 4: Given $t \geqslant 0$, $\forall y \geqslant 0$, the LB of function $\text{In}\left(1+y\right)$ is given by [13]

$$\text{In}\left(1+y\right) \geqslant \hat{\sigma}\text{In}\left(y\right) + \hat{\theta}, \tag{21}$$

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

$$\hat{\sigma} = \frac{t}{1+t},\ \hat{\theta} = \text{In}\left(1+t\right) - \frac{t}{1+t}\text{In}\left(t\right). \tag{22}$$

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 (18a), which enables us to solve **P3**. The main idea is to develop an iterative algorithm where the LB is updated to approximate (18a) 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 (20) and (22), the objective function (18a) 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 (19) and (21) , we acquire the LB of (18a) in the $n+1$-th iteration as

$$\begin{aligned}&\sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\text{In}\left(1+\upsilon_k\right) - \delta P\left(\upsilon_k\right)\right] \\ \geqslant &\sum\nolimits_{k\in\mathcal{K}} \varpi_k \left[\hat{\sigma}_k^{(n)}\text{In}\upsilon_k + \hat{\theta}_k^{(n)} - \delta\sigma_k^{(n)}\text{In}\upsilon_k - \delta\theta_k^{(n)}\right].\end{aligned} \tag{23}$$

Besides, the LB is tight at $\upsilon_k = \upsilon_k^{(n)}$. Thus, the objective function (18a) is replaced by its LB and **P3** is transformed as

$$\mathbf{P4}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}} \sum\nolimits_{k\in\mathcal{K}} \chi_k^{(n)}\text{In}\upsilon_k \tag{24a}$$

$$\text{s.t. } (18b), (18c), \forall k \tag{24b}$$

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. Further, **P4** can be turned into the following GP problem [14]:

$$\mathbf{P5}: \max_{\boldsymbol{\rho},\boldsymbol{\eta},\boldsymbol{\upsilon}} \prod\nolimits_{k\in\mathcal{K}} \upsilon_k^{\chi_k^{(n)}} \tag{25a}$$

$$\text{s.t. } \begin{aligned}&\left(\rho_k\alpha_k + \tfrac{1}{T}\right)\left(\sum\nolimits_{i\in\mathcal{K}}\eta_i\alpha_i + 1 - \tfrac{K}{T}\right)\upsilon_k \\ &\leqslant M\rho_k\eta_k\alpha_k^2,\ \forall k,\end{aligned} \tag{25b}$$

$$\upsilon_k \geqslant 1\Big/\Phi^{-1}\left(\text{In}2\frac{R_{\min}T}{\tau}\right),\ (17c),\ \forall k. \tag{25c}$$

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

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 [16]. The detailed algorithm to solve $\mathbf{P5}$ is shown in Algorithm 1 The convergence of Algorithm 1 can be verified by using the method in [13]. Besides, the computational complexity of the proposed algorithm is on the order of $\mathcal{O}\left(N_{iter} \times \max\left\{27K^3, N_{cost}\right\}\right)$, where N_{iter} is the number of iterations and N_{cost} is the computational complexity of calculating the first-order and second-order derivatives of the objective function and constraint functions of $\mathbf{P5}$ [3].

```
Algorithm 1 Algorithm to solve P5.
1: Input: iteration number $n = 1$, error tolerance $\xi$, 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 (20), (22), 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.
6: Output: Locally optimal power allocation $\{\rho_k^*, \eta_k^*, \forall k\}$.
```

To trigger the iteration process, we need to provide an initial solution for the algorithm, which can be obtained by solving the following optimization problem:

$$\mathbf{P6}: \max_{\boldsymbol{\rho}, \boldsymbol{\eta}, \nu} \ \nu \tag{26a}$$

$$\text{s.t. } \gamma_k \geqslant \nu \Big/ \Phi^{-1}\left(\text{In}2 \frac{R_{\min} T}{\tau}\right), \ (17c), \ \forall k, \tag{26b}$$

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.

IV. 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 $\text{PL}_k = 35.3 + 37.6\log_{10} d_k$ (dB) [17]. The noise power spectral density and the decoding error rate ε are set as -174 dBm/Hz and 10^{-9}, respectively. For comparison, the SP and RP schemes are simulated in the same E and T as the GSP scheme. Note that the simulation results of the SP and RP schemes are optimal, including pilot length and power allocation[2]. For the SP scheme, we consider the practical situation of imperfect pilot interference removal (IPPR) [6]. We average 100 trails to obtain the Monte-Carlo results, in which the locations of users are randomly generated in each trail.

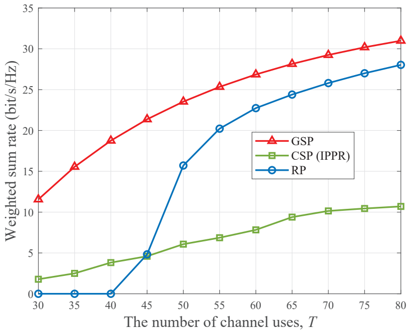

Fig. 3. Impact of the number of channel uses on the weighted sum rate: $K = 20$, $E = -4$ dB, $M = 200$, and $R_{\min} = \frac{T-K}{T\text{In}2}$.

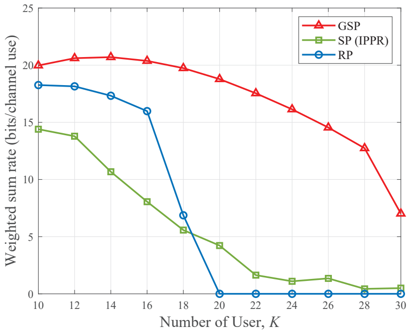

Fig. 4. Impact of the number of users on the weighted sum rate: $T = 40$, $E = -4$ dB, $M = 100$, and $R_{\min} = \frac{T-K}{T\text{In}2}$.

Fig. 3 shows the impact of the number of channel uses on the weighted sum rate. As expected, the rate increases with the growth of the number of channel uses because of the availability of more resources. It should be noted that the GSP scheme is superior to the SP and RP schemes, especially in the small channel uses region, which justifies that the GSP scheme is appropriate for finite blocklength transmission. Besides, 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 $T = 45$.

Fig. 4 depicts the impact of the number of users on the weighted sum rate. Owing to multi-user diversity, it is clear to see that the GSP scheme increases with user number in a small K. Then, 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. In other words, the GSP scheme can support more users at a higher rate than the RP and SP schemes. 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

[2] The optimal pilot length of the RP scheme is equal to the number of users, which can be proved by using Lemma 1. Besides, the algorithms of the RP and SP schemes have the same complexity as Algorithm 1.

overhead, which indicates the RP scheme is not suitable for massive URLLC transmission.

V. Conclusion

In this paper, we investigated the GSP scheme supporting URLLC in mMIMO systems. 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 jointly optimized the data length, pilot and data power allocation to maximize the weighted sum rate. The optimal data length and achievable rate LB for the MRC detection were derived in a closed form, respectively, which shows that the optimal data length of the GSP scheme is equal to $T-K$. Based on the results, the optimization problem can be transformed into a sequence of GP problems, and we then developed an iterative algorithm to obtain a locally optimal solution. Simulation results demonstrated that the GSP scheme is superior to the conventional RP and SP schemes in supporting URLLC with massive connectivity.

Appendix A
Proof of Theorem 1

To calculate the expectations and variances in (11), we have

$$\left|\beta_k\alpha_k^{-1}\mathbb{E}\left\{\mathbf{h}_k^H\mathbf{h}_k\right\}\right|^2 = M^2\alpha_k^2,\ \mathrm{Var}\left(\beta_k\alpha_k^{-1}\mathbf{h}_k^H\mathbf{h}_k\right) = M\alpha_k^2. \tag{27}$$

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

$$\begin{aligned}&\mathbb{E}\left\{\left\|\boldsymbol{\omega}_k^H-\mathbb{E}\left\{\boldsymbol{\omega}_k^H\right\}\right\|^2\right\} = \mathbb{E}\left\{\left\|\boldsymbol{\omega}_k^H\right\|^2\right\} - \left\|\mathbb{E}\left\{\boldsymbol{\omega}_k^H\right\}\right\|^2\\ &= \sum_{i=1}^{3}\mathbb{E}\left\{\left\|\mathbf{u}_{ik}\right\|^2\right\} - \left\|\mathbb{E}\left\{\mathbf{u}_{ik}\right\}\right\|^2\\ &\quad +2\mathbf{R_e}\left\{\sum_{i=1}^{3}\sum_{j=i+1}^{3}\mathbb{E}\left\{\mathbf{u}_{ik}\mathbf{u}_{jk}^H\right\} - \mathbb{E}\left\{\mathbf{u}_{ik}\right\}\mathbb{E}\left\{\mathbf{u}_{jk}^H\right\}\right\},\end{aligned} \tag{28}$$

where $\mathbf{u}_{1k} = \sqrt{\eta_k}T\bar{\mathbf{h}}_k^H\mathbf{h}_k\mathbf{s}_k^H$, $\mathbf{u}_{2k} = \sqrt{\eta_i}T\sum_{i\in\mathcal{K}\setminus k}\hat{\mathbf{h}}_k^H\mathbf{h}_i\mathbf{s}_i^H\mathbf{\Phi}_{[ik]}$, and $\mathbf{u}_{3k} = \hat{\mathbf{h}}_k^H\mathbf{N}\mathbf{W}_k$. Due to the limited space, we only show the final result of each expectation in the expansion (28).

$$\mathbb{E}\left\{\left\|\mathbf{u}_{1k}\right\|^2\right\} = MT^2\eta_k\beta_k\left(\alpha_k-\beta_k\right), \tag{29}$$

$$\mathbb{E}\left\{\left\|\mathbf{u}_{2k}\right\|^2\right\} = M\beta_k\frac{T^2}{\tau}\sum\nolimits_{i\in\mathcal{K}\setminus k}\eta_i\alpha_i r_{ik}, \tag{30}$$

$$\mathbb{E}\left\{\left\|\mathbf{u}_{3k}\right\|^2\right\} = M\beta_k T\tau. \tag{31}$$

The calculative procedure of the remaining terms in (28) is omitted since they are equal to zero. Finally, we can acquire (16) by substituting (27)-(31) into (11).

Appendix B
Proof of Theorem 2

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

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

Proof: According to the known conditions $f\left(x\right) \geqslant \frac{x}{\text{In}2}$, we have $\text{In}\left(1+\frac{a}{bx+c}\right) - \delta\sqrt{1-\frac{1}{\left(1+\frac{a}{bx+c}\right)^2}} \geqslant 1$. Then, taking the first derivative w.r.t. x, we obtain $f'\left(x\right) = \frac{1}{\text{In}2}J\left(x\right) > 0$, $J\left(x\right) = \text{In}\left(1+\frac{a}{bx+c}\right) - \frac{abx}{(bx+c)(bx+c+a)} - \delta\sqrt{1-\frac{1}{\left(1+\frac{a}{bx+c}\right)^2}} + \delta\frac{abx}{\sqrt{\left(\frac{a}{bx+c}\right)^2+\frac{2a}{bx+c}}(bx+c+a)^2}$, which holds due to $\frac{bx}{bx+c}\cdot\frac{a}{bx+c+a} < 1$. Then, 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. $f\left(x\right)$, the value of x should be taken $\frac{T-K}{T}$, which proves Theorem 2.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] D. Verenzuela, 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications, *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402-415, 2019.

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

[17] 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.