

# Optimization of the Downlink Spectral- and Energy-Efficiency of RIS-aided Multi-user URLLC MIMO Systems

Mohammad Soleymani, *Member, IEEE*, Ignacio Santamaria, *Senior Member, IEEE*,  
Eduard Jorswieck, *Fellow, IEEE*, Robert Schober, *Fellow, IEEE*,  
and Lajos Hanzo, *Life Fellow, IEEE*

**Abstract**—Modern wireless communication systems are expected to provide improved latency and reliability. To meet these expectations, a short packet length is needed, which makes the first-order Shannon rate an inaccurate performance metric for such communication systems. A more accurate approximation of the achievable rates of finite-block-length (FBL) coding regimes is known as the normal approximation (NA). It is therefore of substantial interest to study the optimization of the FBL rate in multi-user multiple-input multiple-output (MIMO) systems, in which each user may transmit and/or receive multiple data streams. Hence, we formulate a general optimization problem for improving the spectral and energy efficiency of multi-user MIMO-aided ultra-reliable low-latency communication (URLLC) systems, which are assisted by reconfigurable intelligent surfaces (RISs). We show that an RIS is capable of substantially improving the performance of multi-user MIMO-aided URLLC systems. Moreover, the benefits of RIS increase as the packet length and/or the tolerable bit error rate are reduced. This reveals that RISs can be even more beneficial in URLLC systems for improving the FBL rates than in conventional systems approaching Shannon rates.

**Index Terms**—Energy efficiency, MIMO broadcast channels, reconfigurable intelligent surface, spectral efficiency, ultra-

reliable low-latency communication.

## I. INTRODUCTION

The sixth generation (6G) of wireless communication systems is expected to significantly improve the spectral efficiency (SE), energy efficiency (EE), and reliability of the existing systems, despite of providing a lower latency than 5G [1], [2]. Thus, 6G should employ radical new technologies such as reconfigurable intelligent surface (RIS) to meet these expectations. Moreover, 6G has to support a large variety of applications, which require ultra-reliable and low-latency communications (URLLC) [1], [2]. To attain low latency, realistic finite block length (FBL) codes have to be employed. In this content, the classic Shannon rate is an inaccurate performance metric for URLLC systems. Indeed, the FBL rate is more challenging to optimize than the Shannon rate, especially in multiple-input multiple-output (MIMO) systems, when multi-stream communication is targeted. In fact, to the best of our knowledge, resource allocation has not been designed for multi-user MIMO (MU-MIMO) systems relying on FBL coding in the open literature for the scenario of multiple streams per user. Developing suitable resource allocation schemes is even more challenging in RIS-assisted systems since this requires the joint optimization of the transmit covariance matrices and the channels, which depend on the RIS elements. To close this knowledge gap, we derive a closed-form expression for the rates of users in MU-MIMO systems using realistic FBL coding when multiple streams are allowed. Then, we develop an optimization framework for MU-MIMO RIS-assisted URLLC systems and show that an RIS can substantially improve the SE and EE. Our results show that an RIS can be even more beneficial in MIMO URLLC systems than in systems approaching the classic Shannon rate, since an RIS provides higher gains for short packets and/or for low tolerable bit error rates (BERs).

## A. Literature Review

A main goal of 6G is to drastically enhance the SE and EE, which are even more vital for applications related to FBL coding. To realize this goal, 6G has to employ powerful emerging technologies such as RISs as well as existing MIMO solutions [3], [4]. An RIS was shown to be able to substantially enhance the EE and SE [5]–[10], when studying different

Mohammad Soleymani is with the Signal and System Theory Group, Universität Paderborn, Germany (email: mohammad.soleymani@uni-paderborn.de).

Ignacio Santamaria is with the Department of Communications Engineering, Universidad de Cantabria, Spain (email: i.santamaria@unican.es).

Eduard Jorswieck is with the Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany (email: jorswieck@ifn.ing.tu-bs.de).

Robert Schober is with the Institute for Digital Communications, Friedrich Alexander University of Erlangen-Nuremberg, Erlangen 91058, Germany (email: robert.schober@fau.de).

Lajos Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom (email: lh@ecs.soton.ac.uk).

The work of Ignacio Santamaria was funded by MCIN/ AEI /10.13039/501100011033, under Grant PID2022-137099NB-C43 (MADDIE), and by European Union's (EU's) Horizon Europe project 6G-SENSES under Grant 101139282. The work of Eduard Jorswieck was supported by the Federal Ministry of Education and Research (BMBF, Germany) through the Program of Souverän. Digital. Vernetzt. joint Project 6G-RIC, under Grant 16KISK031, and by European Union's (EU's) Horizon Europe project 6G-SENSES under Grant 101139282. Lajos Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council (EPSRC) projects under grant EP/Y037243/1, EP/W016605/1, EP/X01228X/1, EP/Y026721/1, EP/W032635/1 and EP/X04047X/1 as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). The work of Robert Schober was partly supported by the Federal Ministry of Education and Research of Germany under the programme of "Souverän. Digital. Vernetzt." joint project 6G-RIC (project identification number: PIN 16KISK023) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant SCHO 831/15-1.

performance metrics such as the sum rate, minimum rate, total power consumption required for achieving a specific quality of service (QoS), minimum signal-to-interference-plus-noise ratio (SINR), global EE (GEE), and interference leakage. For instance in [5], the authors proposed algorithms for increasing the GEE and sum rate of a multiple-input single-output (MISO) broadcast channel (BC). In [6], it was shown that an RIS reduces the power consumption of the single-cell MISO BC when a minimum SINR per user has to be ensured. Moreover, the authors of [7] showed that an RIS is capable of increasing the minimum SINR of a single-cell MISO BC for a given power budget. The benefit of RIS in terms of both the minimum rate and the minimum EE was studied in [8] for a multi-cell MISO BC employing non-orthogonal multiple access (NOMA). In a multi-cell BC, each user is associated with only a single base station (BS) at a time. In [11]–[13], it was shown that RISs can also improve the performance of cell-free systems in which each user simultaneously communicates with several access points, rather than a single BS.

The performance of RIS relying on the Shannon capacity achieving codes has been studied in [5]–[9], but naturally, an RIS can also be beneficial in multi-user systems using FBL coding, as shown in [14]–[19]. For instance, in [14], resource allocation schemes were developed for MISO URLLC systems, assisted by simultaneously transmit and reflect (STAR-) RIS, and it was shown that an RIS (either STAR or purely reflective) substantially improves the SE and EE of URLLC systems. In [15], it was demonstrated that RIS and rate splitting can be mutually beneficial tools of enhancing the EE and SE performance of interference-limited URLLC systems. In [17], the advantage of employing an RIS and NOMA in a two-user single-input single-output (SISO) URLLC BC was shown.

The papers [14]–[19] studied multi-user RIS-assisted URLLC systems, but only supported single-stream data transmission per user. However, RIS can also improve the system performance when parallel frequency-domain channels are employed along with FBL coding [20]. Nevertheless, resource allocation has not been designed for multi-user MIMO URLLC systems supporting multiple streams per user in the open literature. This is particularly challenging for RIS-assisted systems. Indeed, only a few treatises exist in multiple-stream data transmission in MIMO systems, which mainly studied a single-user scenario without considering RISs [21]–[23]. Thus, multi-user MIMO systems both with and without RISs require further investigations. MIMO systems support multiple-stream data transmissions per user, which exploit the spectrum efficiently and improve the EE at a specific QoS. Below, we briefly describe the challenges of optimizing the FBL rates when multiple-stream data transmission is supported.

In the FBL regime, the achievable rate depends not only on the Shannon rate,  $C$ , but also on the channel's dispersion,  $V$ , the packet length,  $n_t$ , and the tolerable bit error rate  $\epsilon$ . An accurate approximation for the achievable rate of FBL coding in parallel channels is the normal approximation (NA)<sup>1</sup>, which

<sup>1</sup>Note that the NA may not be accurate when the packet length is extremely short, and/or the tolerable bit error rate is extremely low. For further discussions regarding the accuracy of the NA, please refer to [24]–[26].

**TABLE I:** Overview of most closely related works on RIS-assisted URLLC systems.

	This paper	[15]	[30]	[16]–[20], [31]	[32], [33]
Multi-user	✓	✓	✓	✓	-
Ch. disp. in [34]	✓	✓	✓	-	-
EE	✓	✓	-	-	-
MIMO	✓	-	-	-	-
Multiple streams	✓	-	-	-	-
STAR-RIS	✓	-	-	-	-
Transmission delay	✓	-	-	-	-

is given by [27, Theorem 78]

$$r = \sum_{i=1}^I C_i - Q^{-1}(\epsilon) \sqrt{\frac{\sum_{i=1}^I V_i}{n_t}}, \quad (1)$$

where  $Q^{-1}$  is the inverse of the well-known  $Q$  function of Gaussian distributions,  $I$  is the number of parallel channels,  $C_i$  and  $V_i$  are, respectively, the Shannon rate and the channel dispersion of the  $i$ -th parallel channel. Note that the channel dispersion and Shannon rate of parallel channels are, respectively, a summation of the channel dispersions and Shannon rates of all individual channels, i.e.,  $C = \sum_{i=1}^I C_i$  and  $\sum_{i=1}^I V_i$ . The Shannon rate of MIMO systems can also be represented in a closed-form matrix format, and there are already existing contributions on optimizing the Shannon rates in MIMO RIS-assisted systems [28], [29]. However, the achievable channel dispersion term for Gaussian signals has a fractional structure, which is more challenging to optimize, and its closed-form matrix format has not been derived in the related works. Hence, resource allocation for parallel channels relying on FBL coding can be much more complicated than for single-stream channels. Moreover, the channel's dispersion term makes it impossible to reuse the existing solutions for MIMO RIS-assisted systems, when FBL coding is employed. In the next subsection, we provide a critical review of the existing works on RIS-assisted URLLC systems and discuss the open topics that merit further investigations.

## B. Motivation

The most closely related treatises on RIS-assisted URLLC system designs are compared in Table I, based on the system model, network scenario, performance metrics, and the channel dispersion encountered in multi-user systems. As shown in the table, most of the studies on FBL transmission in RIS-assisted systems have focused on SISO/MISO systems, when only a single-stream data transmission per user is allowed. Additionally, there is a limited number of contributions on EE in RIS-assisted URLLC systems and EE metrics have not been studied in multi-user MIMO systems using FBL coding. Note that in URLLC systems, the EE can be even more vital, since in some applications, it might be impossible to replace the battery of users/nodes, and consequently, the network must be as energy efficient as possible.

Moreover, there is no work on multi-user MIMO systems with FBL considering the achievable channel dispersion term in [34], even for systems without RIS. It should be emphasized that Gaussian signaling cannot achieve the optimal channel dispersion in the presence of interference in multi-user sys-

tems. Hence, when employing Gaussian signaling, it is more accurate to consider the suboptimal channel dispersion term in [34] instead of the optimal one since the channel dispersion in [34] can be attained through Gaussian signals in the presence of both interference and white additive Gaussian noise in multi-user systems.

To sum up, resource allocation schemes should be developed for multi-user RIS-assisted URLLC MIMO systems with more emphasis on EE and by allowing multiple-data-stream transmissions per user and considering the achievable channel dispersion term for Gaussian signals. Thus, a general optimization framework for multi-user RIS-assisted MIMO systems with FBL coding can facilitate future studies in this field.

### C. Contribution

We propose a general optimization framework for maximizing the SE and EE of multi-user MIMO RIS-aided URLLC systems. To the best of our knowledge, this is the first paper on the resource allocation of MU-MIMO RIS-aided URLLC systems supporting multiple stream per users. To develop a general framework, we first formulate a closed-form expression for the channel dispersion of MIMO systems in the presence of interference, based on both the optimal channel dispersion as well as on the channel dispersion term in [34]. Then, we propose specific schemes for optimizing the FBL rate expressions by employing majorization minimization (MM), alternating optimization (AO), and fractional programming (FP) tools such as Dinkelbach-based algorithms. As indicated in Section I-A, due to the channel dispersion term, which has a fractional structure, the FBL rates are much more challenging to optimize than the classic Shannon rates. Moreover, it is impossible to adapt the established works on MIMO RIS-assisted systems with Shannon rates to the systems with FBL coding, and to evaluate how the reliability and latency constraints influence the effectiveness of RISs. Thus, the main novelty of this treatise is the derivation of closed-form expressions for the channel dispersion, followed by the development of algorithms to optimize over the FBL rates, including the dispersion term.

To elaborate, our optimization framework is flexible and can be utilized in a wide range of MU-MIMO URLLC systems aided by RISs. Additionally, the framework may be used for solving a broad spectrum of optimization problems for which the objective function and/or constraints can be, but are not limited to, linear functions of the rates and/or EE of users. The convergence of our framework is ensured towards a stationary point for the general optimization problem, when the feasibility set for the RIS elements conforms to a convex set. We consider a multi-cell MIMO BC as an example of the networks that our framework can be applied to. Furthermore, we consider both the EE and SE metrics as well as the transmission delay for investigating the performance of RIS in MU-MIMO URLLC systems. For the SE metric, we consider the sum rate and the minimum rate of the users, which are among the most common performance metrics for SE. Moreover, we evaluate the EE by optimizing the GEE and the

minimum EE of users. Note that the sum rate and global EE are pivotal overall system performance metrics. By contrast, the minimum rate and the EE of specific users consider the individual performance of the users and can provide reasonable rate/EE-fairness among the users since typically all the users are allocated similar rate/EE when the minimum rate/EE is maximized. Thus, considering all these metrics can provide a complete picture of the performance of MU-MIMO URLLC systems aided by RISs. Moreover, we make realistic assumptions regarding the channel models and the feasible sets of the RIS coefficients for appropriately examining the RIS performance.

In addition to passive and reflective RISs, we also consider simultaneously transmitting and reflecting (STAR) RIS, which provides a full 360° coverage. Moreover, we show that RISs can significantly enhance the EE and SE of a multi-user MIMO URLLC BC. Notably, the advantages of RISs escalate with shorter packet lengths and/or more stringent reliability constraints. This implies that the benefits of RIS can be higher in MU-MIMO URLLC systems. However, it should be noted that the performance of the system may become degraded, if the RIS elements are inaccurately optimized.

### D. Organization and Notations

The structure of the paper is outlined in the following. Section II describes our network scenario, RIS model, and signal model as well as the rate and EE expressions. Moreover, in Section II, we formulate the optimization problem considered. Section III presents the optimization framework proposed. Section IV presents our numerical results. Finally, Section V concludes the paper.

The trace and determinant of the matrix  $\mathbf{X}$  are, respectively, denoted as  $\text{Tr}(\mathbf{X})$  and  $|\mathbf{X}|$ . We represent the conjugate of complex variable/vector/matrix  $x/\mathbf{x}/\mathbf{X}$  as  $x^*/\mathbf{x}^*/\mathbf{X}^*$ . The mathematical expectation is denoted as  $\mathbb{E}\{\cdot\}$ . The identity matrix is represented as  $\mathbf{I}$ . Moreover,  $\mathcal{O}[\cdot]$  is the big-O notation for representing the computational complexity of algorithms.

## II. SYSTEM MODEL

Our proposed framework can be applied to a large family of RIS-assisted MU-MIMO URLLC systems that treat interference as noise at the receivers. As an example of such MU-MIMO systems, we consider a multi-cell MIMO RIS-assisted downlink (DL) BC comprising  $L$  BSs, as shown in Fig. 1. We assume that BS  $l$  has  $N_{BS,l}$  DL transmission antennas (TAs) and serves  $K_l$  multiple-antenna users. The  $k$ -th user associated to BS  $l$ , denoted as  $\mathbf{U}_{lk}$ , has  $N_{u,lk}$  receive antennas (RAs). Additionally, we assume that there are  $M$  reflective passive RISs to assist the BSs, and the  $m$ -th RIS has  $N_{RIS,m}$  elements. Furthermore, we assume perfect, instantaneous, and global CSI, consistent with many other studies on RISs [5], [6], [8], [28], [35], [36]. This assumption is also commonly used in the development of resource allocation solutions for URLLC systems [14], [31], [37]–[40]. These solutions are particularly applicable in systems with large channel coherence time, where the channel state remains stable for extended periods, making channel estimation easier and more accurate.

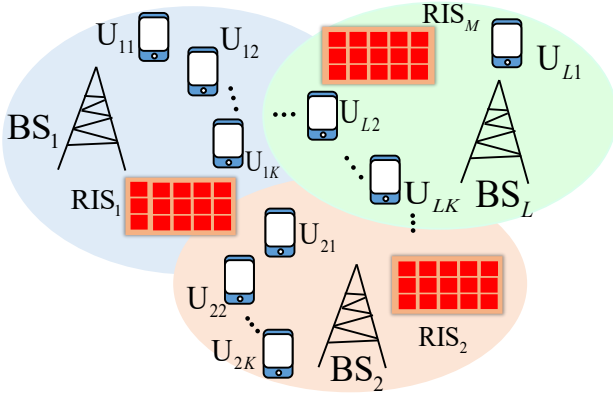


Fig. 1: A multi-cell BC assisted by RISs.

In such systems, resource allocation solutions can be reused across multiple time slots, and the pilot overhead required for acquiring CSI and computing the solutions is relatively low. Additionally, investigating the performance of RISs under the assumption of perfect CSI helps to illustrate the essential tradeoffs in system design and provides an upper bound of the system performance.

#### A. RIS Model

We consider two nearly-passive RIS architectures, namely reflective RIS and STAR-RIS, and employ the RIS model of [35] for the MIMO multi-cell BC.

1) *Reflective RIS*: The channel matrix between BS  $i$  and  $U_{lk}$  as a function of the RIS matrices is given by

$$\mathbf{H}_{lk,i}(\{\Psi\}) = \underbrace{\sum_{m=1}^M \mathbf{G}_{lk,m} \Psi_m \mathbf{G}_{m,i}}_{\text{Link through RIS}} + \underbrace{\mathbf{F}_{lk,i}}_{\text{Direct link}} \in \mathbb{C}^{N_{u,lk} \times N_{BS,i}}, \quad (2)$$

where  $\mathbf{F}_{lk,i} \in \mathbb{C}^{N_{u,lk} \times N_{BS,i}}$  is the channel matrix between the  $i$ -th BS and  $U_{lk}$ ,  $\mathbf{G}_{lk,m} \in \mathbb{C}^{N_{u,lk} \times N_{RIS,m}}$  is the channel matrix between the  $m$ -th RIS and  $U_{lk}$ , and  $\mathbf{G}_{m,i} \in \mathbb{C}^{N_{RIS,m} \times N_{BS,i}}$  is the channel matrix between the  $i$ -th BS and the  $m$ -th RIS. Additionally,  $\{\Psi\} = \{\Psi_m\}_{m=1}^M$  denotes the set of all coefficients of RISs, where  $\Psi_m \in \mathbb{C}^{N_{RIS,m} \times N_{RIS,m}}$  is a diagonal matrix, containing the vector of reflecting coefficients of the  $m$ -th RIS

$$\Psi_m = \text{diag}(\psi_{m1}, \psi_{m2}, \dots, \psi_{mN_{RIS,m}}).$$

Assuming having nearly passive RISs, the absolute value of the RIS coefficients cannot be greater than 1, which results in the following set for the feasible RIS coefficients [3, Eq. (11)]

$$\mathcal{T}_U = \{\psi_{m,n} : |\psi_{m,n}| \leq 1 \quad \forall m, n\}. \quad (3)$$

In this feasibility set, the amplitude and phase of each RIS element are assumed to be independent optimization variables, which might not be realistic. Another common assumption is that the RIS coefficients have to adhere to the unit modulus constraint [3], [4], [6], [7], [35], [41], [42], which leads to

$$\mathcal{T}_I = \{\psi_{m,n} : |\psi_{m,n}| = 1 \quad \forall m, n\}. \quad (4)$$

In this feasibility set, the amplitude of each RIS coefficient is assumed to be equal to 1, while the phases can be optimized. As  $\mathcal{T}_I \subset \mathcal{T}_U$ , it can be expected that the algorithms for  $\mathcal{T}_U$  outperform the algorithms for  $\mathcal{T}_I$ .

2) *STAR-RIS*: STAR-RIS provides an omni-directional 360° full-space coverage. In STAR-RIS, each component can operate in both reflection and transmission mode [43], [44]. Thus, there are two complex-valued optimization parameters per element, when STAR-RIS is employed. We denote the reflection/transmission coefficient for the  $n$ -th element of the  $m$ -th RIS as  $\psi_{m,n}^r/\psi_{m,n}^t$ . Based on the position of the user with respect to STAR-RIS, the STAR-RIS can optimize the channel of the user only through the reflection or transmission coefficients. Therefore, the channel between BS  $i$  and  $U_{lk}$  is

$$\mathbf{H}_{lk,i}(\{\Psi\}) = \sum_{m=1}^M \mathbf{G}_{lk,m} \Psi_m^{t/r} \mathbf{G}_{m,i} + \mathbf{F}_{lk,i}, \quad (5)$$

where we have  $\Psi_m^r = \text{diag}(\psi_{m1}^r, \psi_{m2}^r, \dots, \psi_{mN_{RIS,m}}^r)$  and  $\Psi_m^t = \text{diag}(\psi_{m1}^t, \psi_{m2}^t, \dots, \psi_{mN_{RIS,m}}^t)$ . Assuming operating in a passive mode, the absolute values of the reflection and transmission coefficients have to satisfy

$$|\psi_{m,n}^r|^2 + |\psi_{m,n}^t|^2 \leq 1, \quad \forall m, n, \quad (6)$$

which yields the set

$$\mathcal{T}_{SU} = \{\psi_{m,n}^r, \psi_{m,n}^t : |\psi_{m,n}^r|^2 + |\psi_{m,n}^t|^2 \leq 1 \quad \forall m, n\}. \quad (7)$$

Assuming operating in the passive mode with equal input and output powers, we have

$$|\psi_{m,n}^r|^2 + |\psi_{m,n}^t|^2 = 1, \quad \forall m, n, \quad (8)$$

which results in

$$\mathcal{T}_{SI} = \{\psi_{m,n}^r, \psi_{m,n}^t : |\psi_{m,n}^r|^2 + |\psi_{m,n}^t|^2 = 1 \quad \forall m, n\}. \quad (9)$$

There are three different STAR-RIS schemes, including the energy splitting (ES), mode switching (MS) and time switching (TS) schemes [44], [45]. Since the main focus of this work is on evaluating the impact of employing multiple streams per user regimes, we consider only the MS scheme. Note that the MS scheme has a lower implementation complexity than the ES scheme, but its performance may be comparable to that of the ES scheme as shown in, e.g., [46]–[48]. The framework proposed in this treatise can be extended to include the ES and TS schemes by following an approach similar to [47], [48].

3) *Brief comparison of reflective RIS and STAR-RIS*: The main difference between a reflective RIS and a STAR-RIS is in the coverage area of these RIS architectures, which makes each suitable for a different set of applications. STAR-RIS can provide omni-directional coverage, while reflective RIS can assist the communication between a BS and a user only if they are in the reflection space of the RIS. Thus, when the RIS can be positioned for ensuring that all the transceivers are in the reflection space of the RIS, a reflective RIS could be a more suitable option. However, when the BS is located outdoors and supports both indoor and outdoor users, STAR-RISs are preferable.



From an optimization point of view, each STAR-RIS element has two complex-valued coefficients. Thus, if the ES scheme is employed, algorithms conceived for STAR-RIS could have slightly higher computational complexities compared to reflective RIS. However, if MS or TS schemes are utilized, each RIS element operates either in the reflection mode or in the transmission mode. Thus, only one coefficient for each STAR-RIS element is optimized, which reduces the computational complexity to the same order as that of reflective RIS algorithms.

The channel matrices are assumed to be linear/affine functions of the RIS elements in both reflective RIS and STAR-RIS. To simplify the notations/equations, we remove this dependency and subsequently denote the channels as  $\mathbf{H}_{lk,i}$  for all  $l, k, i$ , hereafter. Additionally, we denote the set of the feasible RIS elements as  $\mathcal{T}$ , unless we refer to a specific set. Note that there are other RIS technologies and/or more practical feasibility sets for STAR/reflective RIS as mentioned in [3], [49], which should be considered in future studies.

### B. Signal Model

We assume that BS  $l$  broadcasts the signal

$$\mathbf{x}_l = \sum_{k=1}^{K_l} \mathbf{x}_{lk} \in \mathbb{C}^{N_{BS,l} \times 1}, \quad (10)$$

where  $\mathbf{x}_{lk}$  is the signal intended for user  $U_{lk}$ , which is a zero-mean complex Gaussian random vector with covariance  $\mathbf{P}_{lk} = \mathbb{E}\{\mathbf{x}_{lk}\mathbf{x}_{lk}^H\}$ , where  $\mathbb{E}\{\mathbf{x}\}$  is the mathematical expectation of  $\mathbf{x}$ . We assume that the zero-mean signals  $\mathbf{x}_{lk}$  are independent from each other, i.e.,  $\mathbb{E}\{\mathbf{x}_{lk}\mathbf{x}_{ij}^H\} = \mathbf{0}$  for  $i \neq l$  and/or  $j \neq k$ . Additionally, we denote the covariance matrix of  $\mathbf{x}_l$  by  $\mathbf{P}_l = \mathbb{E}\{\mathbf{x}_l\mathbf{x}_l^H\}$ . Since the signals  $\mathbf{x}_{lk}$  are zero-mean and independent random vectors for all  $l$  and  $k$ , we have  $\mathbf{P}_l = \sum_k \mathbf{P}_{lk}$ . The set containing all the feasible transmit covariance matrices is denoted as  $\mathcal{P}$  and it is given by

$$\mathcal{P} = \{\mathbf{P}_{lk} : \text{Tr}(\mathbf{P}_l) \leq p_l, \mathbf{P}_{lk} \succeq \mathbf{0}, \forall l, k\}, \quad (11)$$

where  $p_l$  is the power budget of the  $l$ -th BS.

The received signal at  $U_{lk}$  is given by

$$\begin{aligned} \mathbf{y}_{lk} &= \sum_i \mathbf{H}_{lk,i} \mathbf{x}_i + \mathbf{n}_{lk}, \\ &= \underbrace{\mathbf{H}_{lk,l} \mathbf{x}_l}_{\text{Desired Signal}} + \underbrace{\sum_{j \neq k} \mathbf{H}_{lk,l} \mathbf{x}_{lj}}_{\text{Intracell Interference}} + \underbrace{\sum_{i \neq l} \mathbf{H}_{lk,i} \mathbf{x}_i}_{\text{Intercell Interference}} + \underbrace{\mathbf{n}_{lk}}_{\text{Noise}}, \end{aligned} \quad (12)$$

where  $\mathbf{n}_{lk}$  is the zero-mean additive white Gaussian noise at  $U_{lk}$  with covariance matrix  $\sigma^2 \mathbf{I}$ , where  $\mathbf{I}$  denotes the identity matrix. Note that, in (12), the differences between the intercell and intracell links are carefully taken into account, and each user is indeed affected by both *intercell* and *intracell* interference. In this paper, we treat interference as noise, which is optimal for maximizing the sum rate [50] or the generalized degree of freedom [51] when the interference is weak. An alternative strategy for treating interference as noise (TIN) is to detect and cancel interference, which is known as

successive interference cancellation (SIC), and it is optimal when the interference is strong. SIC requires more advanced user devices as well as more sophisticated signaling design. Moreover, to detect and cancel interference at the users takes some time, which may lead to violating the latency constraint in URLLC-related applications.

Note that we employ Gaussian signaling in this work similar to most studies of wireless communication systems both with and without RISs. In practice, typically discrete constellations are employed. Studies based on Gaussian signaling are nevertheless important since they provide valuable insights into the system performance and represent an upper bound for the performance of the technologies studied. Additionally, there are studies on the comparison of Gaussian signals and discrete constellations, e.g., [52]–[54]. The performance gap between the discrete constellations and Gaussian signals grows as the number of bits/symbols in the discrete constellations increases, but it eventually saturates. To account for this performance gap and the idealized assumption, one can employ a signal-to-noise ratio (SNR) offset.

### C. Channel Dispersion, Rate and EE Expressions

A MIMO channel can be modeled as a set of parallel AWGN channels, and [27, Theorem 78] can be employed to obtain the achievable rate of MIMO channels associated with FBL coding. Note that [27, Theorem 78] is based on the optimal power allocation for a point-to-point MIMO communication link; however, the FBL rate expressions can be formulated for any arbitrary power allocation as shown in [27, Section 4.5.4]. In the following lemma, we calculate the achievable FBL rate of users, when the interference is treated as noise for decoding the corresponding signal at the receivers.

**Lemma 1** ([27]). *The second-order rate of user  $U_{lk}$  for FBL coding along with the normal approximation (NA) is given by*

$$r_{lk} = \underbrace{\log |\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk}|}_{\text{Shannon Rate}} - Q^{-1}(\epsilon) \sqrt{\frac{V_{lk}}{n_t}}, \quad (13)$$

where  $n_t$  is the packet length,  $\mathbf{S}_{lk} = \mathbf{H}_{lk,l} \mathbf{P}_{lk} \mathbf{H}_{lk,l}^H$  is the covariance matrix of the desired signal at the user  $U_{lk}$ , while  $\mathbf{D}_{lk}$  is the covariance matrix of the interfering signals plus noise, given by

$$\mathbf{D}_{lk} = \sigma^2 \mathbf{I} + \sum_{i=1, i \neq l}^L \mathbf{H}_{lk,i} \mathbf{P}_i \mathbf{H}_{lk,i}^H + \sum_{j=1, j \neq k}^K \mathbf{H}_{lk,l} \mathbf{P}_{lj} \mathbf{H}_{lk,l}^H. \quad (14)$$

Here, the first-order Shannon rate can also be written as

$$C_{lk} = \log |\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk}| = \log |\mathbf{I} + \mathbf{\Lambda}_{lk}| = \sum_{i=1}^I \log(1 + \lambda_{lki}), \quad (15)$$

where  $\mathbf{\Lambda}_{lk} = \text{diag}(\lambda_{lk1}, \lambda_{lk2}, \dots, \lambda_{lkI})$  is a diagonal matrix, containing the non-zero eigenvalues of the positive semidefinite (PSD) matrix  $\mathbf{D}_{lk}^{-1} \mathbf{S}_{lk}$ , and  $I \leq \min(N_{BS,l}, N_{u,lk})$  is equal to the rank of  $\mathbf{H}_{lk,l} \mathbf{P}_{lk} \mathbf{H}_{lk,l}^H$ , which also represents the number of parallel channels. The parameter  $\lambda_{lki}$  is actually the signal-to-interference-plus-noise ratio (SINR) at the  $i$ -th parallel channel of user  $U_{lk}$ . Finally,  $V_{lk} = \sum_{i=1}^I V_{lki}$  is the

channel dispersion of  $U_{lk}$ , where  $V_{lki}$  is the channel dispersion of the  $i$ -th parallel channel of user  $U_{lk}$ .

The optimal channel dispersion of the  $i$ -th parallel channel is given by [34]

$$V_{lki} = 1 - \frac{1}{(1 + \lambda_{lki})^2}, \quad (16)$$

where  $\lambda_{lki}$  is given in Lemma 1. Unfortunately, the optimal channel dispersion attains the minimum value of  $V_{lki}$  for all  $l$ ,  $k$ , and  $i$ , but Gaussian signals cannot achieve it in the presence of interference. In [34], a coding scheme was proposed for independent, identically distributed (iid) Gaussian signals in interference channels, which has the following channel dispersion

$$V_{lki} = 2 \left( 1 - \frac{1}{1 + \lambda_{lki}} \right). \quad (17)$$

In the following lemma, we present closed-form matrix expressions for the optimal channel dispersion and the achievable channel dispersion in (13).

**Lemma 2.** *The optimal channel dispersion can be written as*

$$V_{lk} = \text{Tr}(\mathbf{I} - (\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk})^{-2}). \quad (18)$$

*Additionally, the achievable channel dispersion for the scheme proposed in [34] can be written in the following matrix format*

$$V_{lk} = 2\text{Tr}(\mathbf{I} - (\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk})^{-1}) \quad (19)$$

$$= 2\text{Tr}(\mathbf{I} - \mathbf{D}_{lk}(\mathbf{D}_{lk} + \mathbf{S}_{lk})^{-1}) \quad (20)$$

$$= 2\text{Tr}(\mathbf{S}_{lk}(\mathbf{D}_{lk} + \mathbf{S}_{lk})^{-1}). \quad (21)$$

*Proof.* It is widely exploited that the trace of a positive semi-definite matrix is equal to the summation of its eigenvalues. It can be readily verified that the non-zero eigenvalues of  $(\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk})^{-2}$  are equal to  $(1 + \lambda_{lki})^{-2}$ ,  $i \in \{1, 2, \dots, I\}$ , which proves the equality in (18). Note that if  $\mathbf{X}$  is a positive semi-definite matrix with non-zero eigenvalues  $\lambda_i$ , then its pseudo-inverse, denoted as  $\mathbf{X}^{-1}$ , is also a positive semi-definite matrix, and its non-zero eigenvalues are  $\lambda_i^{-1}$ .

Similarly, it can be readily verified that the eigenvalues of  $(\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk})^{-1}$  are equal to  $(1 + \lambda_{lki})^{-1}$ ,  $i \in \{1, 2, \dots, I\}$ , which yields (19). Employing a simple matrix factorization, it can be easily verified that (19), (20), and (21) are equivalent.  $\square$

**Remark 1.** *The reliability constraint is modeled by utilizing the maximum tolerable error rate,  $\epsilon$ , in the FBL rates. Moreover, the coding length  $n_t$  should be proportional to the tolerable latency. Indeed, a more stringent latency constraint leads to employing a shorter block length [55], [56]. Additionally, the latency constraint can be modeled as a constraint on the minimum rate as discussed in [57, Remark 1], [15, Sec. II.D], [58].*

The EE of  $U_{lk}$  is defined as [59]

$$e_{lk} = \frac{r_{lk}}{P_c + \eta \text{Tr}(\mathbf{P}_{lk})}, \quad (22)$$

where  $\eta^{-1}$  is the power efficiency of the transmit devices at the BSs,  $P_c$  is the constant power consumption of the system

(including the devices of the BSs, RISs and  $U_{lk}$ ) to transmit data to a user, which is given by [28, Eq. (27)], and  $r_{lk}$  is given by Lemma 1. Note that to compute  $P_c$ , the constant power of the devices of the BSs and RISs is normalized by the number of users served. Moreover, the global EE (GEE) of the network is defined as [59]

$$g = \frac{\sum_{lk} r_{lk}}{LK P_c + \eta \sum_l \text{Tr}(\mathbf{P}_l)}, \quad (23)$$

which quantifies how energy efficient the network is. Finally, the transmission delay of  $U_{lk}$  upon transmitting a packet with length  $n_t$  is  $d_{lk} = \frac{n_t}{r_{lk}}$ .

#### D. Problem Statement

We consider a general optimization problem for URLLC systems formulated as follows

$$\max_{\{\mathbf{P}\} \in \mathcal{P}, \{\Psi\} \in \mathcal{T}} f_0(\{\mathbf{P}\}, \{\Psi\}) \quad (24a)$$

$$\text{s.t. } f_i(\{\mathbf{P}\}, \{\Psi\}) \geq 0, \forall i, \quad (24b)$$

$$r_{lk} \geq r^{th}, \forall l, k, \quad (24c)$$

where constraint (24c) can be interpreted as a latency constraint for each user, as discussed in [14, Sec. II.D]. Moreover, functions  $f_i$ ,  $\forall i$  are, in general, non-linear functions of the optimization variables. These functions can be, but are not restricted to, linear functions of the rates/EEs and/or transmit/receive powers. For instance,  $f_i$  can be a function of the sum rate, minimum rate/EE, transmission/receive power, interference temperature at a user, transmission delay and so on. Additionally,  $f_i$  can be a non-linear function of the rates/EEs such as the geometric mean of the rates as in [60]. Note that (24) may also include minimization problems such as the total power minimization subject to a given target rate, maximum transmission delay minimization, and interference temperature minimization. In this case,  $f_i$  can be chosen as, e.g.,  $-\sum_l \text{Tr}(\mathbf{P}_l)$  or  $-\max_{lk} \{d_{lk}\}$ . Therefore, the general problem in (24) can include an extensive range of optimization scenarios, encompassing the maximization of the minimum weighted rate, sum rate, global EE and minimum EE. We refer the reader to [36, Sec. II.B] for more discussions on the format of the functions  $f_i$ s as well as of the family of optimization problems that can be formulated as (24).

### III. PROPOSED OPTIMIZATION FRAMEWORK

In this section, we propose iterative schemes for solving (24) by leveraging AO, MM-based, and FP algorithms. Specifically, we first fix the RIS coefficients to  $\{\Psi^{(t-1)}\}$  and update the transmit covariance matrices as  $\{\mathbf{P}^{(t)}\}$  by solving (24). We then alternate and update the RIS coefficients, while  $\{\mathbf{P}\}$  is fixed to  $\{\mathbf{P}^{(t)}\}$ . We iterate this procedure until convergence is reached. Unfortunately, the optimization problems are non-convex and complicated even when the RIS elements (or covariance matrices) are fixed. Thus, we propose a suboptimal scheme based on MM to solve the corresponding problems. Below, we present our solutions for updating the transmit covariance matrices and RIS elements in separate subsections.

### A. Updating Transmit Covariance Matrices

To update  $\{\mathbf{P}\}$ , we introduce a new set of variables  $\{\mathbf{Q}\} = \{\mathbf{Q}_{lk}\}_{\forall lk}$ , where  $\mathbf{Q}_{lk}$  is a positive semi-definite matrix and  $\mathbf{P}_{lk} = \mathbf{Q}_{lk} \mathbf{Q}_{lk}^H$ . Equivalently, we can compute  $\mathbf{Q}_{lk}$  as  $\mathbf{P}_{lk}^{1/2}$ . To attain a suboptimal solution for (24), we leverage an MM-based technique. More specifically, we first obtain suitable concave surrogate functions for the rates. Then, we update  $\mathbf{P}_{lk}^{(t)} = \mathbf{Q}_{lk}^{(t)} \mathbf{Q}_{lk}^{(t)H}$ ,  $\forall l, k$ , by solving the corresponding surrogate optimization problems. To derive concave lower bounds for the FBL rates, we utilize the bounds in the following lemmas.

**Lemma 3.** Consider the arbitrary matrices  $\mathbf{\Lambda}$ ,  $\bar{\mathbf{\Lambda}}$  and positive semi-definite matrices  $\mathbf{\Upsilon}$ ,  $\bar{\mathbf{\Upsilon}}$ . Then, the following inequality holds for all feasible  $\mathbf{\Lambda}$ ,  $\bar{\mathbf{\Lambda}}$ ,  $\mathbf{\Upsilon}$ , and  $\bar{\mathbf{\Upsilon}}$ :

$$f(\mathbf{\Lambda}, \mathbf{\Upsilon}) = \text{Tr}(\mathbf{\Upsilon}^{-1} \mathbf{\Lambda} \mathbf{\Lambda}^H) \geq 2\Re\{\text{Tr}(\bar{\mathbf{\Upsilon}}^{-1} \bar{\mathbf{\Lambda}} \mathbf{\Lambda}^H)\} - \text{Tr}(\bar{\mathbf{\Upsilon}}^{-1} \bar{\mathbf{\Lambda}} \bar{\mathbf{\Lambda}}^H \bar{\mathbf{\Upsilon}}^{-1} \mathbf{\Upsilon}), \quad (25)$$

where  $\Re\{x\}$  returns the real value of  $x$ .

*Proof.* Function  $f(\mathbf{\Lambda}, \mathbf{\Upsilon})$  is jointly convex in  $\mathbf{\Lambda}$  and  $\mathbf{\Upsilon}$  [61]. Thus, we can employ the first-order Taylor expansion to obtain an affine lower-bound for  $f(\cdot)$  as follows

$$f(\mathbf{\Lambda}, \mathbf{\Upsilon}) \geq f(\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Upsilon}}) + 2\Re\left\{\frac{\partial f(\mathbf{\Lambda}, \mathbf{\Upsilon})}{\partial \mathbf{\Lambda}}\bigg|_{\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Upsilon}}}(\mathbf{\Lambda} - \bar{\mathbf{\Lambda}}) + \frac{\partial f(\mathbf{\Lambda}, \mathbf{\Upsilon})}{\partial \mathbf{\Upsilon}}\bigg|_{\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Upsilon}}}(\mathbf{\Upsilon} - \bar{\mathbf{\Upsilon}})\right\}, \quad (26)$$

where  $\bar{\mathbf{\Lambda}}$  and  $\bar{\mathbf{\Upsilon}}$  are any arbitrary feasible points, and  $\frac{\partial f(\mathbf{\Lambda}, \mathbf{\Upsilon})}{\partial \mathbf{\Lambda}}\big|_{\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Upsilon}}}$  (or  $\frac{\partial f(\mathbf{\Lambda}, \mathbf{\Upsilon})}{\partial \mathbf{\Upsilon}}\big|_{\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Upsilon}}}$ ) is the derivative of  $f(\cdot)$  with respect to  $\mathbf{\Lambda}$  (or  $\mathbf{\Upsilon}$ ) at  $\bar{\mathbf{\Lambda}}$  and  $\bar{\mathbf{\Upsilon}}$ . Replacing the corresponding derivatives in (26) and simplifying the equation results in (25).  $\square$

**Lemma 4** ([28]). Consider the arbitrary matrices  $\mathbf{\Lambda}$  and  $\bar{\mathbf{\Lambda}}$ , and positive definite matrices  $\mathbf{\Upsilon}$  and  $\bar{\mathbf{\Upsilon}}$ . Then, we have:

$$\ln|\mathbf{I} + \mathbf{\Upsilon}^{-1} \mathbf{\Lambda} \mathbf{\Lambda}^H| \geq \ln|\mathbf{I} + \bar{\mathbf{\Upsilon}}^{-1} \bar{\mathbf{\Lambda}} \bar{\mathbf{\Lambda}}^H| - \text{Tr}(\bar{\mathbf{\Upsilon}}^{-1} \bar{\mathbf{\Lambda}} \bar{\mathbf{\Lambda}}^H) + 2\Re\{\text{Tr}(\bar{\mathbf{\Upsilon}}^{-1} \bar{\mathbf{\Lambda}} \mathbf{\Lambda}^H)\} - \text{Tr}((\bar{\mathbf{\Upsilon}}^{-1} - (\bar{\mathbf{\Lambda}} \bar{\mathbf{\Lambda}}^H + \bar{\mathbf{\Upsilon}})^{-1})^H (\mathbf{\Lambda} \mathbf{\Lambda}^H + \mathbf{\Upsilon})). \quad (27)$$

Upon employing the concave lower bounds in Lemma 3 and Lemma 4, we can obtain a concave lower bound for the FBL rates with the NA approximation as presented in the following lemma.

**Lemma 5.** A concave lower bound for  $r_{lk}$  is given by

$$r_{lk} \geq \tilde{r}_{lk} = a_{lk} + 2 \sum_{ij} \Re\{\text{Tr}(\mathbf{A}_{lk,ij} \mathbf{Q}_{ij}^H \bar{\mathbf{H}}_{lk,i}^H)\} - \text{Tr}(\mathbf{B}_{lk}(\mathbf{H}_{lk,l} \mathbf{Q}_{lk} \mathbf{Q}_{lk}^H \mathbf{H}_{lk,l}^H + \mathbf{D}_{lk})) \quad (28)$$

where

$$a_{lk} = \ln|\mathbf{I} + \bar{\mathbf{D}}_{lk}^{-1} \bar{\mathbf{S}}_{lk}| - \text{Tr}(\bar{\mathbf{D}}_{lk}^{-1} \bar{\mathbf{S}}_{lk}) - \frac{Q^{-1}(\epsilon)(\bar{V}_{lk} + 2I)}{2\sqrt{n_t \bar{V}_{lk}}},$$

$$\mathbf{A}_{lk,ij} = \begin{cases} \bar{\mathbf{D}}_{lk}^{-1} \bar{\mathbf{H}}_{lk,l} \bar{\mathbf{Q}}_{lk} & \text{if } i = l, j = k, \\ \frac{Q^{-1}(\epsilon)}{\sqrt{n_t \bar{V}_{lk}}} (\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1} \bar{\mathbf{H}}_{lk,i} \bar{\mathbf{Q}}_{ij} & \text{otherwise,} \end{cases}$$

$$\mathbf{B}_{lk} = \bar{\mathbf{D}}_{lk}^{-1} - (\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1} + \frac{Q^{-1}(\epsilon)}{\sqrt{n_t \bar{V}_{lk}}} (\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1} \bar{\mathbf{D}}_{lk} (\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1},$$

where  $\bar{\mathbf{D}}_{lk}$ ,  $\bar{\mathbf{S}}_{lk}$ ,  $\bar{\mathbf{Q}}_{ij}$ ,  $\bar{V}_{lk}$ , and  $\bar{\mathbf{H}}_{lk,i}$ ,  $\forall l, k, i, j$  are, respectively, the initial values of  $\mathbf{D}_{lk}$ ,  $\mathbf{S}_{lk}$ ,  $\mathbf{Q}_{ij}$ ,  $V_{lk}$ , and  $\mathbf{H}_{lk,i}$  at the current step, which are obtained upon replacing  $\{\mathbf{P}\}$  by  $\{\mathbf{P}^{(t-1)}\}$  and  $\{\Psi\}$  by  $\{\Psi^{(t-1)}\}$ .

*Proof.* Upon employing Lemma 4, a concave lower bound can be obtained for the first-order Shannon rate as

$$|\mathbf{I} + \mathbf{D}_{lk}^{-1} \mathbf{S}_{lk}| \geq \ln|\mathbf{I} + \bar{\mathbf{D}}_{lk}^{-1} \bar{\mathbf{S}}_{lk}| - \text{Tr}(\bar{\mathbf{D}}_{lk}^{-1} \bar{\mathbf{S}}_{lk}) + 2\Re\{\text{Tr}(\bar{\mathbf{Q}}_{lk}^H \bar{\mathbf{D}}_{lk}^{-1} \mathbf{Q}_{lk})\} - \text{Tr}((\bar{\mathbf{D}}_{lk}^{-1} - (\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1})^H (\bar{\mathbf{H}}_{lk,l} \mathbf{Q}_{lk} \mathbf{Q}_{lk}^H \bar{\mathbf{H}}_{lk,l}^H + \mathbf{D}_{lk})). \quad (29)$$

Next, we obtain a concave lower bound for  $-Q^{-1}(\epsilon)\sqrt{\frac{V_{lk}}{n_t}}$ , which is equivalent to obtaining a convex upper bound for  $\sqrt{V_{lk}}$ . To this end, we first employ the following inequality

$$\sqrt{V_{lk}} \leq \frac{\sqrt{\bar{V}_{lk}}}{2} + \frac{V_{lk}}{2\sqrt{\bar{V}_{lk}}}, \quad (30)$$

which is non-convex since  $V_{lk}$  is not convex in  $\{\mathbf{Q}\}$ . Upon employing Lemma 3, a convex upper bound for  $V_{lk}(\cdot)$  can be obtained as

$$V_{lk} \leq 2\text{Tr}(\mathbf{I}) - 4 \sum_{[ij] \neq [lk]} \Re\{\text{Tr}(\mathbf{A}_{lk,ij} \mathbf{Q}_{ij}^H \bar{\mathbf{H}}_{lk,i}^H)\} + 2\text{Tr}(((\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1})^{-1} \bar{\mathbf{D}}_{lk} (\bar{\mathbf{S}}_{lk} + \bar{\mathbf{D}}_{lk})^{-1})^{-1} \times (\mathbf{H}_{lk,l} \mathbf{Q}_{lk} \mathbf{Q}_{lk}^H \mathbf{H}_{lk,l}^H + \mathbf{D}_{lk})), \quad (31)$$

where  $[ij] \neq [lk]$  includes all possible  $i, j$  pairs, except for the case where  $i = l$  and simultaneously  $j = k$ . Substituting the concave lower bound in (29) and the convex upper bound in (31) into the FBL rate expression proves the lemma.  $\square$

**Remark 2.** The concave lower bound in Lemma 4 is quadratic in  $\{\mathbf{Q}\}$ , and consists of a constant term, a linear/affine term, and a quadratic term.

We denote the surrogate functions for  $f_i$  by  $\tilde{f}_i$ , which are obtained by substituting the concave lower bounds  $\tilde{r}_{lk}$  in (24). For instance, if  $f_i$  is equal to the sum rate, then  $\tilde{f}_i = \sum_{\forall lk} \tilde{r}_{lk}$ . Moreover, if  $f_i$  represents the EE of  $U_{lk}$ , then we have

$$\tilde{f}_i = \tilde{e}_{lk} = \frac{\tilde{r}_{lk}}{P_c + \eta \text{Tr}(\mathbf{Q}_{lk} \mathbf{Q}_{lk}^H)}, \quad (32)$$

which is a fractional function of  $\{\mathbf{Q}\}$  with a concave numerator and convex denominator. Moreover, if  $f_i$  is a function of the transmission delay, then we have  $\tilde{f}_i = -\frac{n_t}{\tilde{r}_{lk}}$ . Note that although the surrogate lower bounds for the rates in Lemma 5 are concave, the  $\tilde{f}_i$ s are not necessarily concave, since they might be a linear function of the EE metrics. Substituting the  $f_i$ s by the  $\tilde{f}_i$ s leads to

$$\max_{\{\mathbf{Q}\}} \tilde{f}_0(\{\mathbf{Q}\}, \{\Psi^{(t-1)}\}) \quad (33a)$$

$$\text{s.t. } \tilde{f}_i(\{\mathbf{Q}\}, \{\Psi^{(t-1)}\}) \geq 0, \forall i, \quad (33b)$$

$$\tilde{r}_{lk} \geq r^{th}, \quad \forall l, k, \quad (33c)$$

$$\sum_k \text{Tr}(\mathbf{Q}_{lk} \mathbf{Q}_{lk}^H) \leq p_l, \forall l. \quad (33d)$$

The optimization problem (33) is convex for the maximization of the minimum and/or sum rates. Hence, it can be efficiently solved by existing numerical tools. Note that our framework can also optimize other SE metrics, such as the geometric mean of users. As shown in [60], the maximization of the geometric mean of the users can be solved by solving a sequence of weighted sum rate maximization problems, which can be efficiently handled by our framework.

Unfortunately, (33) is non-convex for GEE maximization as well as for the maximization of the minimum weighted EE of the users, since the EE and/or GEE functions are not concave in  $\{\mathbf{Q}\}$ . Fortunately, a solution of the minimum weighted EE of the users and/or GEE maximization problems can be obtained by Dinkelbach-based algorithms, since the numerator of  $\tilde{e}_{lk}$  (or  $\tilde{g}$ ) is concave, while its denominator is convex. The problem in (33) is not convex, when minimizing the maximum transmission delay of the users, which can be considered as a latency metric from a physical layer point of view. In the following, we solve (33) for the maximization of the minimum weighted EE of the users, the maximization of the GEE and the minimization of the maximum transmission delay.

1) *Maximization of the Minimum EE:* In this case, (33) can be written as

$$\max_{\{\mathbf{Q}\}, e} \quad e, \quad \text{s.t.} \quad \tilde{e}_{lk} = \frac{\tilde{r}_{lk}}{P_c + \eta \text{Tr}(\mathbf{Q}_{lk} \mathbf{Q}_{lk}^H)} \geq e, \quad \forall i, \quad (34a)$$

$$(33c), (33d). \quad (34b)$$

Upon employing the generalized Dinkelbach algorithm (GDA), we can derive the globally optimal solution of (34) by iteratively solving the convex optimization problem [59]

$$\max_{\{\mathbf{Q}\}, e} \quad e, \quad \text{s.t.} \quad \tilde{r}_{lk} - \mu^{(n)} (P_c + \eta \text{Tr}(\mathbf{Q}_{lk} \mathbf{Q}_{lk}^H)) \geq e, \quad \forall i, \quad (35a)$$

$$(33c), (33d), \quad (35b)$$

and updating  $\mu^{(n)}$  as

$$\mu^{(n)} = \min_{lk} \left\{ \tilde{e}_{lk}^{(n-1)} \right\} = \min_{lk} \left\{ \frac{\tilde{r}_{lk} \left( \mathbf{Q}_{lk}^{(n-1)} \right)}{P_c + \eta \text{Tr} \left( \mathbf{Q}_{lk}^{(n-1)} \mathbf{Q}_{lk}^{(n-1)H} \right)} \right\}. \quad (36)$$

Note that  $n$  is the number of iterations in the inner loop, i.e., the number of GDA iterations.

2) *Maximization of the GEE:* In this case, (33) is equivalent to

$$\max_{\{\mathbf{Q}\}} \frac{\sum_{l,k} \tilde{r}_{lk}}{\sum_{l,k} (P_c + \eta \text{Tr}(\mathbf{Q}_{lk} \mathbf{Q}_{lk}^H))} \quad \text{s.t.} \quad (33c), (33d). \quad (37)$$

Employing the Dinkelbach algorithm, a globally optimal solution of (37) can be found by iteratively solving [59]

$$\max_{\{\mathbf{Q}\}} \sum_{l,k} \tilde{r}_{lk} - \mu^{(n)} \sum_{l,k} (P_c + \eta \text{Tr}(\mathbf{Q}_{lk} \mathbf{Q}_{lk}^H)) \quad \text{s.t.} \quad (33c), (33d),$$

and updating  $\mu^{(n)}$  as

$$\mu^{(n)} = \tilde{g}_{lk}^{(n-1)} = \frac{\sum_{l,k} \tilde{r}_{lk} \left( \mathbf{Q}_{lk}^{(n-1)} \right)}{\sum_{l,k} (P_c + \eta \text{Tr}(\mathbf{Q}_{lk}^{(n-1)} \mathbf{Q}_{lk}^{(n-1)H}))}.$$

3) *Minimization of the maximum transmission delay:* To minimize the maximum delay, we have to solve

$$\min_{\{\mathbf{Q}\}} \max_{\forall lk} \left\{ \frac{n_t}{\tilde{r}_{lk}} \right\} \quad \text{s.t.} \quad (33c), (33d), \quad (38)$$

which is equivalent to maximizing the minimum rate of users as

$$\max_{\{\mathbf{Q}\}} \min_{\forall lk} \{ \tilde{r}_{lk} \} \quad \text{s.t.} \quad (33c), (33d), \quad (39)$$

which is a convex optimization problem. Note that the transmission delay  $d_{lk}$  is a monotonically decreasing function of  $r_{lk}$ , and it is minimized when  $r_{lk}$  is maximized, which makes the solution of (38) equivalent to (39).

4) *Discussion on Single-stream Data Transmission:* In this case, the rank of matrix  $\mathbf{P}_{lk}$  for all  $l, k$  is equal to one. This means that matrix  $\mathbf{P}_{lk}$  can be written as  $\mathbf{P}_{lk} = \mathbf{q}_{lk} \mathbf{q}_{lk}^H$ , where  $\mathbf{q}_{lk} \in \mathbb{C}^{N_{BS,l} \times 1}$ . In other words, when single-stream data transmission is employed, the BSs perform only beamforming to transmit data, and we have to optimize the beamforming vectors instead of transmit covariance matrices. Therefore, the computational complexity of single-stream data transmission is lower than that of employing multiple streams. However, this lower computational complexity is attained at the cost of a significant performance loss, especially when the maximum number of streams, i.e.,  $\min(N_{BS,l}, N_{u,lk})$  for  $\mathbf{U}_{lk}$ , increases. Indeed, as the network size increases, more advanced transmission and resource allocation techniques, involving higher complexities, are needed to provide satisfactory performance.

Note that if the transmitter and/or the receiver are equipped with only a single antenna, we are restricted to single-stream data transmissions. Obviously, the single-stream scheme becomes increasingly suboptimal, when the maximum number of streams grows. To derive single-stream data transmission solutions, we only have to replace matrices  $\mathbf{Q}_{lk}$  by vectors  $\mathbf{q}_{lk}$  and employ the lower bounds in Lemma 5. Indeed, the schemes proposed in this subsection can be applied for both single- and multiple-stream data transmission.

## B. Optimizing the RIS Elements

Now, we update  $\{\Psi\}$  by solving (24) for fixed  $\{\mathbf{P}^{(t)}\}$ , i.e.,

$$\max_{\{\Psi\} \in \mathcal{T}} f_0 \left( \{\mathbf{P}^{(t)}\}, \{\Psi\} \right) \quad (40a)$$

$$\text{s.t.} \quad f_i \left( \{\mathbf{P}^{(t)}\}, \{\Psi\} \right) \geq 0, \quad \forall i, \quad (40b)$$

$$r_{lk} \geq r^{th}, \quad \forall l, k, \quad (40c)$$

which is non-convex since the rates are not concave in  $\{\Psi\}$  and the set  $\mathcal{T}$  might be non-convex. Note that as discussed in Section III-A, the minimization of the maximum transmission delay is equivalent to the maximization of the minimum rate. Hence, our focus in this subsection is on the SE and EE

metrics. In the following, we first consider reflective RISs and then, describe how the solution can be applied to the more sophisticated STAR-RIS, utilizing the MS scheme.

1) *Reflective RISs*: To find a suboptimal solution for (40), we leverage an approach based on MM. Specifically, we first obtain a suitable concave lower bound for the rates and then, convexify  $\mathcal{T}$  if it is not already a convex set. Since the rates have similar structures in  $\{\mathbf{Q}\}$  and in  $\{\Psi\}$ , we can utilize the concave lower bounds in Lemma 5 to attain concave lower bounds for the rates and construct suitable surrogate optimization problems for updating  $\{\Psi\}$ . In the subsequent corollary, we provide the concave lower bounds for the rates. The proof of this corollary closely resembles that of Lemma 5, and thus, we omit it here.

**Corollary 1.** A concave lower bound for  $r_{lk}$  is given by

$$r_{lk} \geq \hat{r}_{lk} = a_{lk} + 2 \sum_{ij} \Re \left\{ \text{Tr} \left( \mathbf{A}_{lk,ij} \mathbf{Q}_{ij}^{(t)H} \mathbf{H}_{lk,i}^H \right) \right\} - \text{Tr} \left( \mathbf{B}_{lk} (\mathbf{H}_{lk,l} \mathbf{Q}_{lk}^{(t)} (\mathbf{H}_{lk,l} \mathbf{Q}_{lk}^{(t)})^H + \mathbf{D}_{lk}) \right), \quad (41)$$

where the constant parameters  $a_{lk}$ ,  $\mathbf{A}_{lk,ij}$ , and  $\mathbf{B}_{lk}$  are defined as in Lemma 5.

Substituting the surrogate functions for the rates, i.e., the  $\hat{r}_{lk}$ s, in (40) yields

$$\max_{\{\Psi\} \in \mathcal{T}} \hat{f}_0 \left( \{\mathbf{P}^{(t)}\}, \{\Psi\} \right) \quad (42a)$$

$$\text{s.t. } \hat{f}_i \left( \{\mathbf{P}^{(t)}\}, \{\Psi\} \right) \geq 0, \forall i, \quad (42b)$$

$$\hat{r}_{lk} \left( \{\mathbf{P}^{(t)}\}, \{\Psi\} \right) \geq r^{th}, \forall l, k, \quad (42c)$$

which is convex if  $\mathcal{T}$  is a convex set, i.e. when  $\mathcal{T}_U$  is considered. For  $\mathcal{T}_U$ , the proposed scheme achieves convergence to a stationary point of (24). Note that the surrogate functions  $\hat{f}_i(\{\mathbf{P}^{(t)}\}, \{\Psi\})$  are concave in  $\{\Psi\}$  even if they are linear functions of the EE metrics. The reason is that the powers (transmit covariance matrices) are fixed, and thus, the EE metrics are not fractional functions of  $\{\Psi\}$ .

Now, we convexify  $\mathcal{T}_I$ . The unit modulus constraint  $|\psi_{m_n}| = 1$  is equivalent to

$$|\psi_{m_n}|^2 \leq 1, \quad (43)$$

$$|\psi_{m_n}|^2 \geq 1. \quad (44)$$

The constraint  $|\psi_{m_n}|^2 \leq 1$  is convex; however,  $|\psi_{m_n}|^2 \geq 1$  is not, which makes (42) a non-convex problem. Thus, we have to approximate (44) with a convex constraint to make (42) convex. To this end, we employ the convex-concave procedure (CCP) and rewrite (44) as

$$|\psi_{m_n}|^2 \geq |\psi_{m_n}^{(t-1)}|^2 - 2\Re\{\psi_{m_n}^{(t-1)*}(\psi_{m_n} - \psi_{m_n}^{(t-1)})\} \geq 1. \quad (45)$$

To avoid potential numerical errors and speed up the convergence, we relax (45) as

$$|\psi_{m_n}^{(t-1)}|^2 - 2\Re\{\psi_{m_n}^{(t-1)*}(\psi_{m_n} - \psi_{m_n}^{(t-1)})\} \geq 1 - \delta, \quad (46)$$

where  $\delta > 0$ . Now, we can approximate (42) as

$$\max_{\{\Psi\}} \hat{f}_0 \left( \{\mathbf{P}^{(t)}\}, \{\Psi\} \right) \text{ s.t. } (42b), (42c), \quad (47a)$$

$$(43), (46), \forall m, n, \quad (47b)$$

which is convex. We denote the solution of (47) as  $\{\Psi^{(*)}\} = \{\Psi_1^{(*)}, \Psi_2^{(*)}, \dots, \Psi_M^{(*)}\}$ . Because of the relaxation in (46), it may happen that  $\psi_{m_n}^{(*)}$ , i.e., the  $n$ -th coefficient of the diagonal matrix  $\Psi_m^{(*)}$ , does not satisfy  $|\psi_{m_n}| = 1$ . Therefore, we normalize  $\{\Psi^{(*)}\}$  as

$$\hat{\psi}_{m_n} = \frac{\psi_{m_n}^{(*)}}{|\psi_{m_n}^{(*)}|}, \quad \forall m, n. \quad (48)$$

To guarantee the convergence, we update  $\{\Psi\}$  as

$$\{\Psi^{(t)}\} = \begin{cases} \{\hat{\Psi}\} & \text{if } f_0 \left( \{\mathbf{P}^{(t)}\}, \{\hat{\Psi}\} \right) \geq f_0 \left( \{\mathbf{P}^{(t)}\}, \{\Psi^{(t-1)}\} \right) \\ \{\Psi^{(t-1)}\} & \text{otherwise.} \end{cases} \quad (49)$$

For  $\mathcal{T} = \mathcal{T}_I$ , our proposed framework converges since a non-decreasing sequence of objective functions (OF)  $f_0$  is generated. For  $\mathcal{T} = \mathcal{T}_U$ , our framework converges to a stationary point of (24) because  $\mathcal{T}_U$  is convex. We summarize our algorithm for maximizing the minimum EE with  $\mathcal{T}_U$  in Algorithm I.

---

**Algorithm I** Maximization of the minimum EE for  $\mathcal{T}_U$ .

---

**Initialization**

Set  $\gamma_1, \gamma_2, t = 1, \{\mathbf{P}\} = \{\mathbf{P}^{(0)}\}$ , and  $\{\Psi\} = \{\Psi^{(0)}\}$

---

**While**  $\left( \min_{\forall lk} e_{lk}^{(t)} - \min_{\forall lk} e_{lk}^{(t-1)} \right) / \min_{\forall lk} e_{lk}^{(t-1)} \geq \gamma_1$

**Optimizing over  $\{\mathbf{P}\}$  by fixing  $\{\Psi^{(t-1)}\}$**

Derive  $\hat{r}_{lk}$  according to Lemma 5

Derive  $\tilde{e}_{lk}$  based on (32)

Compute  $\{\mathbf{Q}\}$  by solving (34), i.e., by running

**While**  $\left( \min_{\forall lk} \tilde{e}_{lk}^{(n)} - \min_{\forall lk} \tilde{e}_{lk}^{(n-1)} \right) / \min_{\forall lk} \tilde{e}_{lk}^{(n-1)} \geq \gamma_2$

Update  $\mu^{(n)}$  based on (36)

Update  $\{\mathbf{Q}\}$  by solving (35)

Compute  $\{\mathbf{P}^{(t)}\}$  as  $\mathbf{P}_{lk}^{(t)} = \mathbf{Q}_{lk}^{(t)} \mathbf{Q}_{lk}^{(t)H} \forall lk$

**Optimizing over  $\{\Psi\}$  by fixing  $\{\mathbf{P}^{(t-1)}\}$**

Derive  $\hat{r}_{lk}^{(t)}$  according to Corollary 1

Calculate  $\{\Psi^{(t)}\}$  by solving (42)

$t = t + 1$

**End (While)**

**Return**  $\{\mathbf{P}^{(*)}\}$  and  $\{\Psi^{(*)}\}$ .

---

2) *STAR-RIS using the MS scheme*: The solution of (40) for STAR-RIS with the MS scheme is very similar to the solution for a reflective RIS. In this case, we can still use the surrogate functions in Corollary 1 to make the rates a jointly concave function of the STAR-RIS coefficients, i.e.,  $\{\Psi^t\}$  and  $\{\Psi^r\}$ . For  $\mathcal{T}_{SU}$ , we can update the STAR-RIS parameters by solving (42), and the algorithm obtains a stationary point of (24). For  $\mathcal{T}_{SI}$ , we have to “convexify” the constraint in (8), which can be done by rewriting it as the two convex constraints in [36, Eq. (34)] and [36, Eq. (36)]. Then we can update the STAR-RIS coefficients similar to the proposed scheme for the reflective RIS.

### C. Computational Complexity Analysis

Our optimization framework operates iteratively, with the actual computational complexity and runtime contingent upon the specific implementation of the algorithms. In this subsection, we calculate an approximate upper bound for the number of multiplications imposed by running our algorithms. To this end, we consider the maximization of the minimum rate with the feasibility set  $\mathcal{T}_U$ . The computational complexity of other optimization problems, including the weighted sum rate, the minimum EE, and global EE, can be similarly computed.

Each iteration of our proposed framework consists of two steps. In the first step, we optimize the transmit covariance matrices by solving (33), which is convex when the minimum rates of the users are maximized. We solve the convex problem in (33) by numerical optimization tools. To numerically solve a convex optimization problem, the number of Newton iterations increases proportionally with the square root of the number of its constraints [62, Chapter 11], which is equal to  $\sum_l (K_l + 1)$  in (33) for the maximization of the minimum rate. Note that the maximization of the minimum rate can be written as in [14, Eq. (30)] and thus, has  $\sum_l (K_l + 1)$  constraints, considering the power budget in (33d). Now, we provide an approximate upper bound for the number of multiplications to find a solution in each Newton iteration. To solve each Newton iteration,  $\sum_l K_l$  surrogate functions have to be computed for the rates,  $\tilde{r}_{lk}$ . The surrogate rates  $\tilde{r}_{lk}$  in Lemma 5 are quadratic in  $\{\mathbf{Q}\}$ , and the computational complexity to compute each surrogate rate can be approximated as  $\mathcal{O}\left[\sum_l \sum_k N_{BS,l}^2 (2N_{BS,l} + N_{u,lk})\right]$ . Note that the coefficients in (28) can be computed once at the beginning of the Newton iterations, and there is no need to recompute them in each Newton iteration to reduce the overall computational complexity of the framework. Finally, the computational complexity to update  $\{\mathbf{P}\}$  can be approximated as  $\mathcal{O}\left[\mathcal{L} \sum_l \sum_k N_{BS,l}^2 \sqrt{\sum_l (K_l + 1)} (2N_{BS,l} + N_{u,lk})\right]$ , where  $\mathcal{L} = \sum_l K_l$  is the total number of users in the system.

Now, we derive an approximation for the number of the multiplications needed to update  $\{\Psi\}$ . To this end, we have to calculate the number of multiplications for solving the surrogate optimization problem in (42), which is convex for the feasibility set  $\mathcal{T}_U$ . The number of constraints in (42) is equal to  $\sum_l K_l + \sum_m N_{RIS,m}$ . Thus, the number of the Newton iterations grows with  $\sqrt{\sum_l K_l + \sum_m N_{RIS,m}}$ . To solve each Newton iteration, we have to compute  $\mathcal{L}$  surrogate functions for the rates,  $\hat{r}_{lk}$ , as well as  $L\mathcal{L}$  equivalent channels, according to (2). To compute each channel,  $\mathbf{H}_{lk,i}$ ,  $\forall l, k, i$ , there are approximately  $\sum_m N_{u,lk} N_{BS,i} N_{RIS,m}$  multiplications, since the matrices  $\Psi_m$  are diagonal, which reduces the computational complexity. Moreover, the structure of the rates in (41) is very similar to the rates in (28). Hence, the computational complexity of calculating  $\hat{r}_{lk}$  in (41) is on the same order of the computational complexity of attaining  $\tilde{r}_{lk}$  as in (28) and can be approximated as  $\mathcal{O}\left[\sum_l K_l N_{BS,l}^2 (2N_{BS,l} + N_{u,lk})\right]$ . Finally, the computational complexity of updating  $\{\Psi\}$  can be approximated as  $\mathcal{O}\left[\sum_l \mathcal{L} \sqrt{\sum_l K_l + \sum_m N_{RIS,m}} (K_l N_{BS,l}^2 (2N_{BS,l} + N_{u,lk}) + \sum_m N_{u,lk} N_{BS,i} N_{RIS,m})\right]$ . Assuming that the maximum num-

ber of iterations is equal to  $T$ , the computational complexity of solving the maximization of the minimum rate for  $\mathcal{T}_U$  using our framework is  $T$  times the summation of the computational complexities of updating  $\{\mathbf{P}\}$  and  $\{\Psi\}$ .

### D. Discussion on Extending the Framework to Uplink

The framework can also be applied to the SE and EE maximization of the uplink along with FBL coding, since the structure of the rates with respect to the beamforming matrices and channels is very similar to the DL scenario considered in the paper. The detailed solution for the UL scenario is beyond the scope of this work. However, we provide some insights on how the proposed solutions can be modified to maximize the sum rate in UL communications in a multi-cell multiple-access channel (MAC), while the intercell interference is treated as noise. In this case, the sum rate of users associated with BS  $l$  is

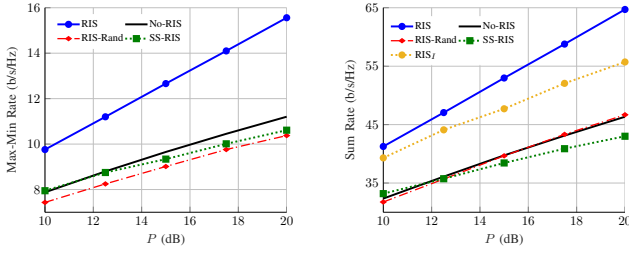
$$r_l = \log |\mathbf{I} + \mathbf{D}_l^{-1} \mathbf{S}_l| - \frac{Q^{-1}(\epsilon)}{n_t} \sqrt{2\text{Tr}(\mathbf{I} - \mathbf{D}_l(\mathbf{D}_l + \mathbf{S}_l)^{-1})}, \quad (50)$$

where  $\mathbf{S}_l = \sum_k \mathbf{H}_{lk,l} \mathbf{Q}_{lk} (\mathbf{H}_{lk,l} \mathbf{Q}_{lk})^H$  is the covariance matrix of the signals decoded at BS  $l$ , and  $\mathbf{D}_l = \sigma^2 \mathbf{I} + \sum_{\forall ik, i \neq l} \mathbf{H}_{lk,l} \mathbf{Q}_{lk} (\mathbf{H}_{lk,l} \mathbf{Q}_{lk})^H$  is the covariance matrix of the noise plus interference, where  $\mathbf{Q}_{ik}$  is the beamforming matrix at  $U_{ik}$ , and  $\mathbf{H}_{lk,l}$  is the uplink channel between  $U_{lk}$  and BS  $l$ . As it can be easily verified,  $\mathbf{S}_l$  and  $\mathbf{D}_l$  in (50) are quadratic in channels and beamforming matrices, which follow a similar structure as the matrices  $\mathbf{S}_{lk}$  and  $\mathbf{D}_{lk}$  in (13). Thus, to obtain a quadratic and concave surrogate function for  $r_l$ , we can employ the bounds in Lemma 3 and Lemma 4, and follow the steps in the proof of Lemma 5. Once the surrogate functions for the rates are calculated, the beamforming matrices and RIS coefficients can be updated by solving the corresponding optimization problems according to our proposed framework.

## IV. NUMERICAL RESULTS

In this section, we provide numerical results based on Monte Carlo simulations. To this end, we consider a two-cell system with one RIS and  $K$  users per cell, similar to [36, Fig. 2]. We also consider that each BS/user/RIS has  $N_{BS}/N_u/N_{RIS}$  TAs/RAs/elements. Moreover, the locations and heights of the users/BSs/RISs are chosen similar to [36]. We assume that the power budgets of the BSs are equal to  $P$ . To generate the channels, we assume that the links with respect to the RISs benefit from a line of sight (LoS) for both BSs and the users that are located in the same cell as the RIS. Therefore, these links follow Rician fading associated with a Rician factor of 3. More particularly, these channels are generated according to [28, (60)-(60)]. On the other hand, we assume that the direct links between the BSs and the users as well as the links through the RISs across the cells are of a non-LoS (NLoS) nature and consequently, follow a Rayleigh distribution. This means that each entry of the corresponding channel matrices follows a complex-valued proper Gaussian distribution with zero mean and unit variance. The large-scale fading of the links is modeled according to [28, (59)]. The path-loss exponents of the LoS and NLoS links are 2.2





**Fig. 2:** Spectral efficiency metrics versus  $P$  for  $N_{BS} = 4$ ,  $N_u = 4$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ .

and 3.75, respectively. The noise power density and channel bandwidth are, respectively, assumed to be  $-174$  dBm per Hz and 1.5 MHz. The other simulation parameters are chosen according to [28], [36].

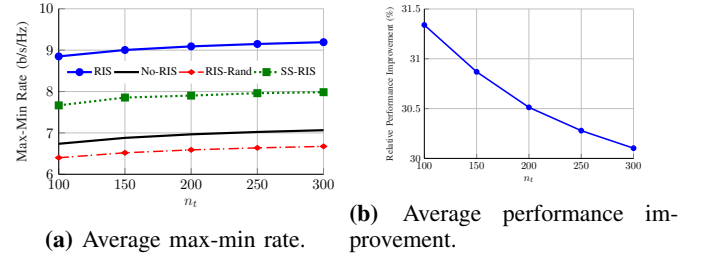
In the following, we provide numerical results for SE and EE maximization in Section IV-A and Section IV-B, respectively. The schemes considered in this section are as follows:

- **RIS (or RIS<sub>I</sub>):** Our algorithms for MIMO RIS-assisted URLLC systems with multiple data streams per users, and  $\mathcal{T}_U$  (or  $\mathcal{T}_I$ ).
- **No-RIS:** The scheme for MIMO URLLC systems with multiple data streams per users, but without RIS.
- **RIS-Rand (or S-RIS-Rand):** The algorithm for MIMO RIS-aided (or STAR-RIS-aided) URLLC systems with multiple data streams per users, but without optimizing RIS elements.
- **STAR-RIS:** Our algorithms for MIMO STAR-RIS-aided URLLC BCs with multiple data streams per users,  $\mathcal{T}_U$ , and the MS scheme.
- **SS-RIS:** Our algorithms for MIMO RIS-assisted URLLC systems with single-stream data transmission per users, and  $\mathcal{T}_U$ .

As emphasized in Section I, there is no other work on multi-user MIMO RIS-aided systems with FBL coding. Thus, we compare the performance of our proposed algorithm to a single-stream data transmission scheme, the multiple-stream data transmission scheme for systems without RIS, and with a non-optimized RIS coefficients as benchmarks.

#### A. Spectral Efficiency Metrics

Here, we present numerical results for SE maximization. To this end, we consider the maximum of the average minimum rate and the average sum rate of users as performance metrics. We refer to the maximum of the minimum achievable rate of users as the max-min rate. Note that it is likely that all the users get the same achievable rate when maximizing the minimum rate, which can provide a reasonable fairness among the users [63]. Nevertheless, when the sum rate is maximized, it could be the case that the users with weaker channels are switched off if QoS constraints are not considered. To provide a comprehensive analysis, we explore the impact of various parameters on the system performance, including the



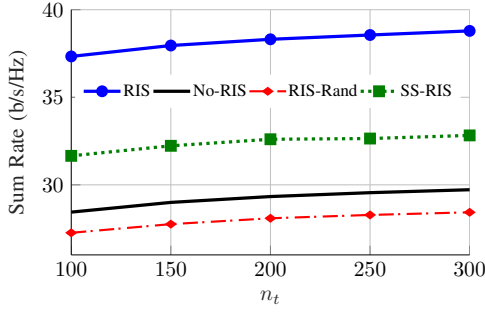
**Fig. 3:** Average max-min rate and relative performance improvement by RIS versus  $n_t$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ .

BS power budgets, packet length, as well as the maximum tolerable packet error rate.

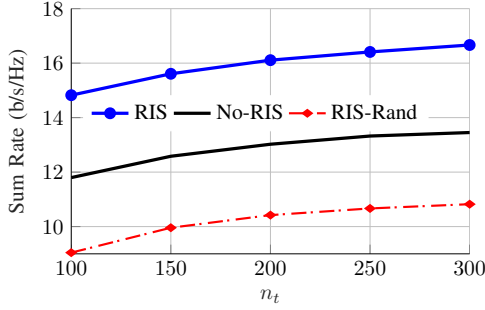
1) *Impact of power budget:* In Fig. 2, we show the average max-min rate and sum rate versus  $P$  for  $N_{BS} = 4$ ,  $N_u = 4$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ . In this figure, the RISs significantly improve the min-max rate and the sum rate, when the RIS elements are optimized. Surprisingly, employing RISs having random elements degrades the max-min rate. Moreover, we can observe that the benefits of RIS increase with  $P$  for these examples. Additionally, the multi-stream scheme substantially improves the SE for these two examples, where even the No-RIS scheme is slightly better than the single-stream scheme, especially at high SNRs. In these examples, the maximum number of streams per user is  $I = \min(N_{BS}, N_u) = 4$ , which makes single-stream transmission gravely suboptimal. Indeed, when the system complexity increases, a more sophisticated resource allocation scheme should be adopted in order to avoid performance degradation.

2) *Impact of packet length:* Fig. 3 shows the average max-min rate and the relative performance improvement by RIS versus  $n_t$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ . Observe that an RIS can substantially increase the average max-min rate. The benefits of deploying RISs decrease with  $n_t$ . Indeed, the shorter the packet length is, the higher the relative improvements provided by RISs can be. As indicated, the packet length correlates with the level of stringency in the latency constraint. Shorter packet lengths are needed, when the latency constraint is more stringent. Thus, this result shows that the RIS benefits increase as the latency constraint becomes more stringent. In other words, RISs can even be more beneficial for URLLC systems. Moreover, the rates increase with  $n_t$  and converge to the Shannon rate when  $n_t$  becomes higher. Furthermore, we can observe that employing multi-stream data transmission substantially increases the average max-min rate for all the values of  $n_t$ . Note that the benefits of multi-stream data transmission in this figure is lower than in Fig. 2 since the maximum number of streams per users,  $I$ , is 3 in this example, which is less than  $I$  in Fig. 2, and the benefits of multi-stream schemes grows with  $I$ .

Fig. 4 shows the average sum rate versus  $n_t$  for  $P = 10$  dB,  $L = 2$ ,  $M = 2$ ,  $\epsilon = 10^{-5}$  and for different values of  $N_{BS}$ ,  $N_u$ ,  $K$  and  $N_{RIS}$ . In this figure, RISs substantially enhance the average sum rate. However, the benefits of RISs

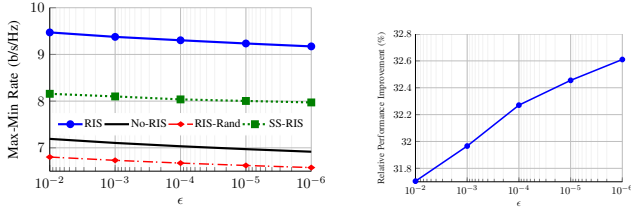


(a)  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ , and  $N_{RIS} = 20$ .



(b)  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 4$ , and  $N_{RIS} = 60$ .

**Fig. 4:** Average sum rate versus  $n_t$  for  $P = 10$  dB,  $L = 2$ ,  $M = 2$ , and  $\epsilon = 10^{-5}$ .

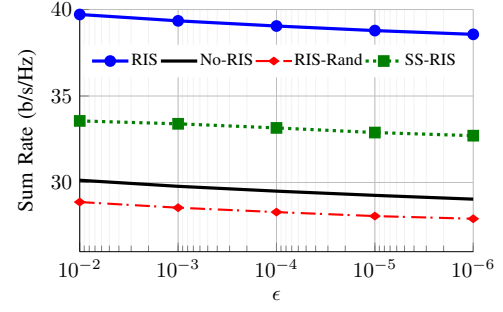


(a) Average max-min rate. (b) Average performance improvement.

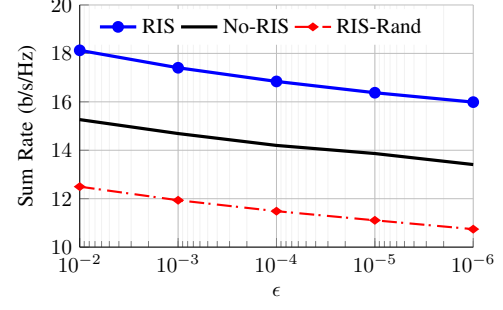
**Fig. 5:** Average max-min rate and relative performance improvement by RIS versus  $\epsilon$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$  bits, and  $N_{RIS} = 20$ .

are much more significant when there are less users in the system (Fig. 4a). Moreover, we can observe that the impact of decreasing the packet length is more severe in the system for a higher number of users. This may show the importance of employing effective interference-management techniques, which should be addressed in future studies. In Fig. 4a, we also observe that the algorithm conceived for the multi-stream data transmission per user outperforms the beamforming scheme, which employs a single-stream data transmission.

3) *Impact of the reliability constraint:* Fig. 5 shows the average max-min rate and relative performance improvement enabled by the deployment of RISs versus  $\epsilon$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$  bits, and  $N_{RIS} = 20$ . In this example, RIS significantly enhances the average max-min rate for all the  $\epsilon$  considered. As expected, the average max-min rate decreases when the reliability constraint is more stringent. In other words, we have to transmit at a lower rate to reduce the decoding error rate. Moreover,



(a)  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ , and  $N_{RIS} = 20$ .



(b)  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 4$ , and  $N_{RIS} = 60$ .

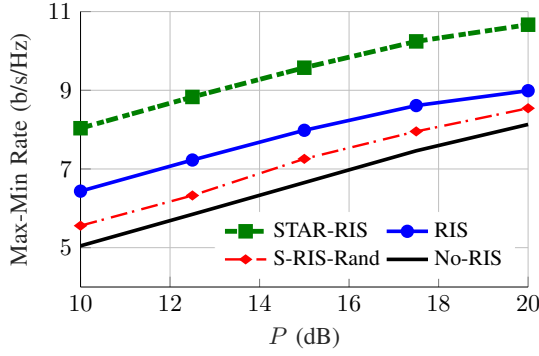
**Fig. 6:** Average sum rate versus  $\epsilon$  for  $P = 10$  dB,  $L = 2$ ,  $M = 2$ , and  $n_t = 256$  bits.

the multi-stream scheme significantly outperforms the single-stream data transmission. Furthermore, we can observe in Fig. 5b that the benefits of RIS increase when  $\epsilon$  decreases. Thus, RISs can be even more beneficial when more reliable communication is required.

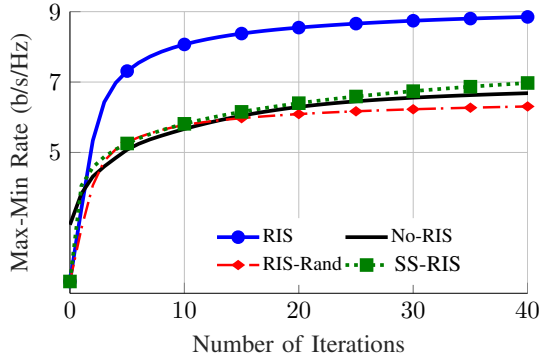
Fig. 6 shows the average sum rate versus  $\epsilon$  for  $P = 10$  dB,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$  bits and different  $N_{BS}$ ,  $N_u$ ,  $K$  and  $N_{RIS}$ . As it can be observed, RISs enhance the average sum rate in both networks, if the RIS elements are optimized. However, RISs with random coefficients decrease the average sum rate in these two examples. We can also observe that the impact of varying  $\epsilon$  is higher in the system supporting more users. Additionally, we can note a significantly greater advantage from RISs when there are fewer users in the network. Moreover, we can observe that the average sum rate substantially increases, if we employ multi-stream data transmission.

4) *Comparison of reflective RIS and STAR-RIS:* A reflective RIS has the same performance as a STAR-RIS if all the users are in the reflection half-space of the reflective/STAR-RIS. Thus, to evaluate the performance differences between these two technologies, we consider a single-cell  $2 \times 2$  MIMO BC in which one of the users is in the reflection space, and the other one is in the transmission space. As shown in Fig. 7, STAR-RIS using the MS scheme can significantly outperform reflective RIS in this example. For instance, STAR-RIS provides about 25% higher average max-min rate at  $P = 10$  dB compared to the reflective RIS.

5) *Convergence behavior:* Fig. 8 shows the average max-min rate versus the number of iterations for  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ ,  $\epsilon = 10^{-4}$ ,  $n_t = 256$  bits and  $N_{RIS} = 20$ . This figure



**Fig. 7:** Average max-min rate versus  $P$  for  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 2$ ,  $L = 1$ ,  $M = 1$ ,  $\epsilon = 10^{-5}$ ,  $n_t = 256$  bits and  $N_{RIS} = 40$ .



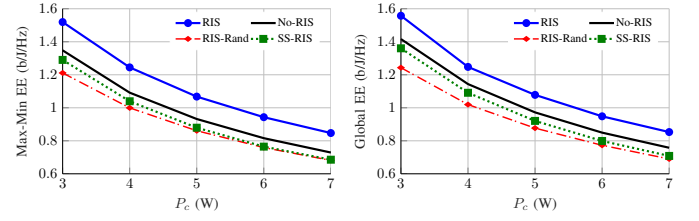
**Fig. 8:** Average max-min rate versus the number of iterations for  $N_{BS} = 4$ ,  $N_u = 3$ ,  $K = 2$ ,  $\epsilon = 10^{-4}$ ,  $n_t = 256$  bits and  $N_{RIS} = 20$ .

illustrates the convergence of the considered algorithms, from which the performance-complexity tradeoff may be inferred. In this example, our scheme proposed for RIS-aided systems employing multiple streams per user outperforms the final solution of the other algorithms after as few as four iterations. Indeed, the lower computational complexity of these alternative algorithms results in significant performance degradation, while to achieve a certain target performance, the RIS scheme requires substantially fewer iterations. Additionally, the No-RIS, RIS-Rand, and SS-RIS algorithms may not support a specific max-min rate, while our framework can achieve it with a reasonable number of iterations. For instance, while the No-RIS, RIS-Rand, and SS-RIS algorithms struggle to support a max-min rate of 8 b/s/Hz, our framework achieves it in only nine iterations.

### B. Energy Efficiency Metrics

Now we investigate the EE of RIS in MU-MIMO URLLC BCs. To this end, we consider the impact of  $P_c$ ,  $\epsilon$ , and  $n_t$ . In the examples provided in this subsection, we assume that each RIS consumes 1 W power. Thus, to make a fair comparison, we consider a lower constant power ( $P_c$ ) for the systems operating without RISs. Moreover, we assume that the power budget of the BSs is  $P = 10$  dB.

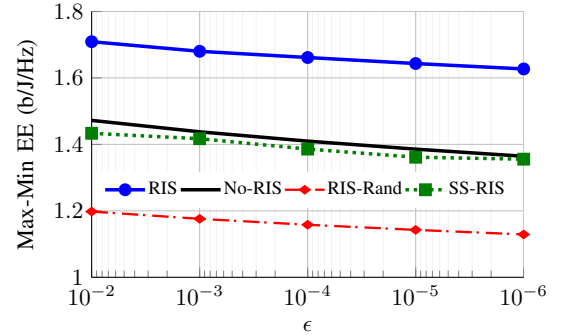
1) *Impact of  $P_c$ :* Fig. 9 shows the average max-min EE and GEE versus  $P_c$  for  $N_{BS} = 5$ ,  $N_u = 5$ ,  $K = 2$ ,  $L = 2$ ,



(a) Average max-min EE.

(b) Average global EE.

**Fig. 9:** Energy efficiency metrics versus  $P_c$  for  $N_{BS} = 5$ ,  $N_u = 5$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ .

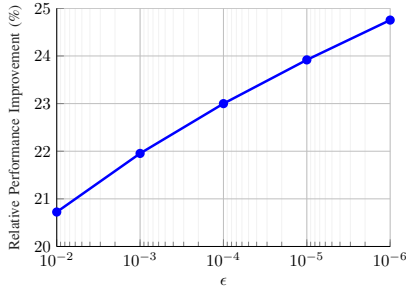


**Fig. 10:** Average max-min EE versus  $\epsilon$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 4$ ,  $K = 2$ ,  $L = 1$ ,  $M = 1$ ,  $n_t = 256$  bits, and  $N_{RIS} = 20$ .

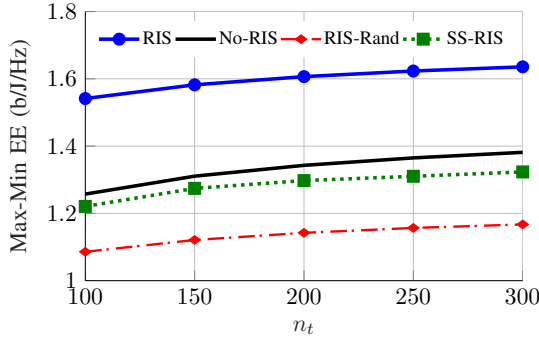
$M = 2$ ,  $n_t = 256$  bits,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ . Note that we use the term max-min EE to refer to the highest minimum EE. As it can be observed, RISs significantly increase the average max-min EE and GEE with FBL. Interestingly, RISs may reduce the EE, if their elements are random. However, our proposed algorithms can statistically enhance the EE of RIS-aided scenarios. For instance, in the particular example of Fig. 9a, an RIS provides more than 10% improvements over the systems disregarding the RISs for all the values of  $P_c$  considered. Additionally, the multi-stream scheme significantly outperforms the single-stream data transmission in the both examples.

2) *Impact of the reliability constraint:* Fig. 10 shows the average max-min EE versus  $\epsilon$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 4$ ,  $K = 2$ ,  $L = 1$ ,  $M = 1$ ,  $n_t = 256$  bits, and  $N_{RIS} = 20$ . In this example, the RIS provides a significant gain, which increases with  $\epsilon$ . Again, RIS decreases the max-min EE when its elements are random. Moreover, we can observe that single-stream data transmission is suboptimal in this  $4 \times 4$  MIMO system. Indeed, the multi-stream systems communicating without RIS outperforms the RIS-aided single-stream data transmission.

Fig. 11 shows the average EE performance improvement attained by RISs versus  $\epsilon$  for  $P = 10$  dB,  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$  bits, and  $N_{RIS} = 20$ . In this example, the RIS substantially increases the average max-min for all values of  $\epsilon$ . Moreover, higher gains are achieved by RISs, when the tolerable bit error rate is lower. Thus, the more reliable the communication has to be, the more energy efficient the RIS-aided systems becomes.



**Fig. 11:** Average EE performance improvement by RIS versus  $\epsilon$  for  $P = 10$  dB,  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $n_t = 256$  bits, and  $N_{RIS} = 20$ .



**Fig. 12:** Average max-min EE versus  $n_t$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 4$ ,  $K = 2$ ,  $L = 1$ ,  $M = 1$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ .

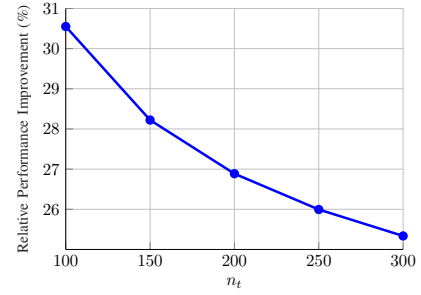
3) *Impact of the packet length:* Fig. 12 shows the average max-min EE versus  $n_t$  for  $P = 10$  dB,  $N_{BS} = 4$ ,  $N_u = 4$ ,  $K = 2$ ,  $L = 1$ ,  $M = 1$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ . In this figure, RIS provides a significant gain when the RIS elements are optimized by our proposed algorithm. However, random RIS coefficients degrade the EE performance.

Fig. 13 shows the average EE improvement by RISs versus  $n_t$  for  $P = 10$  dB,  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ . The average improvements reduce as  $n_t$  increases. This indicates that the lower the tolerable latency, the higher gain the RIS can provide. In other words, RIS-aided systems become more energy efficient when a low latency is required, as in control channels, for example.

## V. SUMMARY, CONCLUSIONS, AND FUTURE RESEARCH

An optimization framework was proposed for MU-MIMO RIS-aided systems with FBL by considering the NA for the rate expressions. To this end, we first calculated closed-form expressions for the FBL rate and then obtained suitable concave lower bounds for the FBL rates. Our proposed framework can be adapted to a large variety of MU-MIMO systems in which interference is treated as noise. Moreover, the framework can obtain a stationary point of a broad spectrum of practical optimization problems such as the maximization of the minimum/sum rate, GEE and minimum EE, when the set of the feasible RIS coefficients adheres to convexity. In summary, the key conclusions of this work are:

- RISs may significantly increase the average max-min rate, sum rate, max-min EE and global EE of the MU-MIMO



**Fig. 13:** Average EE performance improvement by RIS versus  $n_t$  for  $P = 10$  dB,  $N_{BS} = 2$ ,  $N_u = 2$ ,  $K = 2$ ,  $L = 2$ ,  $M = 2$ ,  $\epsilon = 10^{-5}$ , and  $N_{RIS} = 20$ .

systems considered. However, the RIS elements should be optimized to attain the above benefits, since RISs utilizing random elements may even degrade the system performance.

- The benefits of RISs increase when the packet length is reduced and/or the tolerable bit error rate is lower. The packet length can be related to the latency constraint, and the tolerable bit error rate represents the reliability constraint. Thus, these results show that RISs can be even more beneficial in URLLC systems than in non-URLLC systems.
- Multiple-stream data transmission for each user significantly outperforms single-stream data transmission (beamforming) in MU-MIMO RIS-assisted URLLC systems. Indeed, both the reliability and latency can be enhanced, when multiple-stream data transmission is employed in multiple-antenna systems.

In future research, it would be interesting to integrate advanced interference-management techniques, such as RSMA and NOMA, into MU-MIMO URLLC systems. In this regard, the solutions in, e.g., [36], [48], in combinations with the FBL rate expressions in Lemma 1 might be helpful. Another challenging research direction is to extend the optimization framework proposed in this work to scenarios with statistical or imperfect CSI. To this end, the robust designs in [12], [64] can be harnessed. Moreover, studying the performance of other concepts/technologies for RIS, such as holographic RIS [65], active RIS [66], BD-RIS [67], [68], and globally-passive RIS [69], [70] can be another promising direction for extending this work. Furthermore, considering RIS-aided cell-free URLLC systems is worth exploring in future research. Finally, another interesting line of research is to develop computationally more efficient resource allocation schemes that do not compromise performance.

## REFERENCES

- [1] C.-X. Wang *et al.*, "On the road to 6G: Visions, requirements, key technologies and testbeds," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 2, pp. 905–974, 2023.
- [2] T. Gong *et al.*, "Holographic MIMO communications: Theoretical foundations, enabling technologies, and future directions," *IEEE Commun. Surv. Tutor.*, vol. 26, no. 1, pp. 196–257, 2023.
- [3] Q. Wu *et al.*, "Intelligent reflecting surface aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, 2021.



- [4] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, 2020.
- [5] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, 2019.
- [6] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, 2019.
- [7] Q.-U.-A. Nadeem, *et al.*, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, 2020.
- [8] M. Soleymani, I. Santamaria, E. Jorswieck, and S. Rezvani, "NOMA-based improper signaling for multicell MISO RIS-assisted broadcast channels," *IEEE Trans. Signal Process.*, vol. 71, pp. 963–978, March 2023.
- [9] I. Santamaria, M. Soleymani, E. Jorswieck, and J. Gutierrez, "Interference leakage minimization in RIS-assisted MIMO interference channels," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, 2023.
- [10] M. Soleymani, I. Santamaria, A. Sezgin, and E. Jorswieck, "Maximization of minimum rate in MIMO OFDM RIS-assisted broadcast channels," *IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process. (CAMSAP)*, 2023.
- [11] W. Xu, J. An, H. Li, L. Gan, and C. Yuen, "Algorithm unrolling-based distributed optimization for RIS-assisted cell-free networks," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 944–957, 2024.
- [12] J. Yao, J. Xu, W. Xu, D. W. K. Ng, C. Yuen, and X. You, "Robust beamforming design for RIS-aided cell-free systems with CSI uncertainties and capacity-limited backhaul," *IEEE Trans. Commun.*, vol. 71, no. 8, pp. 4636–4649, 2023.
- [13] W. Lyu, Y. Xiu, S. Yang, C. Yuen, and Z. Zhang, "Energy-efficient cell-free network assisted by hybrid RISs," *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 718–722, 2023.
- [14] M. Soleymani, I. Santamaria, and E. Jorswieck, "Spectral and energy efficiency maximization of MISO STAR-RIS-assisted URLLC systems," *IEEE Access*, vol. 11, pp. 70 833–70 852, 2023.
- [15] M. Soleymani, I. Santamaria, E. Jorswieck, and B. Clerckx, "Optimization of rate-splitting multiple access in beyond diagonal RIS-assisted URLLC systems," *IEEE Trans. Wireless Commun.*, 2023.
- [16] Y. Li, C. Yin, T. Do-Duy, A. Masaracchia, and T. Q. Duong, "Aerial reconfigurable intelligent surface-enabled URLLC UAV systems," *IEEE Access*, vol. 9, pp. 140 248–140 257, 2021.
- [17] T.-H. Vu, T.-V. Nguyen, D. B. da Costa, and S. Kim, "Intelligent reflecting surface-aided short-packet non-orthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4500–4505, 2022.
- [18] H. Xie, J. Xu, Y.-F. Liu, L. Liu, and D. W. K. Ng, "User grouping and reflective beamforming for IRS-aided URLLC," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2533–2537, 2021.
- [19] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghayeb, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, 2022.
- [20] W. R. Ghanem, V. Jamali, and R. Schober, "Optimal resource allocation for multi-user OFDMA-URLLC MEC systems," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 2005–2023, 2022.
- [21] B. Makki, T. Svensson, M. Coldrey, and M.-S. Alouini, "Finite blocklength analysis of large-but-finite MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 113–116, Feb. 2019.
- [22] B. Makki, T. Svensson, T. Eriksson, and M.-S. Alouini, "On the required number of antennas in a point-to-point large-but-finite MIMO system: Outage-limited scenario," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1968–1983, May 2016.
- [23] C. Li, Y. Wang, W. Chen, and H. V. Poor, "Ultra-reliable and low-latency multiple-antenna communications in the high SNR regime," *IEEE Wireless Commun. Lett.*, vol. 12, no. 3, pp. 461–465, March 2023.
- [24] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
- [25] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on laplace integrals and their asymptotic approximations," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854–6883, 2016.
- [26] —, "On the evaluation of the Polyanskiy-Poor-Verdú converse bound for finite block-length coding in AWGN," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6578–6590, 2015.
- [27] Y. Polyanskiy, *Channel coding: Non-asymptotic fundamental limits*. Princeton University, 2010.
- [28] M. Soleymani, I. Santamaria, and P. J. Schreier, "Improper signaling for multicell MIMO RIS-assisted broadcast channels with I/Q imbalance," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 2, pp. 723–738, 2022.
- [29] M. Soleymani, I. Santamaria, and E. Jorswieck, "Rate region of MIMO RIS-assisted broadcast channels with rate splitting and improper signaling," *WSA 2023; 26th International ITG Workshop on Smart Antennas*, 2023.
- [30] M. Abughalwa, H. Tuan, D. Nguyen, H. Poor, and L. Hanzo, "Finite-blocklength RIS-aided transmit beamforming," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12 374–12 379, 2022.
- [31] W. R. Ghanem, V. Jamali, and R. Schober, "Joint beamforming and phase shift optimization for multicell IRS-aided OFDMA-URLLC systems," in *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2021, pp. 1–7.
- [32] H. Ren, K. Wang, and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 707–723, 2021.
- [33] B. Zhang, K. Wang, K. Yang, and G. Zhang, "IRS-assisted short packet wireless energy transfer and communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 2, pp. 303–307, 2022.
- [34] J. Scarlett, V. Y. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 81–92, 2016.
- [35] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkhachan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, 2020.
- [36] M. Soleymani, I. Santamaria, and E. Jorswieck, "Rate splitting in MIMO RIS-assisted systems with hardware impairments and improper signaling," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 4580–4597, April 2023.
- [37] 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. on Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2020.
- [38] Y. Wang, V. W. Wong, and J. Wang, "Flexible rate-splitting multiple access with finite blocklength," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1398–1412, May 2023.
- [39] S. Pala, K. Singh, M. Katwe, and C.-P. Li, "Joint optimization of URLLC parameters and beamforming design for multi-RIS-aided MU-MISO URLLC system," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 148–152, 2022.
- [40] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFDMA-URLLC systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7184–7200, Nov. 2020.
- [41] H. Yu *et al.*, "Joint design of reconfigurable intelligent surfaces and transmit beamforming under proper and improper Gaussian signaling," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2589–2603, 2020.
- [42] L. Zhang, Y. Wang, W. Tao, Z. Jia, T. Song, and C. Pan, "Intelligent reflecting surface aided MIMO cognitive radio systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11 445–11 457, 2020.
- [43] H. Zhang and B. Di, "Intelligent omni-surfaces: Simultaneous refraction and reflection for full-dimensional wireless communications," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 4, pp. 1997–2028, 2022.
- [44] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, H. V. Poor, and L. Hanzo, "STAR: Simultaneous transmission and reflection for 360 coverage by intelligent surfaces," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 102–109, 2021.
- [45] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, 2022.
- [46] M. Soleymani, I. Santamaria, A. Sezgin, and E. Jorswieck, "Maximizing spectral and energy efficiency in multi-user MIMO OFDM systems with RIS and hardware impairment," *arXiv preprint arXiv:2401.11921*, 2024.
- [47] M. Soleymani, I. Santamaria, and E. Jorswieck, "Energy-efficient rate splitting for MIMO STAR-RIS-assisted broadcast channels with I/Q imbalance," *Proc. IEEE Eu. Signal Process. Conf. (EUSIPCO)*, pp. 1504–1508, 2023.
- [48] —, "NOMA-based improper signaling for MIMO STAR-RIS-assisted broadcast channels with hardware impairments," *IEEE Global Commun. Conf. (GLOBECOM)*, 2023.

- [49] J. Xu, Y. Liu, X. Mu, R. Schober, and H. V. Poor, "STAR-RISs: A correlated T&R phase-shift model and practical phase-shift configuration strategies," *IEEE J. Sel. Topics Signal Process.*, pp. 1–1, 2022.
- [50] V. S. Annapureddy and V. V. Veeravalli, "Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on the capacity region," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3032–3050, 2009.
- [51] C. Geng, N. Naderializadeh, A. S. Avestimehr, and S. A. Jafar, "On the optimality of treating interference as noise," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1753–1767, 2015.
- [52] G. D. Forney and G. Ungerboeck, "Modulation and coding for linear Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2384–2415, 1998.
- [53] I. Santamaria, P. Crespo, C. Lameiro, and P. Schreier, "Information-theoretic analysis of a family of improper discrete constellations," *Entropy*, vol. 20, no. 1, p. 45, 2018.
- [54] S. Javed, O. Amin, B. Shihada, and M.-S. Alouini, "A journey from improper Gaussian signaling to asymmetric signaling," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 1539–1591, 2020.
- [55] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access with finite blocklength for short-packet and low-latency downlink communications," *IEEE Trans. Veh. Technol.*, 2022.
- [56] —, "Max-min fairness of rate-splitting multiple access with finite blocklength communications," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6816–6821, 2023.
- [57] B. Liu, P. Zhu, J. Li, D. Wang, and X. You, "Energy-efficient optimization in distributed massive MIMO systems for slicing eMBB and URLLC services," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10473–10487, 2023.
- [58] 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.
- [59] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Found Trends in Commun. Inf. Theory*, vol. 11, no. 3–4, pp. 185–396, 2015.
- [60] H. Yu, H. D. Tuan, E. Dutkiewicz, H. V. Poor, and L. Hanzo, "Maximizing the geometric mean of user-rates to improve rate-fairness: Proper vs. improper Gaussian signaling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 295–309, 2022.
- [61] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, 2017.
- [62] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [63] M. Soleymani, I. Santamaria, and P. J. Schreier, "Improper Gaussian signaling for the  $K$ -user MIMO interference channels with hardware impairments," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11632–11645, 2020.
- [64] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A framework of robust transmission design for IRS-aided MISO communications with imperfect cascaded channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 5092–5106, 2020.
- [65] J. An, C. Xu, D. W. K. Ng, G. C. Alexandropoulos, C. Huang, C. Yuen, and L. Hanzo, "Stacked intelligent metasurfaces for efficient holographic MIMO communications in 6G," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2380–2396, 2023.
- [66] Z. Zhang, L. Dai, X. Chen, C. Liu, F. Yang, R. Schober, and H. V. Poor, "Active RIS vs. passive RIS: Which will prevail in 6G?" *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, 2023.
- [67] H. Li, S. Shen, and B. Clerckx, "Beyond diagonal reconfigurable intelligent surfaces: From transmitting and reflecting modes to single-, group-, and fully-connected architectures," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2311–2324, 2023.
- [68] I. Santamaria, M. Soleymani, E. Jorswieck, and J. Gutiérrez, "MIMO capacity maximization with beyond-diagonal RIS," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. IEEE, 2024.
- [69] R. K. Fotock, A. Zappone, and M. Di Renzo, "Energy efficiency optimization in RIS-aided wireless networks: Active versus nearly-passive RIS with global reflection constraints," *IEEE Trans. Commun.*, vol. 72, no. 1, pp. 257–272, 2024.
- [70] M. Soleymani, I. Santamaria, E. Jorswieck, M. Di Renzo, and J. Gutiérrez, "Energy efficiency comparison of RIS architectures in MISO broadcast channels," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. IEEE, 2024.



communications, wireless networking, convex optimization and statistical signal processing.



**Ignacio Santamaria** (M'96-SM'05) received the Telecommunication Engineer degree and the Ph.D. degree in electrical engineering from the Universidad Politécnica de Madrid (UPM), Spain, in 1991 and 1995, respectively. In 1992, he joined the Department of Communications Engineering, Universidad de Cantabria, Spain, where he is Full Professor since 2007. He has co-authored more than 250 publications in refereed journals and international conference papers, and holds two patents. He has co-authored the book *D. Ramirez, I. Santamaria, and L.L. Scharf, "Coherence in Signal Processing and Machine Learning"*, Springer, 2022. His current research interests include signal processing algorithms and information-theoretic aspects of multiuser multiantenna wireless communication systems, multivariate statistical techniques and machine learning theories. He has been involved in numerous national and international research projects on these topics. He has been a visiting researcher at the University of Florida (in 2000 and 2004), at the University of Texas at Austin (in 2009), and at the Colorado State University (in 2015 and 2017). He has been Associate Editor of the *IEEE Transactions on Signal Processing* (2011–2015), and Senior Area Editor of the *IEEE Transactions on Signal Processing* (2013–2015). He has been a member of the *IEEE Machine Learning for Signal Processing Technical Committee* (2009–2014), member of the *IEEE Signal Processing Theory and Methods Technical Committee* (2020–2022), and member of the *IEEE Data Science Initiative (DSI) steering committee* (2020–2022). Prof. Santamaria co-authored a paper that received the 2012 *IEEE Signal Processing Society Young Author Best Paper Award*, and has received the 2022 *IHP International Wolfgang Mehr Fellowship Award*.



**Eduard A. Jorswieck** (Fellow, IEEE) received the Ph.D. degree in computer engineering from TU Berlin, in 2004. From 2006 to 2008, he was a Postdoctoral Fellow and an Assistant Professor with the Signal Processing Group, KTH Stockholm. From 2008 to 2019, he was the Chair for Communication Theory with TU Dresden. He is currently the Managing Director of the Institute of Communications Technology and the Head of the Chair for Communications Systems and a Full Professor with Technische Universität Braunschweig, Germany. His research interest lies in the broad area of communications. He has published more than 170 journal articles, 15 book chapters, one book, three monographs, and more than 320 conference papers. He is a Fellow of the IEEE. He was a recipient of the *IEEE Signal Processing Society Best Paper Award*. Since 2017, he has been the Editor-in-Chief of the *Springer EURASIP Journal on Wireless Communications and Networking*. Since 2022, he has been on the editorial board of the *IEEE TRANSACTIONS ON COMMUNICATIONS*. He was on the editorial boards of the *IEEE SIGNAL PROCESSING LETTERS*, the *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, the *IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS*, and the *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*.





**Robert Schober** (S'98, M'01, SM'08, F'10) received the Diplom (Univ.) and the Ph.D. degrees in electrical engineering from Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Germany, in 1997 and 2000, respectively. From 2002 to 2011, he was a Professor and Canada Research Chair at the University of British Columbia (UBC), Vancouver, Canada. Since January 2012 he is an Alexander von Humboldt Professor and the Chair for Digital Communication at FAU. His research interests fall into the broad areas of Communication Theory,

Wireless and Molecular Communications, and Statistical Signal Processing.

Robert received several awards for his work including the 2002 Heinz Maier Leibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, a 2006 UBC Killam Research Prize, a 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, a 2011 Alexander von Humboldt Professorship, a 2012 NSERC E.W.R. Stacie Fellowship, a 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee, and the 2022 IEEE Vehicular Technology Society Stuart F. Meyer Memorial Award. Furthermore, he received numerous Best Paper Awards for his work including the 2022 ComSoc Stephen O. Rice Prize and the 2023 ComSoc Leonard G. Abraham Prize. Since 2017, he has been listed as a Highly Cited Researcher by the Web of Science. Robert is a Fellow of the Canadian Academy of Engineering, a Fellow of the Engineering Institute of Canada, and a Member of the German National Academy of Science and Engineering.

He served as Editor-in-Chief of the IEEE Transactions on Communications, VP Publications of the IEEE Communication Society (ComSoc), ComSoc Member at Large, and ComSoc Treasurer. Currently, he serves as Senior Editor of the Proceedings of the IEEE and as ComSoc President.



**Lajos Hanzo** is a Fellow of the Royal Academy of Engineering, FIEEE, FIET, Fellow of EURASIP and a Foreign Member of the Hungarian Academy of Sciences. He coauthored 2000+ contributions at IEEE Xplore and 19 Wiley-IEEE Press monographs. He was bestowed upon the IEEE Eric Sumner Technical Field Award.