Resource Allocation for Multi-User Downlink MISO OFDMA-URLLC Systems

Walid R. Ghanem, Vahid Jamali, Yan Sun, and Robert Schober

arXiv:1910.06127v1 [cs.IT] 14 Oct 2019

Abstract

This paper considers the resource allocation algorithm design for downlink multiple-input single-output (MISO) orthogonal frequency division multiple access (OFDMA) ultra-reliable low latency communication (URLLC) systems. To meet the stringent delay requirements of URLLC, short packet transmission is adopted and taken into account for resource allocation algorithm design. The resource allocation is optimized for maximization of the weighted system sum throughput subject to quality-of-service (QoS) constraints regarding the URLLC users' number of transmitted bits, packet error probability, and delay. Despite the non-convexity of the resulting optimization problem, the optimal solution is found via monotonic optimization. The corresponding optimal resource allocation policy can serve as a performance upper bound for sub-optimal low-complexity solutions. We develop such a low-complexity resource allocation algorithm to strike a balance between performance and complexity. Our simulation results reveal the importance of using multiple antennas for reducing the latency and improving the reliability of URLLC systems. Moreover, the proposed sub-optimal algorithm is shown to closely approach the performance of the proposed optimal algorithm and outperforms two baseline schemes by a considerable margin, especially when the users have heterogeneous delay requirements. Finally, conventional resource allocation designs based on Shannon's capacity formula are shown to be not applicable in MISO OFDMA-URLLC systems as they may violate the users' delay constraints.

I. Introduction

The fifth-generation (5G) wireless communication networks impose several different system design objectives including high data rates, high spectral efficiency, reduced latency, higher system capacity, and massive device connectivity. One important objective is to enable ultra-reliable low latency communication (URLLC). URLLC is required for mission critical applications such as factory automation, e-health, autonomous driving, tactile Internet, and

augmented reality to facilitate real-time machine-to-machine and human-to-machine inter-action [2]. URLLC imposes strict quality-of-service (QoS) requirements including a very low latency (e.g., $1\,\text{ms}$) and a low packet error probability (e.g., 10^{-6}) [2]. In addition, the data packet size is typically small, e.g., around 160 bits [3]. Existing mobile communi-cation systems cannot meet these requirements. For example, for the long term evolution (LTE) system, the total frame time is $10\,\text{ms}$, which exceeds the total latency requirement of URLLC applications [4]. The main challenges for the design of URLLC systems are the two contradicting requirements of low latency and ultra high reliability. For this reason, new design strategies are needed to enable URLLC.

Modern communication systems employ multi-carrier transmission, e.g., orthogonal fre-quency division multiple access (OFDMA), due to its ability to exploit multi-user diversity, its robustness to multipath fading, and the flexibility it provides for the allocation of resources, such as power and bandwidth [5]. Furthermore, multiple antenna technology provides more degrees of freedom for resource allocation and facilitates multiplexing and diversity gains [5]. Hence, future communication networks are expected to combine the concepts of multiple antennas, OFDMA, and URLLC.

However, with the exception of our conference paper [1], the resource allocation algorithm design for OFDMA-URLLC systems has not been studied, yet. The authors in [6] studied the weighted sum rate maximization for multi-user downlink OFDMA systems. In [7], the authors studied the resource allocation algorithm design for energy-efficient communication in multi-cell OFDMA systems. The authors in [8] investigated the joint optimal power, sub-carrier, and relay node allocation in multi-relay assisted dual-hop cooperative orthogonal frequency division multiplexing (OFDM) systems. In [9], the authors studied the resource allocation for multiple-input single-output (MISO) OFDMA systems, where a base station (BS) equipped with multiple antennas served multiple single antenna users. However, the resource allocation algorithms proposed in [6]–[9] were based on Shannon's capacity formula for the additive white Gaussian noise (AWGN) channel. Since URLLC systems employ a short frame structure and a small packet size to reduce latency, the relation between the achievable rate, decoding error probability, and transmission delay cannot be captured by Shannon's capacity formula which assumes infinite block length and zero error probability [10]. If Shannon's capacity formula is utilized for resource allocation design for URLLC systems, the latency will be underestimated and the reliability will be overestimated, and as a result, the QoS requirements of the users cannot be met. Therefore, the results in [6]–[9] and the related literature are not applicable for resource allocation in MISO OFDMA-URLLC

systems. Hence, new resource algorithms for MISO OFDMA systems taking into account the specific properties and requirements of URLLC are needed, which is the main motivation for this paper.

In recent years, the performance limits of short packet communication (SPC) [11] have received significant attention in the literature. These performance limits provide a relationship between the achievable rate, decoding error probability, and packet length. The pioneering work in [12] investigated the limits of SPC for discrete memoryless channels, while the authors in [13] extended this analysis to different types of channels, including the AWGN channel and the Gilbert-Elliot channel. SPC for parallel Gaussian channels was analysed in [11], while in [14] an asymptotic analysis based on the Laplace integral was provided for the AWGN channel, parallel AWGN channels, and the binary symmetric channel (BSC). In [15], the authors investigated the maximum achievable rate for SPC over quasi-static multiple-input multiple-output fading channels. The results in [11]–[15] motivated the investigation of resource allocation design for SPC. In particular, optimal power allocation in a multi-user time division multiple access (TDMA) URLLC system was considered in [16]–[18]. In [19], the energy efficiency is maximized by optimizing the antenna configuration, bandwidth allocation, and power control under latency and reliability constraints. In [20], a cross-layer framework based on the effective bandwidth was proposed for optimal resource allocation under QoS constraints. The authors in [21] studied the joint uplink and downlink transmission design for URLLC in MISO systems. In [22], [23], the authors studied a hybrid automatic repeat request (HARQ) scheme for URLLC systems. However, the above works [16]–[24] assumed single carrier transmission which suffers from poor spectrum utilization and requires complex equalization at the receiver. Moreover, the optimization algorithms proposed in [20], [21] are based on a simplified version of the general expression for the achievable rate of SPC [13]. Thus, the optimal resource allocation for MISO OFDMA-URLLC systems is still an open problem.

In this paper, we study the resource allocation algorithm design for broadband downlink MISO OFDMA-URLLC systems, where a BS equipped with multiple antennas serves single antenna URLLC users. This paper makes the following main contributions:

- We propose a novel resource allocation algorithm design for multi-user MISO OFDMA-URLLC systems. The resource allocation algorithm design is formulated as an optimization problem for maximization of the weighted sum throughput subject to QoS constraints for the URLLC users. The QoS constraints include the minimum number of transmitted bits, the maximum packet error probability, and the maximum time for

transmission of a packet, i.e., the maximum delay[1].

- The formulated optimization problem is a non-convex mixed-integer problem which is difficult to solve. However, we transform the problem into the canonical form of a monotonic optimization problem. This reformulation allows the application of the polyblock outer approximation method to find the global optimal solution.

- To strike a balance between computational complexity and performance, we develop a low-complexity sub-optimal algorithm based on difference of convex programming and successive convex approximation to obtain a local optimal solution.

- Computer simulations show that the proposed sub-optimal algorithm closely approaches the performance of the optimal algorithm, despite its significantly lower complexity. Furthermore, both algorithms achieve significant performance gains compared to two baseline schemes, especially if the users have heterogeneous delay requirements, as is expected for Internet-of-Things applications [10]. Moreover, our results reveal that deploying multiple antennas is instrumental for achieving low latency and high reliability in URLLC systems.

We note that this paper expands the corresponding conference version [1] in several directions. First, in [1], resource allocation for single-antenna transceivers was considered, whereas in this paper, we study a system with a multiple-antenna BS. Moreover, in this paper, we derive the *optimal* resource allocation policy for MISO OFDMA-URLLC systems, whereas only a sub-optimal algorithm was provided in [1]. Finally, unlike [1], in this paper, we present extensive simulation results to illustrate the impact of the various system parameters on the performance of the proposed resource allocations algorithms.

The remainder of this paper is organized as follows. In Section II, we present the considered system and channel models. In Section III, the proposed resource allocation problem is formulated. In Section IV, the optimal resource allocation algorithm is derived, whereas the low-complexity sub-optimal algorithm is provided in Section V. In Section VI, the performance of the proposed schemes is evaluated via computer simulations, and finally conclusions are drawn in Section VII.

Notation: In this paper, lower-case letters refer to scalar numbers, while bold lower and upper case letters denote vectors and matrices, respectively. $\log_2(\cdot)$ is the logarithm with base 2. $\text{Tr}(\mathbf{A})$ and $\text{Rank}(\mathbf{A})$ denote the trace and the rank of matrix $\mathbf{A}$, respectively. $\mathbf{A} \succeq 0$

---

[1]We note that the end-to-end (E2E) delay of data packet transmission comprises of various components including the transmission delay, queueing delay, propagation delay, and routing delay in the backhaul and core networks. In this work, we focus on the transmission delay, which is independent of the other components of the E2E delay.

indicates that matrix $\mathbf{A}$ is positive semi-definite. $\mathbf{A}^H$ and $\mathbf{A}^T$ denote the Hermitian transpose and the transpose of matrix $\mathbf{A}$, respectively. $\mathbb{R}_+$ denotes the set of non-negative real numbers. $\mathbb{C}$ is the set of complex numbers. $\mathbf{I}_N$ is the $N \times N$ identity matrix. $\mathbb{H}_N$ denotes the set of all $N \times N$ Hermitian matrices. $|\cdot|$ and $\|\cdot\|$ refer to the absolute value of a complex scalar and the Euclidean vector norm, respectively. The circularly symmetric complex Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu, \sigma^2)$, and $\sim$ stands for "distributed as". $\mathcal{E}\{\cdot\}$ denotes statistical expectation. $\nabla_x f(\mathbf{x})$ denotes the gradient vector of function $f(\mathbf{x})$ and its elements are the partial derivatives of $f(\mathbf{x})$. $\mathbf{u}_d$ is the unit vector whose d-th entry is equal to 1 and all other entries are equal to 0. For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+$, $\mathbf{x} \leq \mathbf{y}$ means $x_i \leq y_i$, $\forall i$, where x_i and y_i are the i-th elements of $\mathbf{x}$ and $\mathbf{y}$, respectively.

II. SYSTEM AND CHANNEL MODELS

In this section, we present the system and channel models adopted for MISO OFDMA-URLLC in this paper.

A. System Model

We consider a single-cell downlink OFDMA system, where a BS equipped with N_T antennas serves K single-antenna URLLC users[2] indexed by $k = \{1, \ldots, K\}$, cf. Fig. 1(a). The frequency band is divided into M orthogonal sub-carriers indexed by $m \in \{1, \ldots, M\}$. We assume that a resource frame has a duration of T_f seconds, and consists of N time slots[3] which are indexed by $n \in \{1, \ldots, N\}$. Thereby, one OFDMA symbol spans one time slot, and in total $M \times N$ resource elements are available for assignment to the K users, cf. Fig. 1(b). We assume that the delay requirements of all users are known at the BS and only users whose delay requirements can potentially be met in the current resource block are admitted into the system. The maximum transmit power of the BS is P_max.

B. Channel Model

In this paper, we assume that the coherence time is larger than T_f. Therefore, the channel gain for a given sub-carrier and a given transmit antenna remains constant for the considered N time slots. The received signal at user k on sub-carrier m in time slot n is given as follows:

---

[2] The URLLC users are assumed to employ a single antenna to ensure low hardware complexity.

[3] In current standards such as LTE, a typical sub-carrier bandwidth is 15 kHz which leads to an OFDM symbol duration of $T_s = 66$ μs. Therefore, to meet a URLLC delay requirement of 1 ms, N has be smaller than 7. For larger sub-carrier spacing, larger values of N are possible.
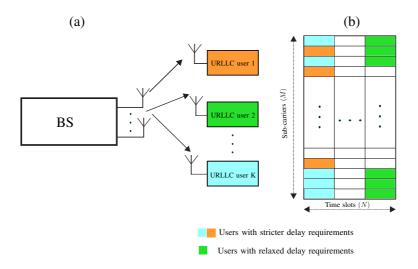
Figure 1. Multi-user downlink MISO OFDMA-URLLC: (a) System model with N_T-antenna BS and K single-antenna users; (b) Frame structure.

$$y_k[m,n] = \mathbf{h}_k^H[m]\mathbf{x}[m,n] + w_k[m,n], \tag{1}$$

where $\mathbf{h}_k[m] \in \mathbb{C}^{N_T \times 1}$ is the channel vector from the BS to user k on sub-carrier m, $\mathbf{x}[m,n] \in \mathbb{C}^{N_T \times 1}$ is the signal vector transmitted by the BS on sub-carrier m in time slot n. Moreover, $w_k[m,n] \sim \mathcal{CN}(0,\sigma^2)$ is the complex AWGN[4]. In this paper, we consider linear transmit precoding at the BS, where each user is assigned a beamforming vector. Hence, the transmit signal of the BS on sub-carrier m in time slot n is given by:

$$\mathbf{x}[m,n] = \sum_{k=1}^{K} \mathbf{w}_k[m,n] u_k[m,n], \tag{2}$$

where $u_k[m,n] \in \mathbb{C}$ and $\mathbf{w}_k[m,n] \in \mathbb{C}^{N_T \times 1}$ are the transmit symbol and the beamforming vector of user k on sub-carrier m in time slot n, respectively. Moreover, without loss of generality, we assume that $\mathcal{E}\{|u_k[m,n]|^2\} = 1, \ \forall k \in \{1,\dots,K\}$. By substituting (2) into (1), the received signal at user k on sub-carrier m in time slot n is given by:

$$
\begin{aligned}
y_k[m,n] &= \mathbf{h}_k^H[m]\left(\sum_{l=1}^{K}\mathbf{w}_l[m,n]u_l[m,n]\right) + w_k[m,n] \\
&= \underbrace{\mathbf{h}_k^H[m]\mathbf{w}_k[m,n]u_k[m,n]}_{\text{desired signal}} + \underbrace{\sum_{l\neq k}\mathbf{h}_k^H[m]\mathbf{w}_l[m,n]u_l[m,n]}_{\text{multi-user interference (MUI)}} + w_k[m,n].
\end{aligned}
\tag{3}
$$

[4]Without loss of generality, we assume that the noise variances for all URLLC users are identical.

Moreover, the signal-to-interference-plus-noise-ratio (SINR) of user k on sub-carrier m in time slot n is given as follows:

$$\gamma_k[m,n] = \frac{|\mathbf{h}_k^H[m]\mathbf{w}_k[m,n]|^2}{\sum_{l \neq k}|\mathbf{h}_k^H[m]\mathbf{w}_l[m,n]|^2 + \sigma^2}. \tag{4}$$

In this paper, we treat the interference caused by other users as noise. Moreover, to obtain a performance upper bound for MISO OFDMA-URLLC systems, for resource allocation, perfect channel state information (CSI) is assumed to be available at the BS.

III. RESOURCE ALLOCATION PROBLEM FORMULATION

In this section, we discuss the achievable rate for SPC, the QoS requirements of the URLLC users, and the adopted system performance metric for resource allocation algorithm design. Furthermore, we formulate the proposed resource allocation optimization problem for MISO OFDMA-URLLC systems.

A. Achievable Rate for SPC

Shannon's capacity theorem, on which most conventional resource allocation designs are based, applies to the asymptotic case where the packet length approaches infinity and the decoding error probability goes to zero [10]. Thus, it cannot be used for resource allocation design for URLLC systems, as URLLC systems have to employ short packets to achieve low latency, which also makes decoding errors unavoidable.

For performance evaluation of SPC, the so-called normal approximation for finite block-length codes was developed in [11]. Mathematically, the maximum number of bits B conveyed in a packet comprising L symbols can be approximated as [11, Eq. (4.277)], [14, Fig. 1]:

$$B = \sum_{i=1}^{L} \log_2(1+\gamma_i) - Q^{-1}(\epsilon)\sqrt{\sum_{i=1}^{L} V_i}, \tag{5}$$

where ϵ is the decoding packet error probability, V_i is the channel dispersion, and $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function which is given by $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-\frac{t^2}{2}\right)\mathrm{d}t$. For the complex AWGN, the channel dispersion is given by [11]

$$V_i = a^2\left(1 - (1+\gamma_i)^{-2}\right), \tag{6}$$

where γ_i is the SINR of the i-th received symbol and $a = \log_2(\mathrm{e})$.

In this paper, we base the resource allocation algorithm design for downlink MISO OFDMA-URLLC systems on (5). Each resource element carries one symbol, and by allocating several resource elements from the available $L = M \times N$ resource elements to a given user, the

number of bits received by the user with packet error probability ϵ can be determined based on (5).

B. QoS and System Performance Metric

The QoS requirements of URLLC users include the minimum number of received bits, B_k, the target packet error probability, ϵ_k, and the maximum number of time slots available for transmission of the user's packet, D_k. According to (5), the total number of bits transmitted over the resources allocated to user k can be written as:

$$\Psi_k(\mathbf{w}_k) = F_k(\mathbf{w}_k) - V_k(\mathbf{w}_k), \tag{7}$$

where

$$F_k(\mathbf{w}_k) = \sum_{m=1}^{M}\sum_{n=1}^{N}\log_2(1 + \gamma_k[m,n]), \tag{8}$$

$$V_k(\mathbf{w}_k) = Q^{-1}(\epsilon_k)\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}V_k[m,n]}, \tag{9}$$

where the channel dispersion $V_k[m,n]$ is given by:

$$V_k[m,n] = a^2\left(1 - (1 + \gamma_k[m,n])^{-2}\right). \tag{10}$$

Furthermore, $\mathbf{w}_k$ is the collection of all beamforming vectors $\mathbf{w}_k[m,n]$, $\forall m,n$, of user k.

The delay requirements of user k can be met by assigning all symbols of user k to the first D_k time slots. In other words, users requiring low latency are assigned resource elements at the beginning of the frame, cf. Fig. 1(b). We note that a user can start decoding as soon as it has received all OFDMA symbols that contain its data, i.e., after D_k time slots.

In order to be able to control the fairness among the URLLC users, we adopt the weighted sum throughput as performance metric. In particular, the weighed sum throughput of the entire system is defined as:

$$U(\mathbf{w}) = \sum_{k=1}^{K}\mu_k\Psi_k(\mathbf{w}_k) = F(\mathbf{w}) - V(\mathbf{w}), \tag{11}$$

where

$$F(\mathbf{w}) = \sum_{k=1}^{K}\mu_k F_k(\mathbf{w}_k), \quad V(\mathbf{w}) = \sum_{k=1}^{K}\mu_k V_k(\mathbf{w}_k), \tag{12}$$

and μ_k is the weight assigned to user k. Larger values of μ_k give a user a higher priority

and, as a result, a higher throughput (i.e., more bits are transmitted to the user) compared to the other users. The value of the μ_k may be specified in the medium access control (MAC) layer and is assumed to be given in the following. Moreover, $\mathbf{w}$ is the collection of the beamforming vectors $\mathbf{w}_k$ of all users.

C. Optimization Problem Formulation

In the following, we formulate a resource allocation optimization problem for maximization of the weighted sum throughput of the system subject to the QoS requirements of each user regarding the received number of bits, the reliability, and the latency. In particular, the proposed resource allocation policies are determined by solving the following optimization problem:

$$\underset{\mathbf{w}}{\text{maximize}} \ F(\mathbf{w}) - V(\mathbf{w}) \tag{13}$$

$$\text{s.t. C1: } F_k(\mathbf{w}_k) - V_k(\mathbf{w}_k) \geq B_k, \ \forall k,$$

$$\text{C2: } \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \|\mathbf{w}_k[m,n]\|^2 \leq P_{\max},$$

$$\text{C3: } \mathbf{w}_k[m,n] = 0, \quad \forall n > D_k, \forall k.$$

In (13), constraint C1 guarantees the transmission of a minimum number of B_k bits to user k. Constraint C2 is the total power budget constraint of the BS. Finally, constraint C3 ensures that user k is served within the first D_k time slots to meet its delay requirements. The problem in (13) is a non-convex optimization problem. The non-convexity is caused by the form of the SINR in (4) and the non-convex normal approximation in (5) which appear in the cost function and constraint C1.

Remark 1. Resource allocation algorithm design for conventional, non-URLLC OFDMA systems is typically based on Shannon's capacity formula, i.e., $V(\mathbf{w})$ and $V_k(\mathbf{w}_k)$ in (13) are absent [6]–[9]. The presence of $V(\mathbf{w})$ and $V_k(\mathbf{w}_k)$ makes problem (13) significantly more challenging to solve but is essential to capture the characteristics of OFDMA-URLLC systems.

There is no systematic approach to solving general non-convex problems optimally. However, in Section IV, we will show that based on a sequence of transformations, problem (13) can be solved optimally, by employing monotonic optimization. Moreover, in Section V, we develop a sub-optimal algorithm based on successive convex approximation and difference of convex programming to obtain close-to-optimal performance with low computational

complexity.

IV. OPTIMAL SOLUTION OF THE OPTIMIZATION PROBLEM

In this section, we solve the optimization problem in (13) optimally based on monotonic optimization [25], which leads to an iterative resource allocation algorithm, where a semi-definite relaxation (SDR) problem is solved in each iteration.

A. Semi-Definite Programming Relaxation

To facilitate the application of semi-definite programming (SDP), we define new variables $\mathbf{W}_k[m, n] = \mathbf{w}_k[m, n]\mathbf{w}_k^H[m, n]$ and $\mathbf{H}_k[m] = \mathbf{h}_k[m]\mathbf{h}_k^H[m]$, $\forall k, m, n$, and rewrite (13) in equivalent form as follows:

$$\underset{\mathbf{W}}{\text{maximize}} \; F(\mathbf{W}) - V(\mathbf{W}) \tag{14}$$

$$\text{s.t. C1: } F_k(\mathbf{W}_k) - V_k(\mathbf{W}_k) \geq B_k, \; \forall k,$$

$$\text{C2: } \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \text{Tr}(\mathbf{W}_k[m, n]) \leq P_{\max},$$

$$\text{C3: } \text{Tr}(\mathbf{W}_k[m, n]) = 0, \forall n > D_k, \forall k,$$

$$\text{C4: } \mathbf{W}_k[m, n] \succeq 0, \quad \forall k, m, n,$$

$$\text{C5: } \text{Rank}(\mathbf{W}_k[m, n]) \leq 1, \quad \forall k, m, n,$$

where

$$F(\mathbf{W}) = \sum_{k=1}^{K} \mu_k \sum_{m=1}^{M} \sum_{n=1}^{N} \log_2 \left(1 + \gamma_k[m, n]\right), \tag{15}$$

$$V(\mathbf{W}) = \sum_{k=1}^{K} \mu_k a Q^{-1}(\epsilon_k) \sqrt{\sum_{m=1}^{M} \sum_{n=1}^{N} \left(1 - (1 + \gamma_k[m, n])^2\right)}, \tag{16}$$

and

$$\gamma_k[m, n] = \frac{\text{Tr}(\mathbf{H}_k[m]\mathbf{W}_k[m, n])}{\sum_{l \neq k} \text{Tr}(\mathbf{H}_k[m]\mathbf{W}_l[m, n]) + \sigma^2}. \tag{17}$$

We note that $\mathbf{W}_k[m, n] \succeq 0$ and $\text{Rank}(\mathbf{W}_k[m, n]) \leq 1, \forall k, m, n$, in constraints C4 and C5 are imposed to ensure that $\mathbf{W}_k[m, n] = \mathbf{w}_k[m, n]\mathbf{w}_k^H[m, n]$ holds after optimization. Moreover, for simplicity of notation, we define $\mathbf{W}_k$ as the collection of all optimization variables $\mathbf{W}_k[m, n], \forall m, n$, and $\mathbf{W}$ as the collection of all $\mathbf{W}_k, \forall k$.

B. Problem Transformation

The objective function and constraint C1 in (14) have a complicated structure. To handle this complexity and to facilitate the application of monotonic optimization, we introduce a set of auxiliary variables $z_k[m,n], \forall k, m, n$, to bound the SINR from below, i.e.,

$$0 \leq z_k[m,n] \leq \gamma_k[m,n] = \frac{f_k[m,n](\mathbf{W})}{g_k[m,n](\mathbf{W})}, \forall k, m, n, \quad (18)$$

where $f_k[m,n](\mathbf{W})$ and $g_k[m,n](\mathbf{W})$ are the numerator and denominator of the SINR in (17) and are given respectively by

$$f_k[m,n](\mathbf{W}) = \text{Tr}(\mathbf{H}_k[m]\mathbf{W}_k[m,n]), \forall k, m, n, \quad (19)$$

$$g_k[m,n](\mathbf{W}) = \sum_{l \neq k} \text{Tr}(\mathbf{H}_k[m]\mathbf{W}_l[m,n]) + \sigma^2, \forall k, m, n. \quad (20)$$

Let us replace $\gamma_k[m,n]$ by $z_k[m,n]$ in $F(\mathbf{W})$, $V(\mathbf{W})$, $F_k(\mathbf{W}_k)$, and $V_k(\mathbf{W}_k)$ and denote the resulting functions by $F(\mathbf{z})$, $V(\mathbf{z})$, $F_k(\mathbf{z}_k)$, and $V_k(\mathbf{z}_k)$, respectively, i.e.,

$$F(\mathbf{z}) = \sum_{k=1}^{K} \mu_k F_k(\mathbf{z}_k), \quad (21)$$

$$V(\mathbf{z}) = \sum_{k=1}^{K} \mu_k V(\mathbf{z}_k), \quad (22)$$

$$F_k(\mathbf{z}_k) = \sum_{m=1}^{M} \sum_{n=1}^{N} \log_2(1 + z_k[m,n]), \forall k, \quad (23)$$

$$V(\mathbf{z}_k) = aQ^{-1}(\epsilon_k)\sqrt{\sum_{m=1}^{M} \sum_{n=1}^{N} (1 - (1 + z_k[m,n])^{-2})}, \quad (24)$$

where $\mathbf{z}_k$ denotes the collection of optimization variables $z_k[m,n]$, $\forall m, n$, and $\mathbf{z}$ denotes the collection of optimization variables $\mathbf{z}_k, \forall k$. Using these notations, and after dropping rank constraint C5 in (14), we formulate a new optimization problem as follows:

$$\underset{\mathbf{W}, \mathbf{z}}{\text{maximize}} \; F(\mathbf{z}) - V(\mathbf{z}) \quad (25)$$

$$\text{s.t. C1: } F_k(\mathbf{z}_k) - V_k(\mathbf{z}_k) \geq B_k, \forall k,$$

$$\text{C2-C4,}$$

$$\text{C6: } z_k[m,n] \leq \frac{f_k[m,n](\mathbf{W})}{g_k[m,n](\mathbf{W})}, \forall k, m, n,$$

$$\text{C7: } z_k[m,n] \geq 0.$$

In the following, we first find an optimal solution for problem (25). Subsequently, we prove that problems (25) and (14) are equivalent, cf. **Proposition** 1. Hence, the solution obtained for problem (25) constitutes an optimal solution for problem (14), too.

The main condition required for applying monotonic optimization is the monotonicity of the objective function and the constraints. We note that the objective function and constraint C1 in (25) are differences of two monotonic concave functions in the optimization variables $\mathbf{z}$, cf. Appendix A. Hence, problem (25) can be transformed into the canonical form of a monotonic optimization problem in two steps:

- **Step 1:** To transform the objective function in (25) into a monotonic function, we note that the SINR in (18) is upper bounded by $z_{\max,k}[m,n]$ [5]:

$$z_k[m,n] \leq z_{\max,k}[m,n] \triangleq \frac{P_{\max}}{\sigma^2} \operatorname{Tr}(\mathbf{H}_k[m,n]), \forall k, m, n. \tag{26}$$

Let us define $\mathbf{z}_{\max}$ as the collection of all $z_{\max,k}[m,n]$. Since $V(\mathbf{z})$ is monotonically increasing in $\mathbf{z}$, $\mathbf{z} \leq \mathbf{z}_{\max}$ leads to $V(\mathbf{z}) \leq V(\mathbf{z}_{\max})$. Therefore, $V(\mathbf{z}) + t = V(\mathbf{z}_{\max})$ holds, for some positive t. Hence, substituting $V(\mathbf{z})$ by $V(\mathbf{z}_{\max}) - t$, the optimization problem in (25) can be rewritten as follows:

$$\underset{\mathbf{W}, \mathbf{z}, t}{\operatorname{maximize}} \ F(\mathbf{z}) + t - V(\mathbf{z}_{\max}) \tag{27}$$

$$\text{s.t.} \ \ \text{C1-C4, C6, C7,}$$

$$\text{C8: } t + V(\mathbf{z}) \leq V(\mathbf{z}_{\max}),$$

$$\text{C9: } t \geq 0.$$

We note that at the optimal point constraint C8 holds with equality due to the monotonicity of the objective function with respect to auxiliary optimization variable t.

- **Step 2:** We use a similar approach as for transforming the cost function to transform constraint C1 into a standard monotonic constraint. In particular, $V_k(\mathbf{z}_k) + \zeta_k = V_k(\mathbf{z}_{\max,k})$ holds for some positive auxiliary optimization variable ζ_k, where $\mathbf{z}_{\max,k}$ is the collection of the $z_{\max,k}, \forall m, n$. Therefore, by substituting $V_k(\mathbf{z}_k)$ by $V_k(\mathbf{z}_{\max,k}) - \zeta_k$, constraint C1 can be transformed into two monotonic constraints as follows:

$$\text{C1a:} \quad F_k(\mathbf{z}_k) + \zeta_k \geq V_k(\mathbf{z}_{\max,k}) + B_k, \forall k, \tag{28}$$

---

[5]The right hand side of (26) is obtained by allocating all available power $P_{\max}$ to time slot n, sub-carrier m, and user k.

$$\text{C1b:} \quad V_k(\mathbf{z}_k) + \zeta_k \leq V_k(\mathbf{z}_{\text{max},k}), \forall k. \tag{29}$$

We note that the left hand sides of (28) and (29) are monotonically increasing functions. Hence, problem (27) has been transformed to an equivlant monotonic optimization problem as follows:

$$\underset{\mathbf{W},\mathbf{z},t,\boldsymbol{\zeta}}{\text{maximize}} \ F(\mathbf{z}) + t \tag{30}$$
$$\text{s.t. C1a, C1b, C2-C4, C6-C9,}$$

where $\boldsymbol{\zeta}$ is the collection of optimization variables $\zeta_k, \ \forall k$. Note that, in (30), we removed the constant $V(\mathbf{z}_{\text{max}})$ from the objective function, because it has no effect on the optimal solution. Optimization problem (30) has a monotonically increasing objective function and all constraints are monotonically increasing functions (C1b, C6, C8) or convex functions (C1a, C2-C4, C7, C9). Therefore, (30) belongs to the class of monotonic optimization problems [26], [27], which can be solved using algorithms such as outer polyblock approximation. To facilitate the presentation of the proposed solution, we rewrite the problem (30) in the canonical form of a monotonic optimization problem as follows:

$$\underset{\mathbf{W},\mathbf{z},t,\boldsymbol{\zeta}}{\text{maximize}} \ F(\mathbf{z}) + t \tag{31}$$
$$\text{s.t. } (\mathbf{W}, \mathbf{z}, t, \zeta) \in \mathcal{V},$$

where the feasible set $\mathcal{V} = \mathcal{G} \cap \mathcal{H}$ is the intersection of the normal set $\mathcal{G}$ and the co-normal set $\mathcal{H}$ [28]. The normal set $\mathcal{G}$ is given by:

$$\mathcal{G} = \left\{ (\mathbf{z},t)|0 \leq z_k[m,n] \leq \frac{f_k[m,n](\mathbf{W})}{g_k[m,n](\mathbf{W})}, \forall k,m,n, \ \mathbf{z} \in \mathcal{Z}, \mathbf{W} \in \mathcal{W} \right\}, \tag{32}$$

with $\mathcal{Z}$ and $\mathcal{W}$ being the feasible set spanned by constraints (C1b, C2-C4, C6-C9). The co-normal set $\mathcal{H}$ is defined by constraint C1a. Now, we are ready to design the optimal resource allocation algorithm based on the polyblock outer approximation algorithm [26].

C. Polyblock Algorithm

Due to the monotonicity of the objective function and the constraints, the optimal solution of optimization problem (31) is at the boundary of the feasible set $\mathcal{V}$ [28]–[30]. However, the boundary of the feasible set is unknown. Thus, we approach the boundary from above by enclosing the feasible set $\mathcal{V}$ by an initial polyblock $\mathcal{B}^{(1)}$ with an initial vertex $\mathbf{v}^{(1)} = \left(\mathbf{z}_1^{(1)}, t_1^{(1)}\right)$ as shown in Fig. 2(a), where for simplicity of illustration, we consider a case
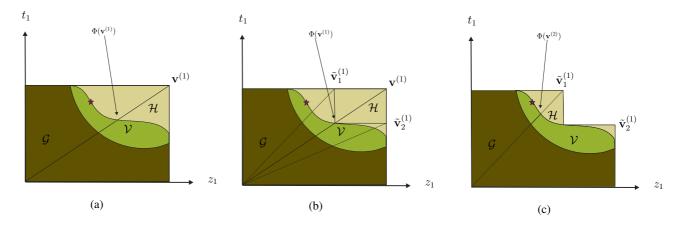
Figure 2. The polyblock outer approximation algorithm. $\mathcal{G}$ is the normal set, $\mathcal{H}$ is the co-normal set, and $\mathcal{V} = \mathcal{G} \cap \mathcal{H}$. $\Phi(\cdot)$ is the intersection point between the boundary and the line connected a vertex with the origin. The star is the optimal solution located at the boundary of the feasible set $\mathcal{V}$.

with only two dimensions t_1 and z_1 to depict the polyblock algorithm. Subsequently, the intersection point $\Phi(\cdot)$ between the vertex and the origin is calculated, and the new polyblock $\mathcal{B}^{(2)}$ is now defined by three vertices $\mathbf{v}^{(1)}, \tilde{\mathbf{v}}^{(1)}$, and $\tilde{\mathbf{v}}^{(2)}$, see Fig. 2(b). Since, vertex $\mathbf{v}^{(1)}$ has no effect on polyblock $\mathcal{B}^{(2)}$, we can remove it, see Fig. 2(c). This process is continued until the feasible set $\mathcal{V}$ is enclosed by a final polyblock $\mathcal{B}^{(1)} \supset \mathcal{B}^{(2)} \supset \cdots \supset \mathcal{V}$. Finally, we select the vertex that maximizes the objective function in (31). This procedures is summarized in **Algorithm** 1. **Algorithm** 1 requires the calculation of intersection point $\Phi(\cdot)$ in each iteration, which is performed by **Algorithm** 2, as explained in the following.

D. Calculation of Intersection Point

In each iteration of **Algorithm** 1, we determine the intersection of the line connecting $\mathbf{v}^{(j)}$ and the origin with the boundary of the feasible set, cf. Fig. 2(a). In other words, we have to find a $\lambda > 0$ which satisfies $\Phi(\mathbf{v}^{(j)}) = \lambda \mathbf{v}^{(j)}$. λ can be obtained based on the following optimization problem:

$$\text{maximize} \quad \lambda \tag{33}$$
$$\text{s.t.} \quad \lambda \mathbf{v}^{(j)} \in \mathcal{V}.$$

To solve (33), we use the bisection method which is formally presented in **Algorithm** 2. In particular, in line 5 of **Algorithm** 2, we solve the following SDP problem:

$$\underset{\mathbf{W}, \zeta}{\text{maximize}} \, 1 \tag{34}$$

Algorithm 1 Polyblock Outer Approximation Algorithm

1: Initialize polyblock $\mathcal{B}^{(1)}$ with vertex set $\mathbf{v}^{(1)} = \{\mathbf{z}^{(1)}, t^{(1)}\}$, where the elements of $\mathbf{z}^{(1)}$ are $z_k^{(1)}[m,n] = \text{Tr}(\mathbf{H}_k[m])P_{\max}, \forall k, m, n$, and $t^{(1)} = V(\mathbf{z}_{\max})$,

2: Set error tolerance $\rho \ll 1$ and iteration index $j = 1$.

3: **Repeat**{Main Loop}

4: Construct a smaller polyblock $\mathcal{B}^{(j+1)}$ with vertex set $\mathbf{v}^{(j+1)}$ by replacing $\mathbf{v}^{(j)}$ with $D = K \times M \times N + 1$ new vertices $\{\tilde{\mathbf{v}}_1^{(j)}, \ldots, \tilde{\mathbf{v}}_D^{(j)}\}$, $d \in \{1, \ldots, D\}$. The new vertex $\tilde{\mathbf{v}}_d^{(j)}$ is generated as
$$\tilde{\mathbf{v}}_d^{(j)} = \mathbf{v}^{(j)} - \left(v_d^{(j)} - \phi_d(\mathbf{v}^{(j)})\right)\mathbf{u}_d,$$
where $v_d^{(j)}$ and $\phi_d(\mathbf{v}^{(j)})$ are the d-th elements of $\mathbf{v}^{(j)}$ and $\Phi(\mathbf{v}^{(j)})$, respectively. $\Phi(\mathbf{v}^{(j)})$ is obtained by **Algorithm** 2.

5: Find $\mathbf{v}^{(j+1)}$ as that vertex of $\mathcal{V}^{(j+1)}$ whose intersection maximizes the objective function of problem
$$\mathbf{v}^{(j+1)} = \underset{\mathbf{v} \in \mathcal{V}^{(j+1)}}{\arg\max}\{F(\mathbf{z}) + t\},$$

6: Set $j = j + 1$

7: **until** $\frac{\|\mathbf{v}^{(j)} - \Phi(\mathbf{v}^{(j)})\|}{\|\mathbf{v}^{(j)}\|} \leq \rho$

8: Optimal vertex $\mathbf{v}^* = \Phi(\mathbf{v}^{(j)})$ and optimal beamforming matrix $\mathbf{W}^*$ are obtained when calculating $\Phi(\mathbf{v}^{(j)})$ with **Algorithm** 2.

Algorithm 2 Optimal Intersection Algorithm via Bisection Method

---

1: Initialize feasible set $\mathcal{V}$, vertex $\mathbf{v}^{(j)} = \{\mathbf{z}^{(j)}, t^{(j)}\}$. $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$

2: Set error tolerance $\delta \ll 1$.

3: **while** $(\lambda_{\max} - \lambda_{\min}) \geq \delta$ **do**

4: $\quad \lambda = (\lambda_{\max} + \lambda_{\max})/2$.

5: $\quad$ Check the feasibility of problem (34) using, e.g., CVX, and check if $\lambda t^{(j)} + V(\lambda\mathbf{z}^{(j)}) \leq V(\mathbf{z}_{\max}), 0 \leq \lambda t^{(j)} \leq V(\mathbf{z}_{\max})$.

6: $\quad$ **if** the two conditions in line 5 are satisfied **then**

7: $\quad\quad$ set $\lambda = \lambda_{\min}$

8: $\quad$ **else**

9: $\quad\quad$ set $\lambda = \lambda_{\max}$

10: $\quad$ **end if**

11: **end while**

12: $\lambda = \lambda_{\min}$, $\Phi(\mathbf{v}^{(j)}) = \lambda\mathbf{v}^{(j)}$.

$$\text{s.t. C1a: } F_k(\lambda\mathbf{z}_k^{(j)}) + \zeta_k \geq V_k(\mathbf{z}_{\max,k}) + B_k, \forall k,$$

$$\text{C1b: } V_k(\lambda\mathbf{z}_k^{(j)}) + \zeta_k \leq V_k(\mathbf{z}_{\max,k}), \forall k,$$

$$\text{C2-C4, C7,}$$

$$\text{C6: } (\lambda z_k^{(j)}[m,n])g_k[m,n](\mathbf{W}) - f_k[m,n](\mathbf{W}) \leq 0, \forall k, m, n.$$

Problem (34) is a convex optimization problem which can be solved using standard optimization software tools such as CVX [31]. Moreover, the tightness of the applied SDR is revealed in the following theorem.

Theorem 1. The optimal $\mathbf{W}_k[m,n], \forall k, m, n$, as the solution of (34) has a rank less than or equal to one, i.e., $\mathrm{Rank}(\mathbf{W}_k[m,n]) \leq 1, \forall k, m, n$.

Proof. Please refer to Appendix B. ∎

Proposition 1. Optimization problems (25) and (14) are equivalent in the sense that they yield the same solution for the beamforming matrix $\mathbf{W}_k[m,n], \forall k, m, n$.

Proof. The solution of (25) is the same as that of (14) if i) constraint C6 in (25) holds with equality and ii) $\mathbf{W}_k[m,n], \forall k, m, n$, obtained from (25) has rank smaller than or equal to one. Problem (25) is solved with **Algorithm** 1 where in each iteration problem (34) is solved. In Theorem 1, we showed that the $\mathbf{W}_k[m,n], \forall k, m, n$, obtained from (34) have rank equal to or smaller than one. This implies that the solution of (25) has also rank equal to or smaller than one, i.e., condition ii) holds. Moreover, in Section IV-B, we showed that (25) is a monotonic optimization problem. This implies that the optimal solution lies on the boundary of the feasible set of (25). As a consequence, constraint C6 in (25) has to hold with equality, i.e., $z_k[m,n] = \gamma_k[m,n], \forall k, m, n$. Hence, condition i) is also satisfied. This completes the proof. ∎

The computational complexity of the optimal scheme is exponential in the number of vertices, D, used in each iteration. Nevertheless, the obtained global solution constitutes a valuable performance upper bound for any sub-optimal resource allocation algorithm. In the next section, we propose a sub-optimal resource allocation algorithm which has polynomial time computational complexity and yields close-to-optimal performance.

V. LOW-COMPLEXITY RESOURCE ALLOCATION ALGORITHM

In this section, we propose a low-complexity resource allocation algorithm based on penalized successive convex approximation providing a locally optimal solution of optimization problem (13).

A. Difference of Convex Programming

In this subsection, we solve optimization problem (25), as (25) is equivalent to (13). We solve optimization problem (25) in two steps. First, we transform the problem into the canonical form needed for application of difference of convex programming. Second, we

apply a Taylor series expansion to obtain a convex approximation of the non-convex terms. As a result, we obtain a convex optimization problem that can be efficiently solved using convex optimization software. In the following, we explain these two steps in detail.

Step 1: We note that non-convex constraint C6 in (25) can be rewritten as follows:

$$\text{C6: } z_k[m,n]g_k[m,n](\mathbf{W}) = z_k[m,n](I_k[m,n](\mathbf{W}) + \sigma^2) \le f_k[m,n](\mathbf{W}), \ \forall k, m, n, \quad (35)$$

where $g_k[m,n] = I_k[m,n](\mathbf{W}) + \sigma^2$. We note that $z_k[m,n]I_k[m,n](\mathbf{W})$ in (35) is a bilinear term which is non-convex. In fact, the Hessian matrix of a bilinear function is neither positive nor negative semi-definite. Thus, bilinear functions are neither convex nor concave in general, which is an obstacle for designing computationally efficient resource allocation algorithms. The product of two convex function $f_1(x)$ and $f_2(x)$ can be written as a difference of two convex functions as follows [32]:

$$f_1(x)f_2(x) = 0.5(f_1(x) + f_2(x))^2 - 0.5f_1(x)^2 - 0.5f_2(x)^2. \quad (36)$$

Exploiting (36) with $z_k[m,n]$ and $I_k[m,n](\mathbf{W})$ as f_1 and f_2, respectively, we can express the product term $z_k[m,n]I_k[m,n](\mathbf{W})$ in (35) as follows:

$$z_k[m,n]I_k[m,n](\mathbf{W}) = Q(z_k[m,n], \mathbf{W}) - T(z_k[m,n], \mathbf{W}), \quad (37)$$

where

$$Q(z_k[m,n], \mathbf{W}) = \frac{1}{2}(z_k[m,n] + I_k[m,n](\mathbf{W}))^2, \forall k, m, n, \quad (38)$$

$$T(z_k[m,n], \mathbf{W}) = \frac{1}{2}(z_k[m,n])^2 + \frac{1}{2}(I_k[m,n](\mathbf{W}))^2, \forall k, m, n. \quad (39)$$

Furthermore, substituting (37) into (35), we obtain an equivalent representation for constraint C6 in (35) as follows:

$$\text{C6: } Q(z_k[m,n], \mathbf{W}) - T(z_k[m,n], \mathbf{W}) \le f_k[m,n](\mathbf{W}) - \sigma^2 z_k[m,n], \forall k, m, n, \quad (40)$$

where the left hand side is a difference of two convex functions. Hence, optimization problem (25) can now be rewritten as follows:

$$\underset{\mathbf{W}, \mathbf{z}}{\text{minimize}} \ - [F(\mathbf{z}) - V(\mathbf{z})] \quad (41)$$

s.t. C1-C4, C7,

$$\text{C6: } Q(z_k[m,n], \mathbf{W}) - T(z_k[m,n], \mathbf{W}) \le f_k[m,n](\mathbf{W}) - \sigma^2 z_k[m,n], \forall k, m, n.$$

The optimization problem in (41) belongs to the class of difference of convex programming problems, since its objective function can be written as a difference of two convex functions and constraints C1 and C6 can also be expressed as the differences of two convex functions. In particular, functions $-F(\mathbf{z})$, $-V(\mathbf{z})$, $Q(z_k[m,n],\mathbf{W})$, and $T(z_k[m,n],\mathbf{W})$ are convex functions.

Step 2: To obtain a convex optimization problem that can be efficiently solved, we have to handle the non-convex objective function and non-convex constraints C1 and C6. To this end, we determine the first order approximations of functions $V_k(\mathbf{z}_k)$ and $T(z_k[m,n],\mathbf{W})$ using Taylor series as follows:

$$V_k(\mathbf{z}_k) \;\; \leq \bar{V}_k(\mathbf{z}_k) = \;\; V(\mathbf{z}_k^{(j)}) + \nabla_{\mathbf{z}_k} V_k(\mathbf{z}_k^{(j)})^T (\mathbf{z}_k - \mathbf{z}_k^{(j)}), \tag{42}$$

and

$$T(z_k[m,n],\mathbf{W}) \geq \bar{T}(z_k[m,n],\mathbf{W}) = T(z_k^{(j)}[m,n],\mathbf{W}^{(j)}) +$$
$$\nabla_{z_k[m,n]} T(z_k^{(j)}[m,n],\mathbf{W}^{(j)})(z_k[m,n] - z_k^{(j)}[m,n])$$
$$+ \mathrm{Tr}(\nabla_{\mathbf{W}} T(z_k^{(j)}[m,n],\mathbf{W}^{(j)})^T)(\mathbf{W} - \mathbf{W}^{(j)}), \forall k,m,n, \tag{43}$$

where $\mathbf{W}^{(j)}$, $\mathbf{z}_k^{(j)}$, and $z_k^{(j)}[m,n]$ are initial feasible points, and

$$\nabla_{\mathbf{z}_k} V_k(\mathbf{z}_k) = \frac{a^2 Q^{-1}(\epsilon_k)}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N} V_k[m,n]}} \begin{pmatrix} \frac{1}{(1+z_k[1,1])^3} \\ \frac{1}{(1+z_k[2,1])^3} \\ \vdots \\ \frac{1}{(1+z_k[M,N])^3} \end{pmatrix}, \tag{44}$$

$$\nabla_{z_k[m,n]} T(z_k[m,n],\mathbf{W}) = z_k[m,n], \tag{45}$$

and

$$\nabla_{\mathbf{W}} T(z_k[m,n],\mathbf{W}) = I_k[m,n](\mathbf{W})\mathbf{H}_k[m]. \tag{46}$$

The right hand sides of (42) and (43) are affine functions, and by substituting them in (41), we obtain the following convex optimization problem:

$$\underset{\mathbf{W},\mathbf{z}}{\text{minimize}} \;\; -[F(\mathbf{z}) - \bar{V}(\mathbf{z})] \tag{47}$$
$$\text{s.t. C1: } F_k(\mathbf{z}_k) - \bar{V}_k(\mathbf{z}_k) \geq B_k, \;\; \forall k,$$

C2- C4, C7,

C6: $Q(z_k[m,n], \mathbf{W}) - \bar{T}(z_k[m,n], \mathbf{W}) \leq f_k[m,n](\mathbf{W}) - z_k[m,n], \forall k, m, n.$

Optimization problem (47) can be efficiently solved by standard convex solvers such as CVX [31]. Problem (47) can be solved iteratively where the solution of (47) in iteration j is used as the initial point for the next iteration $j + 1$. The algorithm produces a sequence of improved feasible solutions until convergence to a local optimum point of problem (47) or equivalently problem (13) in polynomial time [33], [34]. Moreover, one can show that the solution to (47) yields a matrix that has a rank equal to or smaller than one, i.e., $\text{Rank}(\mathbf{W}_k[m,n]) \leq 1, \forall k, m, n$. The corresponding proof is similar to the one presented in Appendix B.

B. Penalized Successive Convex Approximation

In order to solve (47) using successive convex approximation, we require a feasible initial point that satisfies QoS constraint C1. Since it is not easy to find such initial feasible points, we propose a corresponding algorithm which is based on penalizing optimization problem (47) when the QoS is violated. The basic idea is to relax the considered problem by adding slack variables to constraint C1 and penalizing the sum of the violations of the constraints. Thereby, using this technique, optimization problem (47) can be rewritten in equivalent form as follows:

$$\underset{\mathbf{W}, \mathbf{z}, \boldsymbol{\tau}}{\text{minimize}} \quad -[F(\mathbf{z}) - \bar{V}(\mathbf{z})] + \beta^{(j)} \sum_{k=1}^{K} \tau_k \qquad (48)$$

$$\text{s.t. C1: } F_k(\mathbf{z}_k) - \bar{V}_k(\mathbf{z}_k) + \tau_k \geq B_k, \ \forall k,$$

$$\text{C2-C4, C6, C7,}$$

where $\beta^{(j)}$ is the penalizing weight in iteration j, $\tau_k, \forall k$, are slack variables, and $\boldsymbol{\tau}$ is the collection of slack variables $\tau_k, \forall k$. **Algorithm** 3 presents an iterative algorithm for solving (48). In the first iteration, by choosing a small penalty weight $\beta^{(1)} > 0$, we allow the QoS constraint to be violated such that the feasible set is large. Then, in each subsequent iteration j, we use the solution from the previous iteration as initial point, increase the penalty factor $\beta^{(j)}$, and solve the problem again. Thus, if a feasible point exists, continuing this iterative procedure eventually yields solutions where $\tau_k = 0, \forall k$, holds, i.e., (48) becomes equivalent to (47). Otherwise, if τ_k does not converge to zero, the original problem is not feasible. Moreover, a maximum value for the penalty weight $\beta_{\max}$ is imposed to avoid numerical

---

Algorithm 3 Penalized Successive Convex Approximation

1: Initialize: The maximum number of iterations $J_{\max}$, iteration index $j = 1$, initial points $\mathbf{W}^{(1)}$, $\mathbf{z}^{(1)}$, initial penalty factor $\beta^{(1)} \gg 1$, $\beta_{\max}$, $\eta > 1$.
2: **Repeat**
3: Solve convex problem (48) for a given $\mathbf{W}^{(j)}$ and $\mathbf{z}^{(j)}$ and store the intermediate resource allocation policy $\{\mathbf{W}, \mathbf{z}\}$
4: Set $j = j + 1$ and update $\mathbf{W}^{(j)} = \mathbf{W}$, $\mathbf{z}^{(j)} = \mathbf{z}$, and $\beta^{(j)} = \min(\eta\beta^{(j-1)}, \beta_{\max})$.
5: **Until** convergence or $j = J_{\max}$
6: $\mathbf{W}^* = \mathbf{W}^{(j)}$,

Table I
SYSTEM PARAMETERS USED IN SIMULATIONS.

Parameter	Value
Number and bandwidth of sub-carriers	64 and 15 kHz
Noise power density	-174 dBm/Hz
Number of bits per packet	160 bits
Maximum BS transmit power $P_{\max}$	45 dBm
Error tolerances ρ and δ for **Algorithms** 1 and 2	0.01
Penalty factors $\beta^{(1)}$, $\beta_{\max}$ for **Algorithm** 3	1000, 5000, 1.1
Value of η for **Algorithm** 3	1.5
Packet error probability for user k	$\epsilon_k = 10^{-6}$

issues.

VI. PERFORMANCE EVALUATION

In this section, we provide simulation results to evaluate the effectiveness of the proposed resource allocation design for MISO OFDMA-URLLC systems. We adopt the simulation parameters given in Table I, unless specified otherwise. In our simulations, a single cell is considered with inner and outer radius $r_1 = 50$ m and $r_2 = 250$ m, respectively. The BS is located at the centre of the cell. The path loss is calculated as $35.3 + 37.6 \log_{10}(d_k)$ [20], where d_k is the distance from the BS to user k. The small scale fading gains between the BS and the users are modelled as independent and identically Rayleigh distributed. For simplicity, all user weights are set to $\mu_k = 1, \forall k$. All simulation results are averaged over 1000 realizations of the path loss and multipath fading, unless specified otherwise.

A. Performance Metric

For performance evaluation of the system, we define the sum throughput of the system for a given channel realization as follows:

$$\bar{R} = \begin{cases} \frac{1}{MN} \sum_{k=1}^{K} \Psi_k(\mathbf{W}_k), & \text{if } \mathbf{W} \text{ is feasible} \\ 0 & \text{otherwise.} \end{cases} \tag{49}$$

If the optimization problem is infeasible for a given channel realization, we set the corresponding sum throughput to zero. The average system sum throughput is obtained by averaging $\bar{R}$ over all considered channel realizations.

B. Performance Bound and Benchmark Schemes

We compare the performance of the proposed resource allocation algorithm design with the following benchmark and baseline schemes[6]:
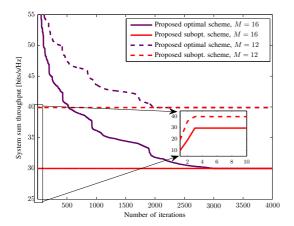
- **Upper bound**: To obtain an (unachievable) performance upper bound, Shannon's capacity formula is adopted in problem (13), i.e., $V(\mathbf{W})$ and $V_k(\mathbf{W}_k)$ are set to zero in the objective function and constraint C1, respectively, and all other constraints are retained. The resulting optimization problem is solved optimally and sub-optimally using modified versions of **Algorithms** 1 and 3, respectively. The corresponding sum throughput is obtained from (49) where $\Psi_k(\mathbf{W}_k) = F_k(\mathbf{W}_k)$ is used.

- **Baseline scheme 1:** For this scheme, as for the performance upper bound, the solution obtained for Shannon's capacity formula is used to compute the sum throughput. However, now $\Psi_k(\mathbf{W}_k) = F_k(\mathbf{W}_k)$ is used in (49), and $\Psi_k(\mathbf{W}_k) = F_k(\mathbf{W}_k) - V_k(\mathbf{W}_k) \geq B_k$ is used to check the feasibility of the solution.

- **Baseline scheme 2:** For this scheme, we employ maximum ratio transmission (MRT) beamforming, where $\mathbf{w}_k[m,n] = \sqrt{p_k[m,n]}\frac{\mathbf{h}_k[m]}{\|\mathbf{h}_k[m]\|}$, and optimize the powers $p_k[m,n]$. The resulting optimization problem is solved using successive convex approximation employing a similar approach as for deriving **Algorithm** 3.

C. Simulation Results

In this subsection, we illustrate the effectiveness of the proposed resource allocation algorithms for MISO OFDMA-URLLC systems via computer simulations.

Figs. 3 and 4 show the convergence of the proposed optimal (**Algorithm** 1) and sub-optimal (**Algorithm** 3) algorithms for different numbers of sub-carriers M and different numbers of users K. We show the system sum throughput as a function of the number of iterations for a given channel realization. As can be observed from Fig. 3, the proposed optimal scheme and the proposed low-complexity scheme converge to the global optimum solution after a finite number of iterations. However, the low-complexity scheme reaches the optimal point much faster than the optimal scheme. In particular, **Algorithm** 1 converges to

---

[6]We do not show simulation results for single-input single-output (SISO) OFDMA-URLLC systems to avoid overloading the figures. We refer interested readers to [1] for such results.
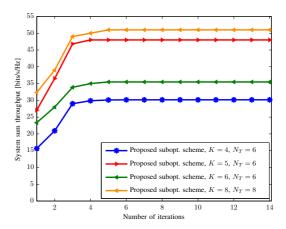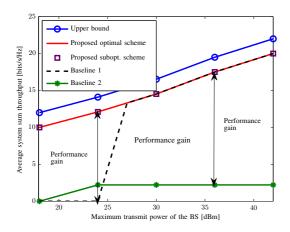
Figure 3. Convergence of the proposed optimal (**Algorithm** 1) and low-complexity sub-optimal (**Algorithm** 3) algorithms. $P_{\max} = 45$ dBm, $K = 2$, $N = 2$, $N_T = 4$, $D_1 = 1$, $D_2 = 2$, and $d_1 = d_2 = 50$m.

Figure 4. Convergence of the proposed low-complexity scheme. $P_{\max} = 45$ dBm, $M = 64$, $N = 4$, $D_1 = D_2 = 2$, $D_k = 4$, $\forall k \neq \{1, 2\}$. The users are randomly distributed within the inner and the outer radius.

the optimal solution after approximately 2000 and 3000 iterations for $M = 12$ and $M = 16$, respectively, while **Algorithm** 3 converges in less than 5 iterations for both $M = 12$ and $M = 16$. For the proposed optimal scheme, the number of iterations required for convergence increases significantly as the number of sub-carriers increase since increasing the number of sub-carriers increases the dimensionality of the search space. The convergence speed of the proposed low-complexity scheme is less sensitive to the problem size and the number of users compared to that of the optimal scheme. Furthermore, Fig. 3 shows that the proposed low-complexity sub-optimal scheme achieves the same performance as the optimal scheme.

In Fig. 3, we chose relatively small values for M, N, N_T, and K since the complexity of optimal **Algorithm** 1 increases rapidly with the dimensionality of the problem. In Fig. 4, we investigate the convergence behaviour of the proposed sub-optimal scheme for larger values of these parameters. As can be observed from Fig. 4, for all considered combinations of parameter values, the proposed low-complexity suboptimal scheme requires at most 5 iterations to converge.

In Figs. 5 and 6, we show the average sum throughput versus the maximum transmit power at the BS. As expected, the average system sum throughput improves with increasing maximum transmit power $P_{\max}$ because the SINR of the users can be enhanced by allocating more power. In particular, in Fig. 5, the proposed low-complexity scheme attains virtually the same performance as the proposed optimal scheme for all considered transmit powers. In Fig. 5, we also show results for the two baseline schemes. Baseline schemes 1 and 2 achieve lower throughputs compared to the proposed schemes. For baseline scheme 2, this
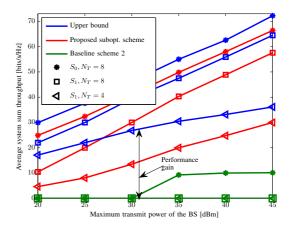
Figure 5. Average system sum throughput versus BS transmit power, $P_{\max}$, for different resource allocation schemes. $M = 16$, $N_T = 2$, $N = 2$, $K = 2$, $d_k = 50$ m, $\forall k \in \{1, 2\}$, and $D_1 = 1, D_2 = 2$.

Figure 6. Average system sum throughput versus BS transmit power, $P_{\max}$, for different resource allocation schemes. $M = 64$, $N = 4$, and $K = 6$. The users are randomly distributed within the inner and the outer radius.
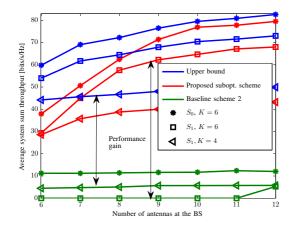
performance loss is caused by the sub-optimality of the fixed beamformer. This causes the average system sum throughput to quickly saturate for transmit powers exceeding 25 dBm. For baseline scheme 1, the resource allocation policies $\mathbf{W}$ obtained based on Shannon's capacity formula often violate constraint C1 in (13), especially for small $P_{\max}$, leading to a non-feasible solution. Therefore, Shannon's capacity formula should not be used for the design of MISO OFDMA-URLLC systems, since the QoS requirements of the users cannot be guaranteed. For high $P_{\max}$, for the proposed schemes, all non-zero $\text{Tr}(\mathbf{W}_k[m,n])$ assume large values. Hence, the corresponding $\gamma_k[m,n]$ in (7) are large and $V_k(\mathbf{w}_k)$ becomes negligible compared to $F_k(\mathbf{w}_k)$. Therefore, in this case, baseline scheme 1, which assumes $V_k(\mathbf{w}_k)$ is zero, yields a similar performance as the proposed schemes.

In Fig. 6, we show the average system sum throughput for different numbers of antennas and different delay requirements. For delay scenario S_0, none of the users has delay restrictions, i.e., $D_k = N = 4, \forall k$. In contrast, for delay scenario S_1, two users have strict delay constraints while the remaining users do not, i.e., $D_1 = D_2 = 2, D_k = N = 4, \forall k \neq \{1, 2\}$. As can be observed from Fig. 6, adding more antennas at the BS considerably increases the average system sum throughput, as additional antennas offer additional degrees of freedom for resource allocation and enable more efficient and more precise beamforming. In particular, for delay scenario S_1 and transmit power $P_{\max} = 40$ dBm, increasing the number of antennas at the BS from $N_T = 4$ to $N_T = 8$ improves the average system sum throughput by 96.77 %. Fig. 6 also reveals the impact of the delay requirements on the average system sum throughput. As can be observed, the stricter delay requirements for S_1 reduce the average system sum

throughput compared to S_0 because the BS is forced to allocate more power to the two delay sensitive users even if their channel conditions are poor to ensure their delay requirements are met. For example, for $P_{\max} = 40$ dBm and $N_T = 8$, the strict delay requirements of S_1 decreases the upper bound and the average system throughput of the proposed scheme compared to S_0 by 6.2 and 10 bits/s/Hz, respectively.

In Fig. 7, we investigate the average system sum throughput versus the number of antennas at the BS, N_T, for delay scenarios S_0 and S_1 considered in Fig. 6, and different numbers of URLLC users. As can be observed from Fig. 7, the average system sum throughput significantly improves as the number of antennas at the BS increases. This is due to the fact that more antennas offer additional degrees of freedom for resource allocation which leads to higher received SINRs at the users. Furthermore, the proposed scheme approaches the performance upper bound as the number of BS antennas increases since the value of $V_k(\mathbf{w}_k)$ in (7) becomes small compared to that of $F_k(\mathbf{w}_k)$ for large SINRs. Hence, the impact of finite blocklength coding on the average system sum throughput can be compensated by using large numbers of antennas at the BS. Moreover, as expected, changing the delay requirements from S_0 to S_1 reduces the throughput for all considered schemes. To compensate for this effect, the BS can increase the number of antennas in order to be able to serve the users with stricter delay requirement in a more efficient manner. Fig. 7 also elucidates the impact of the numbers of users on the average system sum throughput. As can be seen, since the proposed scheme can exploit multi-user diversity, increasing the number of users from $K = 4$ to $K = 6$ increases the throughput. In contrast, baseline scheme 2 cannot support $K = 6$ users for delay scenario S_1 because this scheme does not exploit all available degrees of freedom for resource allocation, and hence, the two users with strict delay requirements may lead to infeasible solutions, which has a negative impact on the average system sum throughput.

In Fig. 8, we investigate the effect of different delay requirements on the average system sum throughput. We consider the following delay scenarios: $\tilde{S}_0 = \{D_k = N = 5, \forall k\}$ (i.e., no delay sensitive users), $\tilde{S}_1 = \{D_1 = D_2 = D, D_k = N = 5, \forall k \neq \{1, 2\}\}$, $\tilde{S}_2 = \{D_k = D, \forall k \in \{1, 2, 3, 4\}, D_5 = D_6 = N = 5\}$, $\tilde{S}_3 = \{D_k = D, \forall k \in \{1, 2, 3, 4, 5\}, D_6 = N = 5\}$. In Fig. 8, we show the average system sum throughout versus delay parameter D. As can be observed, the average system sum throughout increases with D, which is due to the fact that a large D increases the feasible set of problem (13). Furthermore, the average system sum throughput decreases as the number of delay sensitive users requiring a delay of $D < N = 5$ increases, since having to serve more delay sensitive users reduces the flexibility in resource allocation. Moreover, the performance of baseline scheme 2 decreases significantly
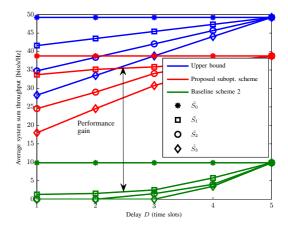
Figure 7. Average system sum throughput [bits/s/Hz] vs. number of antennas at the BS. $P_{\max} = 45$ dBm, $M = 64$, and $N = 4$. The users are randomly distributed within the inner and the outer radius.

Figure 8. Average system sum throughput [bits/s/Hz] vs. delay in time slots. $K = 6$, $P_{\max} = 45$ dBm, $M = 64$, $N = 5$, and $N_T = 6$. The users are randomly distributed within the inner and the outer radius.

if delay sensitive users are present. In particular, for baseline scheme 2, changing the delay requirements from $\tilde{S}_0$ to $\tilde{S}_1$ significantly decreases the average system sum throughput, as fixed MRT beamforming is not able to adequately support delay sensitive users.

In Fig. 9, we show the average system sum throughput versus the number of users for delay scenarios S_0 and S_1 considered in Fig. 6. The average system sum throughput for the proposed low-complexity scheme is close to the upper bound for small numbers of users for both considered delay scenarios. This is due to the fact that if there are only few users, they can be assigned a sufficiently large number of resource blocks to make the impact of finite blocklength coding negligible. As the number of users increases, the average system sum throughput increases due to multi-user diversity. However, at the same time, the impact of finite blocklength coding becomes more pronounced, and hence, the gap between the proposed scheme and the upper bound widens. Thus, there exists a trade-off between the performance degradation caused by short blocklengths and the performance gain induced by multi-user diversity. On the other hand, baseline scheme 2 cannot support more than $K = 6$ users for delay scenario S_1 because this scheme does not exploit all available degrees of freedom for resource allocation.

VII. CONCLUSION

In this paper, we investigated the optimal resource allocation algorithm design for broadband MISO OFDMA-URLLC systems. The resource allocation algorithm design was formulated as a non-convex optimization problem for maximization of the weighted system sum throughput subject to QoS constraints for the URLLC users. The global optimum solution was
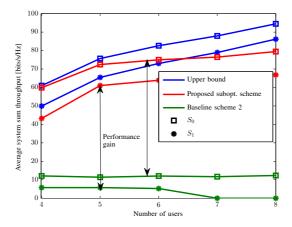
Figure 9. Average system sum throughput versus number of users. $P_{\max} = 45$ dBm, $M = 64$, $N = 4$, and $N_T = 12$. The users are randomly distributed within the inner and the outer radius.

obtained exploiting monotonic optimization theory. Moreover, to strike a balance between complexity and performance, we proposed a low-complexity sub-optimal algorithm to solve the optimization problem using successive convex approximation. Our simulation results revealed that the proposed sub-optimal algorithm achieves a close-to-optimal performance with low computational complexity. Furthermore, deploying multiple antennas at the BS was shown to be an effective approach to improve the reliability and to reduce the latency of URLLC systems. Moreover, our results revealed that stringent delay requirements have a negative impact on the throughput of MISO OFDMA-URLLC systems. Our results also showed that resource allocation based on Shannon's capacity formula, as is typically done in MISO OFDMA systems, may lead to infeasible solutions if URLLC is desired. Finally, the proposed optimal and sub-optimal algorithms were shown to significantly outperform two heuristic baseline schemes emphasizing the importance of optimal resource allocation in MISO OFDMA-URLLC systems.

APPENDIX A

In the following, we show that the objective function of (25) is a difference of two monotonic and concave functions. To this end, we rewrite the objective function as follows:

$$U(\mathbf{z}) = F(\mathbf{z}) - V(\mathbf{z}). \tag{50}$$

$U(\mathbf{z})$ is the difference of two concave functions if both $F(\mathbf{z})$ and $V(\mathbf{z})$ are monotonic and concave. Function $F(\mathbf{z})$ is a sum of logarithmic functions, and hence, it is a monotonic and concave function [35]. Furthermore, to prove that $V(\mathbf{z})$ is monotonic and concave, we rewrite

it as follows:

$$V(\mathbf{z}) = \sum_{k=1}^{K} \mu_k Q^{-1}(\epsilon_k) \sqrt{\sum_{m=1}^{M} \sum_{n=1}^{N} V_k[m,n]}, \tag{51}$$

where

$$V_k[m,n] = a^2 \left(1 - (1 + z_k[m,n])^{-2}\right). \tag{52}$$

Note that $V(\mathbf{z})$ is always positive, because for $\epsilon_k \in (0, 0.5)$, $Q^{-1}(\epsilon_k) > 0$ holds. To prove the monotonicity and the concavity of $V(\mathbf{z})$, first we will show that $V_k[m,n]$ is concave by taking the first and second derivatives with respect to $z_k[m,n]$ as follows:

$$\frac{\mathrm{d}V_k[m,n]}{\mathrm{d}z_k[m,n]} = \frac{2a^2}{(1 + z_k[m,n])^3}, \tag{53}$$

$$\frac{\mathrm{d}^2V_k[m,n]}{\mathrm{d}(z_k[m,n])^2} = \frac{-6a^2}{(1 + z_k[m,n])^4}. \tag{54}$$

Function $V_k[m,n]$ is a monotonic increasing and concave function because the first derivative is positive and the second derivative is negative for any $z_k[m,n] > 0$, respectively. Moreover, since a sum of monotonic functions is monotonic, and the sum of concave functions is also concave, $\sum_{m=1}^{M} \sum_{n=1}^{N} V_k[m,n]$ is a monotonic and concave function. By using the composition rules of convex analysis, the square root is concave and the extended-value extension on the real line is non-decreasing [35]. Thus, the square root of a monotonic and concave function is monotonic and concave. Finally, a weighted sum of monotonic and concave functions is also a monotonic and concave. This concludes the proof.

APPENDIX B

The SDP problem in (34) is jointly convex in the optimization variables and satisfies Slater's constraint qualifications. Therefore, strong duality holds and solving the dual problem is equivalent to solve the primal problem [35]. To formulate the dual problem, we write the Lagrangian of problem (34) as follows:

$$
\begin{aligned}
\mathcal{L} = &-\sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \theta_k[m,n][f_k[m,n](\mathbf{W}) - \lambda z_k^{(j)}[m,n]g_k[m,n](\mathbf{W})] \\
&+\alpha \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathrm{Tr}(\mathbf{W}_k[m,n]) \\
&+\sum_{k=1}^{K} \sum_{n=1}^{N} \eta_k[n] \mathrm{Tr}(\mathbf{W}_k[m,n]) - \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathrm{Tr}(\mathbf{W}_k[m,n]\mathbf{Y}_k[m,n]) + \Lambda,
\end{aligned} \tag{55}
$$

where Λ represents the collection of all terms that are independent of $\mathbf{W}$. Variables $\theta_k[m, n]$, α, and $\eta_k[n]$ are the Lagrange multipliers associated with constraints C6, C2, and C3, respectively. Matrices $\mathbf{Y}_k[m, n] \in \mathbb{C}^{N_T \times N_T}$ are the Lagrange multipliers for the positive semi-definite constraint C4 for matrices $\mathbf{W}_k[m, n]$. Therefore, the dual problem for the SDP problem in (34) is given as follows:

$$\underset{\substack{\theta_k[m,n],\alpha,\eta_k[n]\geq 0, \\ \mathbf{Y}_k[m,n]\succeq 0}}{\text{maximize}} \underset{\mathbf{W}_k[m,n],\boldsymbol{\zeta}}{\text{minimize}} \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}, \theta_k[m,n], \alpha, \eta_k[n], \mathbf{Y}_k[m,n]). \tag{56}$$

In the following, we reveal the structure of the optimal $\mathbf{W}$ of (34) by studying the Karush Kuhn Tucker (KKT) optimality conditions. The KKT conditions for the optimal solution $\mathbf{W}^*$ are given by:

$$\mathbf{Y}_k^*[m, n] \succeq 0, \quad \theta_k^*[m, n], \alpha^*, \eta_k^*[n] \geq 0 \tag{57}$$

$$\mathbf{Y}_k^*[m, n]\mathbf{W}_k^*[m, n] = \mathbf{0}, \tag{58}$$

$$\nabla_{\mathbf{W}_k^*[m,n]}\mathcal{L} = \mathbf{0}, \tag{59}$$

where $\mathbf{Y}_k^*[m, n]$, $\theta_k^*[m, n]$, α^*, and $\eta_k^*[n]$ are the optimal Lagrange multipliers for dual problem (56), and $\nabla_{\mathbf{W}_k^*[m,n]}\mathcal{L}$ denotes the gradient with respect to matrices $\mathbf{W}_k^*[m, n]$. The KKT condition in (59) can be rewritten as follows:

$$-\theta_k^*[m, n]\mathbf{H}_k[m] + \alpha\mathbf{I}_{N_T} + \eta_k[n]\mathbf{I}_{N_T} - \mathbf{Y}_k^*[m, n] = \mathbf{0}. \tag{60}$$

By rearranging the terms in (60), we obtain:

$$(\alpha + \eta_k[n])\mathbf{I}_{N_T} = \theta_k^*[m, n]\mathbf{H}_k[m] + \mathbf{Y}_k^*[m, n]. \tag{61}$$

Multiplying both sides of (61) with $\mathbf{W}_k^*[m, n]$ and exploiting (58), we get:

$$(\alpha + \eta_k[n])\mathbf{W}_k^*[m, n] = \theta_k^*[m, n]\mathbf{H}_k[m, n]\mathbf{W}_k^*[m, n]. \tag{62}$$

Now, we consider two cases for the value of $\alpha + \eta_k[n]$, namely $\alpha + \eta_k[n] = 0$ and $\alpha + \eta_k[n] > 0$. For the first case, since both α and $\eta_k[n]$ are non-negative $\alpha + \eta_k[n] = 0$ implies that $\eta_k[n] = 0$, and as a result, constraint C3 holds with equality. This means that $\mathbf{W}_k^*[m, n] = \mathbf{0}$ and hence $\text{Rank}(\mathbf{W}_k^*[m, n])$ is zero. For the second case, when $\alpha + \eta_k[n] > 0$ holds, using basic rank

inequalities for matrices, we obtain the following relations:

$$\text{Rank}((\alpha + \eta_k[n])\mathbf{W}_k^*[m, n]) = \text{Rank}(\mathbf{W}_k^*[m, n])$$

$$= \text{Rank}(\theta_k^*[m, n]\mathbf{H}_k[m]\mathbf{W}_k^*[m, n]) \leq \text{Rank}(\theta_k^*[m, n]\mathbf{H}_k[m]) \leq 1. \qquad (63)$$

This implies that the beamforming matrix is either rank one or $\mathbf{W}_k^*[m, n] = \mathbf{0}$, i.e., no transmission to user k on subcarrier m on time slot n. This completes the proof of Theorem 1.

REFERENCES

[1] W. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink URLLC-OFDMA systems," in *Proc. IEEE Int. Commun. Conf.*, Shanghai, P.R. China, May 2019.

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

[3] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. IEEE Int. Conf. 5G Ubiq. Connect*, Nov 2014, pp. 146–151.

[4] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.

[5] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless. Commun*, vol. 11, no. 9, pp. 3292–3304, September 2012.

[6] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. IEEE Intern. Sympos. on Inf. Theory*, July 2006, pp. 1394–1398.

[7] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Trans. Wireless. Commun*, vol. 11, no. 10, pp. 3618–3631, October 2012.

[8] W. Dang, M. Tao, H. Mu, and J. Huang, "Subcarrier-pair based resource allocation for cooperative multi-relay OFDM systems," *IEEE Trans. Wireless Commun*, vol. 9, no. 5, pp. 1640–1649, May 2010.

[9] V. D. Papoutsis, I. G. Fraimis, and S. A. Kotsopoulos, "User selection and resource allocation algorithm with fairness in MISO-OFDMA," *IEEE Commun. Lett.*, vol. 14, no. 5, pp. 411–413, May 2010.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[11] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University.

[12] V. Strassen, "Asymptotische Abschatzungen in Shannon's Informationstheorie," *In Proc. 3rd Trans. Prague Conf. Inf. Theory*, vol. 56, no. 5, pp. 689–723, May 1962.

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

[14] 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, Dec 2016.

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

[16] Y. Hu, M. Ozmen, M. C. Gursoy, and A. Schmeink, "Optimal power allocation for QoS-constrained downlink multi-user networks in the finite blocklength regime," *IEEE Trans. Wireless Commun*, vol. 17, no. 9, pp. 5827–5840, Sept 2018.

[17] J. Chen, L. Zhang, Y. Liang, X. Kang, and R. Zhang, "Resource allocation for wireless-powered IoT networks with short packet communication," *IEEE Trans. Wireless Commun*, vol. 18, no. 2, pp. 1447–1461, Feb 2019.

[18] S. Xu, T. H. Chang, S. C. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun*, vol. 15, no. 8, pp. 5527–5540, Aug 2016.

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

[20] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Commun*, vol. 17, no. 1, pp. 127–141, Jan 2018.

[21] C. Shen, T. Chang, H. Xu, and Y. Zhao, "Joint uplink and downlink transmission design for URLLC using finite blocklength codes," in *Proc. ISWCS 2018, Lisbon, Portugal, August 28-31, 2018*, 2018, pp. 1–5.

[22] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun*, vol. 36, no. 11, pp. 2475–2485, Nov 2018.

[23] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Throughput-delay tradeoffs with finite blocklength coding over multiple coherence blocks," *IEEE Trans. Commun*, pp. 1–1, 2019.

[24] M. Haghifam, M. Robat Mili, B. Makki, M. Nasiri-Kenari, and T. Svensson, "Joint sum rate and error probability optimization: Finite blocklength analysis," *IEEE Wireless Commun. Lett*, vol. 6, no. 6, pp. 726–729, Dec 2017.

[25] Y. Sun, D. W. K. Ng, J. Zhu, and R. Schober, "Robust and secure resource allocation for full-duplex MISO multicarrier NOMA systems," *IEEE Trans. Commun*, vol. 66, no. 9, pp. 4119–4137, Sep. 2018.

[26] Y. Sun, D. W. K. Ng, and R. Schober, "Optimal resource allocation for multicarrier MISO-NOMA systems," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–7.

[27] H. Tuy, F. Al-Khayyal, and P. T. Thach, *Monotonic optimization: branch and cut methods*. Boston, MA: Springer US, 2005, pp. 39–78. [Online]. Available: https://doi.org/10.1007/0-387-25570-2_2

[28] Y. J. A. Zhang, L. Qian, and J. Huang, "Monotonic optimization in communication and networking systems," *Found. Trends Netw.*, vol. 7, no. 1, pp. 1–75, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1561/1300000038

[29] H. Tuy, "Monotonic optimization: Problems and solution approaches," *SIAM J. on Optimization*, vol. 11, no. 2, pp. 464–494, Feb. 2000. [Online]. Available: https://doi.org/10.1137/S1052623499359828

[30] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Found. Trends Commun. Inf. Theory*, vol. 9, no. 23, pp. 113–381, 2013. [Online]. Available: http://dx.doi.org/10.1561/0100000069

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

[32] H. Tuy, *Convex Analysis and Global Optimization*, 2nd ed. Springer Publishing Company, Incorporated, 2016.

[33] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun*, vol. 65, no. 3, pp. 1077–1091, March 2017.

[34] E. Che, H. D. Tuan, and H. H. Nguyen, "Joint optimization of cooperative beamforming and relay assignment in multi-user wireless relay networks," *IEEE Trans. Wirel. Commun*, vol. 13, no. 10, pp. 5481–5495, Oct 2014.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.