

Rudra: An Algorithm for Optimizing Spectrum Efficiency with Data Rate Guarantee in Next-G Communications

Shiva Acharya* Shaoran Li[†] Yubo Wu* Nan Jiang[‡] Wenjing Lou* Y. Thomas Hou*

*Virginia Tech, Blacksburg, VA, USA [†]Nvidia Corp., Santa Clara, CA, USA

[‡]Georgia Institute of Technology, Atlanta, GA, USA

Abstract—The limited spectrum and data-intensive applications in 5G/Next-G networks call for data rate guarantees for UEs with minimum spectrum usage. This problem is challenging due to the complexity of MU-MIMO system, unknown Channel State Information (CSI), and estimation errors. We propose Rudra, which aims to minimize spectrum usage while offering probabilistic guarantee for each UE’s data rate. The essence of Rudra is to employ Chance-Constrained Programming (CCP) by leveraging limited CSI samples. By reformulating the CCP into a deterministic one, Rudra offers an iterative solution that addresses Resource Block Group (RBG) allocation, Modulation and Coding Scheme (MCS) selection, and Beamforming (BF) design. Simulations show that Rudra can meet our design objective and outperforms a modified state-of-the-art algorithm.

Index Terms—5G/Next-G, MU-MIMO, CSI, errors, data sample, spectrum

I. INTRODUCTION

Many 5G/Next-G applications (e.g., eMBB) require steady data rate to ensure UEs’ quality of experience (QoE) [1]. With such an increase in demand and limited radio spectrum, it is critical to optimize radio resource allocation so as to minimize spectrum usage while providing the expected data rates [2]. There are several well-known challenges to this problem: First, the base station (BS) must (in the downlink direction) efficiently allocate Resource Blocks (RBs) or RB Groups (RBGs), select the appropriate Modulation and Coding Scheme (MCS) for each UE, and design MU-MIMO beamforming (BF) for its antennas during each transmission time interval (TTI). Optimizing these tasks jointly is not trivial, especially when the objective is to minimize spectrum usage and the constraint involves per-UE data rate guarantee. Second, optimal RB(G) allocation, MCS selection, and BF design all hinges upon knowledge of accurate Channel State Information (CSI). But accurate CSI is hardly available in practice due to unknown CSI distribution as well as estimation errors [3]. Such a lack of CSI poses a fundamental challenge to optimal RBG allocation, MCS selection, and BF design.

The goal of this paper is to minimize spectrum usage while providing the expected data rates through optimal RBG allocation, MCS selection, and BF design, all under CSI uncertainty. To the best of our knowledge, none of the existing works has successfully addressed this problem. Most of the

existing research did not offer a comprehensive solution for RBG allocation, MCS selection, and MU-MIMO BF design to meet UE data rate requirements [4]–[11]. Also, minimizing spectrum usage is typically not an objective in these works. Further, most scheduling solutions for MU-MIMO assumed either perfect CSI knowledge or some well-known distributions [4]–[10]. But perfect CSI knowledge is hardly available in practice; neither are the assumed models accurate, rendering limited practical significance of these approaches.

The main contributions of this paper are summarized below:

- We study a spectrum minimization problem for 5G MU-MIMO. Our problem aims to address RBG allocation, MCS selection, and MU-MIMO BF design to provide a probabilistic data rate guarantee to UEs. To address CSI uncertainty, we resort to a limited number of CSI data samples, without any assumption or knowledge of CSI distribution. This approach is especially appealing to cope with time-varying channel conditions, as what typically occur in practice.
- To meet UEs’ data rate requirements with a high probability (statistical guarantee), we formulate our problem based on chance-constrained programming (CCP). We propose Rudra, a data-driven solution that relies on limited CSI data samples. In the first stage, Rudra transforms the CCP into a deterministic Mixed Integer Non-Linear Programming (MINLP) problem by constructing an error-embedded (EE)-Wasserstein ambiguity set based on limited CSI samples, bridging the gap between the true, but unknown empirical distribution and the empirical distribution from collected CSI samples.
- For the reformulated deterministic optimization problem, Rudra offers an iterative algorithm to minimize spectrum usage, along with finding a feasible solution to RBG allocation, MCS selection, and MU-MIMO BF. Specifically, Rudra allocates RBGs based on UEs’ data rate requirements, prioritizing UEs with higher rate demands, and selects suitable MCS to satisfy these demands. For BF design, Rudra leverages Zero-forcing (ZF) basis vectors based on available CSI samples, linearly combining them and subsequently scaling them to match UE power allocations.
- Through extensive simulations, we demonstrate that Rudra meets our design objective. Specifically, the results show that Rudra is capable of keeping data rate violation below the prescribed risk level for all UEs, whereas

This research was supported in part by NSF grant CNS-2312447, ONR MURI grant N00014-19-1-2621, Virginia Commonwealth Cyber Initiative (CCI), Virginia Tech Institute for Critical Technology and Applied Science (ICTAS).

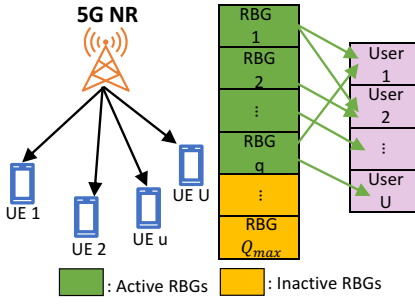


Fig. 1. System Model: MU-MIMO transmission in 5G NR.

the benchmark algorithm (based on the state-of-the-art) often exceeds the required risk level significantly. In the meantime, Rudra's RBG usage is only marginally higher than the benchmark algorithm.

II. PROBLEM DESCRIPTION

Consider a 5G BS serving a set of UEs, as shown in Fig. 1. The BS has multiple antennas, and each UE has a single antenna. We consider a downlink scenario where each UE has a data rate requirement. The BS aims to minimize its radio resources while offering guarantee to these data rate requirements.

Following 3GPP standards [12], the time and frequency domains are slotted into Transmission Time Intervals (TTIs) and subcarriers (SCs). A group of 12 SCs in one TTI is called a Resource Block (RB). To reduce control overhead, it is common to combine a few RBs into an RB Group (RBG) [12] when scheduling radio resources to the UEs. We will employ RBG for scheduling in this paper.

Scheduling decisions at a BS in each TTI involves three components.

- First, BS needs to allocate the RBGs to different UEs. Under MU-MIMO, one RBG can be scheduled to serve multiple UEs, while one UE can be served on multiple RBGs.
- Second, the BS needs to choose from one of the 29 MCS candidates for each UE, and a UE must use the same MCS for all its scheduled RBGs [12]. A higher MCS means a higher data rate per RBG but also requires a higher SINR to support it. Since the same MCS must be used across all scheduled RBGs for a given UE, BS must consider the channel quality of all scheduled RBGs to that UE.
- Third, the BS needs to do BF to mitigate inter-user interference and improve SINR, enabling UEs to decode intended signals successfully.

We assume each UE has a data rate requirement. However, the inherent uncertainty in CSI between the BS and a UE makes it impractical to offer a deterministic performance guarantee. In this case, a probabilistic performance guarantee of data rate is plausible and is what we will pursue in this work.

From the network operator's perspective, it is desired to minimize the spectrum resources used by the BS to meet

these data rate requirements. Thus, the objective of the paper is to minimize the spectrum resources used by the BS while satisfying the probabilistic data rate guarantees for the UEs.

There are a number of technical challenges to our problem. First, we want to address underlying CSI uncertainty between the BS and a UE in a realistic manner, without assuming any knowledge of CSI distribution or statistics. This is different from the vast amount of research that either assumes knowledge of perfect CSI or assumes CSI follows some well-known distributions. Second, the three components in scheduling, namely RBG allocation, MCS selection, and BF design, are coupled with each other. For instance, since multiple RBGs can be assigned to a given UE, and a UE must use the same MCS across all assigned RBGs, RBG allocation directly affects MCS selection. Additionally, under MU-MIMO, an RBG can be assigned to multiple UEs, which will directly impact the design of BF vectors. Third, each of these scheduling components is mathematically challenging on its own. RBG allocation and MCS selection involve large search space with integer variables, while BF design involves complex vectors with non-convex terms.

III. MODELING AND FORMULATION

In this section, we analyze and formulate our problem.

At the BS, denote $Q_{\max}$ as the upper limit of RBGs that can be allocated, and N_T as the number of antennas. Denote $\mathcal{Q} = \{1, 2, \dots, g, \dots, q\}$ as the required set of contiguous RBGs (see Fig. 1) that will be allocated to the UEs to meet their data rate requirements, where q represents the maximum number of RBGs required. We have:

$$1 \leq q \leq Q_{\max}. \quad (1)$$

Denote $\mathcal{U} = \{1, 2, \dots, u, \dots, U\}$ as the set of UEs, where U is the maximum number of UEs served by the BS. Denote x_u^g as a binary variable to indicate whether RBG g is assigned to UE u , i.e.,

$$x_u^g = \begin{cases} 1, & \text{if RBG } g \text{ is assigned to UE } u, \\ 0, & \text{otherwise.} \end{cases}$$

Under MU-MIMO, since each RBG can transmit to up to N_T UEs simultaneously, we have:

$$\sum_{u \in \mathcal{U}} x_u^g \leq N_T \quad (g \in \mathcal{Q}). \quad (2)$$

At UE u , denote z_u^m as a binary variable to indicate whether or not it uses MCS level $m \in \mathcal{M}$, where $\mathcal{M} = \{1, 2, \dots, m, \dots, M\}$ represents the set of possible MCS levels, with M being the maximum MCS level. We have:

$$z_u^m = \begin{cases} 1, & \text{if UE } u \text{ uses MCS level } m, \\ 0, & \text{otherwise.} \end{cases}$$

Per 5G specifications [12], a UE can only use one (same) MCS level for all the RBGs allocated to it. We have,

$$\sum_{m \in \mathcal{M}} z_u^m = 1 \quad (u \in \mathcal{U}). \quad (3)$$

For BF at the BS, denote $\mathbf{w}_u^g$ as an $N_T \times 1$ complex precoding vector for UE u on RBG g . $\mathbf{w}_u^g$'s for all UEs will be designed by the BS based on CSI (more on CSI later). We want to ensure $\|\mathbf{w}_u^g\|_2^2$ to be zero when RBG g is not allocated to UE u (where $\|\cdot\|_2$ is the L^2 -norm). We have:

$$\|\mathbf{w}_u^g\|_2^2 \leq x_u^g P_{\max} \quad (u \in \mathcal{U}, g \in \mathcal{Q}), \quad (4)$$

where $P_{\max}$ is the BS's total power budget for all UEs. To ensure the aggregate $\|\mathbf{w}_u^g\|_2^2$ over all UE u and RBG g does not exceed $P_{\max}$, we have:

$$\sum_{u \in \mathcal{U}} \sum_{g \in \mathcal{Q}} \|\mathbf{w}_u^g\|_2^2 \leq P_{\max}. \quad (5)$$

Probabilistic Data Rate Guarantee Since a UE uses the same MCS across all the RBGs allocated to it, each RBG will contribute the same data rate as long as its SINR exceeds the required threshold by the MCS. Denote γ_u^g as the achieved SINR from RBG g at UE u . We have:

$$\gamma_u^g = \frac{|(\mathbf{w}_u^g)^\dagger \mathbf{h}_u^g|^2}{\sum_{i \in \mathcal{U}, i \neq u} |(\mathbf{w}_i^g)^\dagger \mathbf{h}_u^g|^2 + \sigma^2}, \quad (6)$$

where $\mathbf{h}_u^g$ is the unknown CSI from RBG g to UE u , $(\cdot)^\dagger$ is the conjugate transpose of complex vector, σ^2 is the thermal noise.

Denote η^m as the required SINR threshold for MCS level m and r^m as the achievable data rate corresponding to MCS level m , respectively (see Table 5.1.3.1-1 in [12]). Denote $r_{u,m}^g$ as the achieved instantaneous data rate by RBG g on UE u under MCS level m . We have:

$$r_{u,m}^g = \begin{cases} r^m, & \text{if } \gamma_u^g \geq \eta^m, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

That is, if the achieved SINR γ_u^g exceeds the required threshold η^m for MCS level m , we will be able to obtain a data rate of r^m or 0 otherwise. The total data rate for UE u can be calculated by summing up the achieved data rate across all RBGs allocated to u and all m 's, which is $\sum_{g \in \mathcal{Q}} \sum_{m \in \mathcal{M}} z_{u,m}^g r_{u,m}^g$.

Denote R_u as the data rate requirement from UE u . The computation of the expected data rate requires calculating the achieved SINR (6), which depends on the unknown CSI $\mathbf{h}_u^g$. In contrast to the existing approaches that either assume perfect knowledge of CSI or assume that CSI follows some well-known distributions, we do not make any of these assumptions. Instead, we employ chance-constrained programming (CCP), which allows us to incorporate the unknown CSI distribution through the use of a chance constraint. This approach allows for occasional violation of the constraint up to a small probability, known as risk level. We have:

$$\mathbb{P}\left\{\sum_{g \in \mathcal{Q}} \sum_{m \in \mathcal{M}} z_{u,m}^g r_{u,m}^g \geq R_u\right\} \geq 1 - \epsilon_u \quad (u \in \mathcal{U}), \quad (8)$$

where ϵ_u is the violation probability that can be tolerated by UE u . Constraint (8) says that the achieved data rate must be no smaller than the required data rate (R_u) with probability at least $1 - \epsilon_u$.

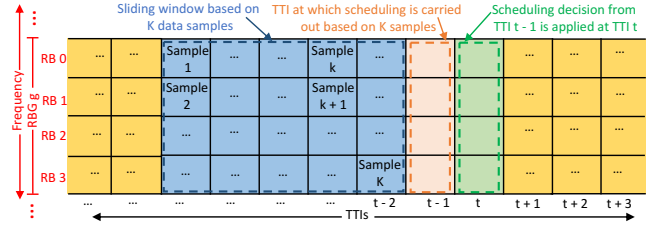


Fig. 2. An illustration of sliding window used to collect K data samples for scheduling.

Objective In this work, we are interested in minimizing the spectrum usage (i.e., number of contiguous active RBGs starting from RBG 1) at the BS, such that the data rate requirements of all UEs are met with high probability $1 - \epsilon_u$. The solution to this problem is the key for a carrier network operator to properly provision bandwidth for the cellular access network (e.g., through network slicing [13]).

Our optimization problem can be formulated as follows:

$$(P1) \quad \min_{x_u^g, z_{u,m}^g, \mathbf{w}_u^g} q$$

s.t. RBG constraint (1),
 UE assignment under MU-MIMO (2),
 MCS selection constraint (3),
 BS power constraints (4) and (5),
 Calculation of data rate (6) and (7),
 Data rate guarantee constraint (8).

P1 is called a CCP, with (8) being a chance constraint. As discussed in Section II, our problem is very challenging due to the presence of the chance constraint (8) that involves the unknown CSI distributions. Additionally, the problem is highly complex due to the coupling of RBG allocation, MCS assignment, and BF design (via constraints (4), (7), and (8)), as well as the large subspace of the binary variables x_u^g and $z_{u,m}^g$, and $\mathbf{w}_u^g$ being complex vectors.

Next, we present a solution, codenamed Rudra,¹ for problem P1. Rudra solves the problem in two stages, as illustrated in Section IV and Section V, respectively.

IV. RUDRA: REFORMULATION WITH LIMITED DATA SAMPLES

In this section, we show how Rudra reformulates the CCP (involving unknown CSI) into a deterministic problem solely based on limited data samples while aiming for the same probabilistic data rate guarantee that the original problem (P1) aims to offer.

In a nutshell, Rudra first forms an empirical distribution based on a very small set of recently collected CSI data samples. Rudra then bridges this empirical distribution to the unknown empirical distribution, which is also based on K

¹Rudra, a Hindu god of auspiciousness, is known as a benevolent deity who fulfills any wish. We use Rudra as the codename of our proposed solution as it aims to fulfill UEs' data rate requirements with high probabilities.

CSI data samples that are free of any CSI estimation errors, by forming an *error-embedded (EE)-Wasserstein ambiguity set* with an appropriate radius. Finally, based on the established ambiguity set, Rudra aims at transforming the chance constraint in P1 into a deterministic constraint that only depends on the small-set of collected CSI data samples. We elaborate on the details of these steps in the rest of this section.

Collecting CSI Data Samples As mentioned, our approach to address unknown CSI is through collecting a very small set of CSI data samples similar to the idea in [14]. Denote $\mathcal{K} = \{1, 2, \dots, k, \dots, K\}$ as the set of K CSI data samples, which come from most recent TTIs and are closely related, say within the same RBG (see Fig. 2). In this paper, we assume an RBG consists of 4 contiguous RBs [12].

As shown in Fig. 2, Rudra collects K CSI data samples from a sliding TTI window (in blue-shaded boxes) and these K data samples are immediately used to design the scheduling solution at the next TTI (i.e., slot $t - 1$). The scheduling solution found in TTI $t - 1$ is then applied to TTI t (shown in green-shaded box in the figure).

Given the small window size used for collecting the K CSI data samples, it is reasonable to consider that all these K samples within a given RBG follow the same distribution. However, it should be noted that different RBGs may have different CSI distribution. Denote $\hat{\mathbf{h}}_u^g(k)$ as the k -th CSI data sample for $\mathbf{h}_u^g$ (the channel from RBG g to UE u). With the K data samples, we can construct an *empirical distribution*, denoted as $\mathbb{P}_{\hat{\mathbf{h}}_u^g}$, for the unknown channel $\mathbf{h}_u^g$. We have:

$$\mathbb{P}\{\hat{\mathbf{h}}_u^g = \hat{\mathbf{h}}_u^g(k)\} = \frac{1}{K} \quad (u \in \mathcal{U}, g \in \mathcal{Q}, k \in \mathcal{K}). \quad (9)$$

Bridging Empirical Distribution of Collected CSI to Unknown Error-Free CSI Empirical Distribution Let $\mathbf{h}_u^{g*}$ be the discrete random variable corresponding to K data samples that are free from any estimation errors. Denote $\mathbb{P}_{\mathbf{h}_u^{g*}}$ as the unknown empirical distribution of $\mathbf{h}_u^{g*}$ based on K error-free CSI data samples. There is a “distance” between unknown empirical distribution of error-free CSI $\mathbf{h}_u^{g*}$ ($\mathbb{P}_{\mathbf{h}_u^{g*}}$) and the empirical distribution $\mathbb{P}_{\hat{\mathbf{h}}_u^g$ that is based on collected CSI data sample $\hat{\mathbf{h}}_u^g$. Such a distance between two distributions $\mathbb{P}_{\mathbf{h}_u^{g*}}$ and $\mathbb{P}_{\hat{\mathbf{h}}_u^g}$ can be represented by the so-called ∞ -Wasserstein distance [15], which we denote as $W_\infty(\mathbb{P}_{\mathbf{h}_u^{g*}}, \mathbb{P}_{\hat{\mathbf{h}}_u^g})$.

Although we have no idea about the unknown empirical distribution $\mathbb{P}_{\mathbf{h}_u^{g*}}$, we could construct a “ball”, with $\mathbb{P}_{\hat{\mathbf{h}}_u^g}$ as its center and a ∞ -Wasserstein distance (or radius) sufficiently large so that the unknown error-free empirical distribution is contained in the ball with a high probability. Denote $\mathcal{F}_{W_\infty}(\theta_u^g)$ as the set of all possible distributions such that

$$\mathcal{F}_{W_\infty}(\theta_u^g) = \left\{ \mathbb{P}_{\mathbf{h}_u^{g*}} : W_\infty(\mathbb{P}_{\mathbf{h}_u^{g*}}, \mathbb{P}_{\hat{\mathbf{h}}_u^g}) \leq \theta_u^g, \mathbf{h}_u^{g*} \in \mathbb{C}^{N_T \times 1} \right\},$$

where θ_u^g is the radius of the EE-Wasserstein ambiguity set for UE u on RBG g and $\mathbb{C}^{N_T \times 1}$ is the complex space for $N_T \times 1$ column vector. We call $\mathcal{F}_{W_\infty}(\theta_u^g)$ as error-embedded (EE)-Wasserstein ambiguity set as the center of the ball embeds errors.

Note that the EE-Wasserstein ambiguity set is fundamentally different than the so-called ∞ -Wasserstein ambiguity set [15], which assumes the collected data samples are error-free. Under the ∞ -Wasserstein ambiguity set, as the number of data samples approaches ∞ , the empirical distribution converges to the true unknown distribution, reducing the radius of the ∞ -Wasserstein ambiguity set to zero. In contrast, our EE-Wasserstein ambiguity set considers CSI data samples with external errors arising from hardware impairments, environmental noise, thermal noise, and other factors. Specifically, the empirical distribution at the center of our ambiguity set is based on the collected CSI data samples containing these errors. Therefore, even as the number of collected CSI data samples approaches ∞ , the empirical distribution based on these collected CSI data samples ($\mathbb{P}_{\hat{\mathbf{h}}_u^g}$) does not converge to the unknown empirical distribution ($\mathbb{P}_{\mathbf{h}_u^{g*}}$).²

In this context, The radius θ_u^g , a crucial hyperparameter, should be set according to the network setting to ensure that $\mathbb{P}_{\mathbf{h}_u^{g*}}$ falls within $\mathcal{F}_{W_\infty}(\theta_u^g)$. The larger the θ_u^g , the larger the ball and more conservative the solution will be (in terms of the size of the feasible search space and achieved objective value). Therefore, it is important to choose θ_u^g appropriately to meet our needs (i.e., to ensure $\mathbb{P}_{\mathbf{h}_u^{g*}}$ falls inside the ball). We propose to set the radius of the EE-Wasserstein ambiguity set through long-term offline measurement.

Reformulating Probabilistic Chance Constraint into Deterministic Constraint Now, we reformulate the chance constraint (8) into a deterministic constraint based on the established EE-Wasserstein ambiguity set. Since $\mathbb{P}_{\mathbf{h}_u^{g*}}$ belongs to $\mathcal{F}_{W_\infty}(\theta_u^g)$, as long as we guarantee the probabilistic data rate for all distributions in $\mathcal{F}_{W_\infty}(\theta_u^g)$, we can claim the same guarantee for $\mathbb{P}_{\mathbf{h}_u^{g*}}$. This means we can substitute (8) by the following:

$$\inf_{\mathbb{P}_{\mathbf{y}_u^g} \in \mathcal{F}_{W_\infty}(\theta_u^g)} \mathbb{P}\left\{ \sum_{g \in \mathcal{Q}} \sum_{m \in \mathcal{M}} z_u^m r_{u,m}^g \geq R_u \right\} \geq 1 - \epsilon_u \quad (u \in \mathcal{U}), \quad (10)$$

where $\mathbf{y}_u^g$ represents any $N_T \times 1$ complex column vector and $\mathbb{P}_{\mathbf{y}_u^g}$ represents any unknown empirical distribution inside $\mathcal{F}_{W_\infty}(\theta_u^g)$ based on K samples of $\mathbf{y}_u^g$.

To reformulate (10), we resort to the SINR γ_u^g , as $r_{u,m}^g$ is related to γ_u^g via (7). Inspired by the idea in [15], we aim to reformulate (10) based on limited data samples using EE-Wasserstein ambiguity set. As per [15], under ∞ -Wasserstein distance, guaranteeing (10) (i.e., guaranteeing chance constraint for any distribution $\mathbb{P}_{\mathbf{y}_u^g}$ in $\mathcal{F}_{W_\infty}(\theta_u^g)$) can be achieved by new constraints regarding the K CSI data samples. Specifically, solving (10) is equivalent to solving the worst-case value of the function associated with the chance constraint (i.e., $\sum_{g \in \mathcal{Q}} \sum_{m \in \mathcal{M}} z_u^m r_{u,m}^g$) for all available K data samples under norm constraint and requiring at least $K \cdot (1 - \epsilon_u)$ of these worst-case value to meet the data rate requirement R_u . By norm constraint, we require that the ∞ -Wasserstein distance

²On the other hand, as the number of CSI data samples goes to ∞ , the unknown error-free empirical distribution ($\mathbb{P}_{\mathbf{h}_u^{g*}}$) converges to the true error-free unknown CSI distribution ($\mathbb{P}_{\mathbf{h}_u^g}$).

between the empirical CSI distribution ($\mathbb{P}_{\mathbf{h}_u^g}$) and any unknown empirical distribution ($\mathbb{P}_{\mathbf{y}_u^g}$) be bounded by the radius θ_u^g .

When L^2 norm is used in the definition of ∞ -Wasserstein distance, finding the ∞ -Wasserstein distance between these two distributions is equivalent to finding the Euclidian distance between the collected CSI data sample $\mathbf{h}_u^g$ in $\mathbb{P}_{\mathbf{h}_u^g}$ and any CSI sample $\mathbf{y}_u^g$ in $\mathbb{P}_{\mathbf{y}_u^g}$. To compute the worst-case data rate for each UE, we first need to determine the worst-case SINR for these UEs. Let $\hat{\gamma}_u^g(k)$ denote the worst-case SINR for $u \in \mathcal{U}$, $g \in \mathcal{Q}$, and $k \in \mathcal{K}$. $\hat{\gamma}_u^g(k)$ can be written as:

$$\hat{\gamma}_u^g(k) = \min \left\{ \frac{|\mathbf{w}_u^g \mathbf{y}_u^g|^2}{\sum_{i \in \mathcal{U}, i \neq u} |\mathbf{w}_i^g \mathbf{y}_i^g|^2 + \sigma^2} : \|\mathbf{y}_u^g - \hat{\mathbf{h}}_u^g(k)\|_2 \leq \theta_u^g \right\}. \quad (11)$$

Based on the values of $\hat{\gamma}_u^g(k)$ for each UE, we can evaluate the expected data rate achieved by the UEs. Using a similar data rate calculations as in (7), let $\hat{r}_{u,m}^g(k)$ represent the expected data rate for UE u on RBG g based on $\hat{\gamma}_u^g(k)$, i.e.,

$$\hat{r}_{u,m}^g(k) = \begin{cases} r^m, & \text{if } \hat{\gamma}_u^g(k) \geq \eta^m, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

Since we only have K collected CSI data samples, to satisfy (10), we only need to guarantee that the achieved data rate for UE u is greater than or equal to R_u for at least $K \cdot (1 - \epsilon_u)$ data samples. Based on this understanding, substituting (12) into (10), we can reformulate (10) using the CSI data samples as:

$$\sum_{k \in \mathcal{K}} \mathbb{I} \left\{ \sum_{g \in \mathcal{Q}} \sum_{m \in \mathcal{M}} z_{u,m}^m \hat{r}_{u,m}^g(k) \geq R_u \right\} \geq K(1 - \epsilon_u) \quad (u \in \mathcal{U}), \quad (13)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function returning 1 if its argument is true and 0 otherwise.

Substitute (6) and (7) with (11) and (12) respectively, and similarly replace (8) with (13), we then reformulate P1 as follows:

$$\begin{aligned} \text{(P2)} \quad & \min_{x_u^g, z_u^m, \mathbf{w}_u^g} q \\ \text{s.t.} \quad & \text{Constraints (1), (2), (3), (4), (5),} \\ & \text{Data rate calculation (11) and (12),} \\ & \text{Data rate guarantee constraint (13).} \end{aligned}$$

P2 is a deterministic Mixed Integer Non-Linear Programming (MINLP) problem. It is still hard to solve due to the coupling of variables x_u^g , z_u^m , and $\mathbf{w}_u^g$ as discussed in Section II. In fact, solving joint RBG allocation and MCS selection was shown to be NP-hard [4]. Further, even with a given RBG allocation and MCS selection, BF design remains difficult due to constraints (11) and (13). In particular, it is hard to find the optimal solution in (11).

V. RUDRA: SOLVING THE DETERMINISTIC FORMULATION

In this section, we design a solution to P2. Our approach, shown in Fig. 3, consists of the following steps. First, Rudra identifies a lower bound Q^{LB} for the total number of RBGs

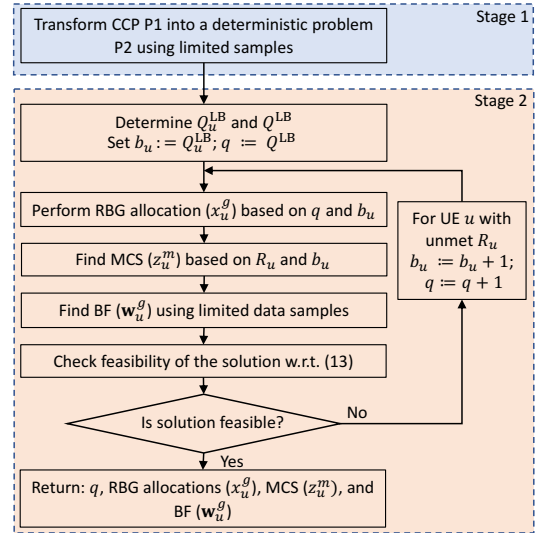


Fig. 3. A flow chart for solving the deterministic problem formulation by Rudra.

required and the minimum number of RBGs Q_u^{LB} for each UE to meet its data rate requirements. Using these values as a starting point, Rudra allocates RBGs based on each UE's data rate requirement R_u . Next, Rudra selects an appropriate MCS level for each UE based on the data rate requirements R_u and the assigned number of RBGs b_u . It then determines a suitable BF solution by first creating Zero-forcing (ZF) basis vectors from the collected CSI sample. These vectors are linearly combined to form an initial BF solution, which is then scaled to meet the assigned transmission power, calculated to achieve the minimum SINR required for the selected MCS. Rudra then checks the feasibility of the solution w.r.t. (13). If feasible, Rudra terminates and returns the current RBG usage q , RBG allocations x_u^g , MCS levels z_u^m , and BF vectors $\mathbf{w}_u^g$ for each UE. If the solution is not feasible, Rudra increases the RBG usage b_u for UEs that have not met their data rate requirements by 1, increases the total RBG usage q by 1, and repeats the process until all UEs' data rate requirements are satisfied. The remainder of this section details each step.

A. Step 1: Determining Lower Bound for RBG

A simple lower bound for the required number of RBGs can be found by assuming the highest MCS level (and thus its corresponding achieved data rate) for each UE. Based on this, we can determine the minimum number of RBGs required to meet the data rate requirement R_u of each UE u .

Denote r_u^{max} as the maximum data rate achieved by UE u on an RBG based on the highest MCS level (i.e., 29). r_u^{max} can be easily calculated based on the spectral efficiency (SE) Table 5.1.3.1-1 in [12]. Denote Q_u^{LB} as the minimum number of RBGs required for UE u . We have:

$$Q_u^{\text{LB}} = \left\lceil \frac{R_u}{r_u^{\text{max}}} \right\rceil, \quad (u \in \mathcal{U}).$$

Since Q_u^{LB} is based on the highest possible MCS level for each UE $u \in \mathcal{U}$, our lower bound for total number of required

RBGs, denoted as Q^{LB} , should be at least $\max_{u \in \mathcal{U}} Q_u^{\text{LB}}$. On the other hand, under MU-MIMO, each RBG can transmit up to N_T UEs on each RBG. In practice, due to channel rank conditions, the actual number of UEs supported by an RBG is typically smaller than N_T . In this paper, we assume each RBG can transmit to $N_T/2$ UEs, where N_T is the number of antennas at the BS. Then we have:

$$Q^{\text{LB}} = \max \left\{ \max_{u \in \mathcal{U}} Q_u^{\text{LB}}, \left\lceil \frac{\sum_u Q_u^{\text{LB}}}{N_T/2} \right\rceil \right\}.$$

Q^{LB} and Q_u^{LB} will be the starting points for q and b_u , respectively, in the iterative search.

B. Step 2: RBG allocation

In this step, we assign RBGs to UEs based on their data rate requirements R_u 's. The main idea is to prioritize UEs with the highest R_u 's with RBGs that enjoy the best channel quality (based on CSI data samples).

First, Rudra sorts UEs in $\mathcal{U}$ based on a descending order of R_u 's.³ Denote this new sorted set as $\mathcal{U}_{\text{sort}}$. Then we start with the first UE in $\mathcal{U}_{\text{sort}}$, i.e., the one with the highest R_u value in the set. Then for this UE u , we assign b_u RBGs with the best channel quality among the available RBGs. (If the number of available RBGs is less than b_u , Rudra increments q by one more RBG and restart the RBG allocation process for all UEs.)

To compare channel quality among the available RBGs and chose the top b_u RBGs with the best channel quality, we resort to the collected CSI data samples. Specifically, denote $\hat{\mathbf{h}}_u^g$ as the average squared Frobenius norm of channel gain for UE u across RBGs g based on the K CSI data samples. We have:

$$\hat{\mathbf{h}}_u^g = \frac{1}{K} \sum_{k \in \mathcal{K}} \left\| \mathbf{h}_u^g(k) \right\|_2^2. \quad (14)$$

The RBGs are then sorted in descending order based on their $\hat{\mathbf{h}}_u^g$ values, and the top b_u distinct RBGs with the highest gains are allocated to UE u .

Once Rudra is done with UE u , it removes u from the set $\mathcal{U}_{\text{sort}}$ and then consider the next UE in the set (with the highest R_u requirement). Throughout the RBG allocation process, we maintain a counter for each RBG to ensure it will not be allocated to more than $N_T/2$ UEs.

Once all UEs are allocated with their b_u 's, Rudra terminates with this RBG allocation step.

C. Step 3: Finding MCS

The goal of this step is to find a suitable MCS based on the RBG allocation results from the previous step. To meet each UE's data rate requirements R_u , each RBG allocated to an UE must deliver a minimum data rate of $\frac{R_u}{b_u}$, where R_u is UE u 's data rate requirement and b_u is the number of RBGs allocated to UE u . Since UEs must use the same MCS level for all scheduled RBGs, we set MCS level based on $\frac{R_u}{b_u}$. Then we choose the lowest MCS level, say m , from Table 5.1.3.1-1 [12] that satisfies $\frac{R_u}{b_u}$ and set $z_u^m = 1$ and $z_u^i = 0$ when $i \neq m$.

³If there is a tie during sorting, then break the tie randomly.

D. Step 4: Finding BF Solution

Our goal is to determine a suitable BF solution based on the current RBG allocation and MCS selection. We propose to exploit the K CSI data samples (between each RBG g and UE u) to compute ZF basis vectors, which will be combined linearly to form the initial BF solutions. To scale these solutions appropriately, we resort to power allocation. To do this, we allocate power on each RBG w.r.t. its UEs according to the UEs' minimum SINR requirements, which can be determined by the selected MCS levels from the previous step. Once we have the power allocation among the RBGs, the final BF solution is then obtained by scaling the initial BF solution in proportion to the power allocation. We elaborate on the details of our approach in the rest of this section.

Denote $\mathcal{U}^g$ as the set of scheduled UEs on RBG g (from RBG allocation step). Denote $\mathbf{H}^g(k)$ as the k -th CSI sample matrix for $\mathcal{U}^g$, which is a $|\mathcal{U}^g| \times N_T$ complex channel matrix. Let $\mathbf{W}_{\text{basis}}^g(k)$ denote the ZF basis matrix for all UEs $u \in \mathcal{U}^g$ based on CSI sample k , forming a complex weight matrix of size $N_T \times |\mathcal{U}^g|$. Each column of $\mathbf{W}_{\text{basis}}^g(k)$ represents the ZF basis vector for a UE $u \in \mathcal{U}^g$. $\mathbf{W}_{\text{basis}}^g(k)$ can be efficiently computed by pseudo-inverting $\mathbf{H}^g(k)$.

We generate the initial BF solution by linearly combining the basis matrices $\mathbf{W}_{\text{basis}}^g(k)$ across K CSI samples, assigning each ZF basis matrix an equal weight. For $g \in \mathcal{Q}$, denote $\hat{\mathbf{W}}^g$ as the initial BF matrix for all scheduled UEs in RBG g . It is computed as:

$$\hat{\mathbf{W}}^g = \sum_{k=1}^K \mathbf{W}_{\text{basis}}^g(k),$$

where $\hat{\mathbf{W}}^g$ is of size $N_T \times |\mathcal{U}^g|$. Denote $\hat{\mathbf{w}}_u^g$ as the initial precoding vector for UE $u \in \mathcal{U}^g$, corresponding to a column in $\hat{\mathbf{W}}^g$.

To scale these initial BF solutions appropriately, we consider how power is allocated to each UE over its scheduled RBGs. Specifically, We propose to perform power allocation for each UEs $u \in \mathcal{U}^g$ to meet their minimum SINR requirement γ_u^{th} . Here, γ_u^{th} is based on the selected MCS level from the previous step (using Table 5.1.3.1-1 [12]).

Denote $\hat{p}_u^g$ as the minimum required power to be allocated to UE $u \in \mathcal{U}^g$ to meet the minimum SINR requirement γ_u^{th} . That is, $\gamma_u^{\text{th}} = \frac{\hat{p}_u^g \hat{\mathbf{h}}_u^g}{\sigma^2}$, where $\hat{\mathbf{h}}_u^g$ is defined in (14) and σ^2 is the noise power. Clearly, this is an estimate and assumes the BF solution can null all interference. We have $\hat{p}_u^g = \frac{\gamma_u^{\text{th}} \sigma^2}{\hat{\mathbf{h}}_u^g}$.

Denote p_u^g as the final power allocated to UE u on RBG g . Clearly, it should be allocated in proportion to $\hat{p}_u^g$. We have:

$$p_u^g = \frac{\hat{p}_u^g}{\sum_{g \in \mathcal{Q}} \sum_{u \in \mathcal{U}^g} \hat{p}_u^g} P_{\text{max}}.$$

Based on the above power allocation, we now perform scaling on $\hat{\mathbf{w}}_u^g$ to obtain $\mathbf{w}_u^g$ for all $u \in \mathcal{U}^g$ as:

$$\mathbf{w}_u^g = \hat{\mathbf{w}}_u^g \cdot \sqrt{\frac{p_u^g}{\|\hat{\mathbf{w}}_u^g\|_2^2}},$$

and for $u \notin \mathcal{U}^g$, set $\mathbf{w}_u^g = 0$.

E. Step 5: Reality Check and Updates

In this step, we verify if the current RBG allocation, MCS selection, and BF design meet the data rate constraint (13) for each UE. That is, we calculate the achieved data rate for each UE w.r.t. K data samples and check whether or not at-least $\lceil K(1 - \epsilon_u) \rceil$ indicator functions in (13) are true for each UE $u \in \mathcal{U}$. If all UEs meet their requirement, Rudra terminates with the current solution. Otherwise, for those UEs whose data rate requirements are not satisfied, their required number of RBGs b_u is incremented by 1. Then Rudra increases the total RBG usage q by 1 and loops back to step 2 (RBG allocation). We elaborate on the details of this step in the rest of this section.

To evaluate (13) for UE u , we first need to compute $\hat{\gamma}_u^g(k)$. This entails solving the minimization problem (11). Given its enormous complexity, we instead find a lower bound for $\hat{\gamma}_u^g(k)$ and use this lower bound for $\hat{\gamma}_u^g(k)$.

An easy lower bound can be obtained by minimizing the numerator $|(\mathbf{w}_u^g)^\dagger \mathbf{y}_u^g|^2$ and maximizing each $|(\mathbf{w}_i^g)^\dagger \mathbf{y}_u^g|^2$ term in the denominator in (11), both subject to the constraint $\|\mathbf{y}_u^g - \hat{\mathbf{h}}_u^g(k)\|_2 \leq \theta_u^g$. We have:

$$\begin{aligned} \text{(OPT-N)} \quad & \min_{\mathbf{y}_u^g} |(\mathbf{w}_u^g)^\dagger \mathbf{y}_u^g|^2 \\ \text{s.t.} \quad & \|\mathbf{y}_u^g - \hat{\mathbf{h}}_u^g(k)\|_2 \leq \theta_u^g. \end{aligned}$$

$$\begin{aligned} \text{(OPT-D)} \quad & \max_{\mathbf{y}_u^g} |(\mathbf{w}_i^g)^\dagger \mathbf{y}_u^g|^2 \\ \text{s.t.} \quad & \|\mathbf{y}_u^g - \hat{\mathbf{h}}_u^g(k)\|_2 \leq \theta_u^g. \end{aligned}$$

Since both OPT-N and OPT-D are convex, we can easily find their optimal solutions. For OPT-N, the closed-form expression for the optimal solution $\mathbf{y}_u^{g,*}$ is:

$$\mathbf{y}_u^{g,*} = \hat{\mathbf{h}}_u^g(k) - \theta_u^g \frac{\mathbf{w}_u^g}{\|\mathbf{w}_u^g\|_2} e^{j \cdot \angle((\mathbf{w}_u^g)^\dagger \hat{\mathbf{h}}_u^g(k))}, \quad (15)$$

where $\angle$ represents the phase of the argument in $(\cdot)$. Similarly, for OPT-D, the closed-form expression for the optimal solution for $i \neq u$, denoted as $\mathbf{y}_{u(i)}^{g,*}$, is:

$$\mathbf{y}_{u(i)}^{g,*} = \hat{\mathbf{h}}_u^g(k) + \theta_u^g \frac{\mathbf{w}_i^g}{\|\mathbf{w}_i^g\|_2} e^{j \cdot \angle((\mathbf{w}_i^g)^\dagger \hat{\mathbf{h}}_u^g(k))}. \quad (16)$$

We can substitute the optimal solutions (15) and (16) into the objective function of OPT-N and OPT-D respectively to derive their optimal objective values. By inserting these objective values into the numerator and denominator of (11), we can find a lower bound for $\hat{\gamma}_u^g(k)$.

With the lower bound for achieved SINR $\hat{\gamma}_u^g(k)$ for UE u , we compute the data rate $\hat{r}_{u,m}^g(k)$ across all scheduled RBGs based on (12). We then check the feasibility by evaluating the indicator function in (13) across K CSI samples for each UE. If the sum of the indicator function values is at least

TABLE I
SIMULATION PARAMETERS

BS	
Number of transmit antennas (N_T)	8
Maximum transmit power ($P_{\max}$)	46 dBm
5G numerology	0
Bandwidth per RB	180 kHz
Number of RBs per RBG	4
UE	
Number of UEs (U)	20 (Randomly distributed)
Data rate requirement (R_u)	Random from $[4, 7]$ Mbps
Risk level (ϵ_u)	0.1
Thermal noise (σ^2)	-150 dBm/Hz

$\lceil K(1 - \epsilon_u) \rceil$, the data rate requirement is satisfied and Rudra returns the current x_u^g , z_u^m , and $\mathbf{w}_u^g$ as the final solution with objective value q . Otherwise, for UEs whose data rates are not satisfied, we increment their b_u 's by 1, update the total RBG q by 1, and repeat steps 2–5 until all UEs' data rate requirements are satisfied.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of Rudra.

A. Simulation Settings

A network topology with one BS and 20 UEs used in the simulation is shown in Fig. 4(a). The BS is placed at the center of the circle while the 20 UEs are randomly deployed within a radius of 300 m from the center. The key parameters at the BS and the UEs are given in Table. I.

The wireless channel is modeled by path loss and Rayleigh fading. Specifically, the path loss (in dB) is given by: $38 + 30 \times \log_{10}(d_u)$, where d_u is the distance between the BS and UE u (in meter). To simulate the estimation error in the collected CSI samples, we use a truncated Gaussian distribution [9]. We initialize the Gaussian distribution with a mean of 0 and a variance of 0.1, followed by truncation at three standard deviations from the mean. It should be noted that both the channel model and estimation error model described are only used for parameter generation in our simulation studies. Rudra relies solely on limited CSI samples and are “blindfolded” (unaware) to any distribution knowledge.

We collect $K = 48$ CSI samples per $\mathbf{h}_u^g$. We set radius for our ambiguity set as $\theta_u^g = 9.48 \times 10^{-7}$ for all $u \in \mathcal{U}, g \in \mathcal{Q}$.

All simulations were conducted on a MacBook Pro laptop with a 2.3 GHz 8-core Intel Core i9 processor and 32 GB of 2667 MHz DDR4 RAM, using MATLAB R2021a.

B. Benchmark

To the best of our knowledge, no existing algorithm directly addresses our problem. The closest is the Unified Algorithm by Zhang, *et al.* in [4], which maximizes system throughput via greedy RBG allocation and MCS selection, assuming perfect channel knowledge. To extend this algorithm for our problem, we use the average of available CSI data samples for the channel. Since the Unified Algorithm does not address BF design, we incorporate ZF for BF. To make a fair comparison with Rudra, we modify the Unified Algorithm to meet the UEs' data rate requirements and change its objective to minimize

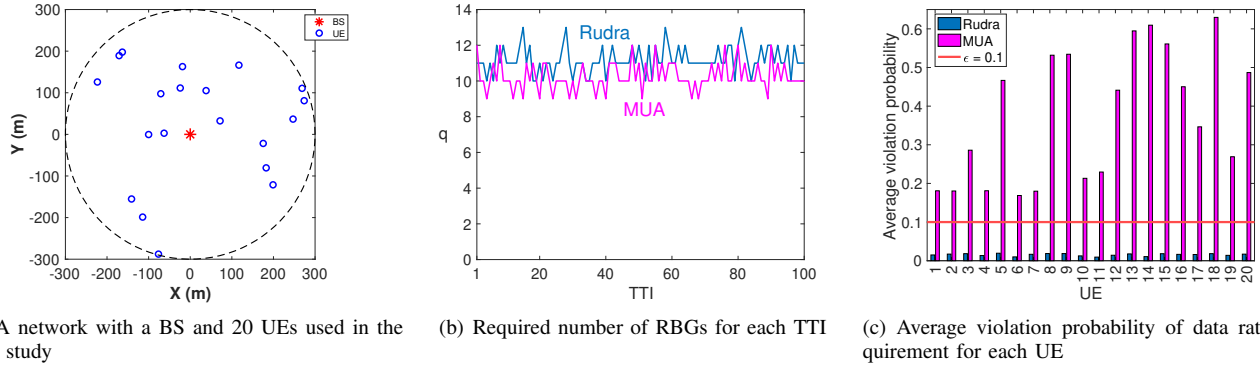


Fig. 4. Performance evaluation of Rudra.

bandwidth usage. We call the modified algorithm “Modified Unified Algorithm” (or MUA).

C. Case Study

In this section, we analyze Rudra’s performance through a case study. We set $\epsilon = 0.1$ and simulate over 10,000 TTIs.

Bandwidth Performance: Firstly, we present the bandwidth performance of Rudra (q). Figure 4(b) shows the required number of RBGs (q) from Rudra and that under MUA over 100 TTIs. As shown in the figure, the required number of RBGs (q) under MUA is slightly lower than Rudra across the 100 TTIs. This is because MUA does not offer any guarantee on UEs’ data rate requirement (as we shall see later in Fig. 4(c)). But the difference between the two is not very significant.

Probabilistic Data Rate Guarantee: Next, we show the average violation probability of data rate requirement for each UE under Rudra and those under MUA. Figure 4(c) shows the results for each UE, which are averaged over 10,000 TTIs. Clearly, Rudra’s violation probability meets the target $\epsilon = 0.1$ for all UEs, whereas MUA’s violation probabilities exceeds $\epsilon = 0.1$ for all UEs (with some exceeding 0.5, or 50%). This is because MUA does not incorporate chance constraints (by taking into considerations of channel and estimation uncertainty) in its algorithm design.

VII. CONCLUSIONS

In this paper, we investigated a spectrum usage minimization problem with probabilistic data rate guarantees to UEs. Our proposed solution, codenamed Rudra, addressed this problem with the following innovative features: i) coping with CSI uncertainty by using only limited CSI data samples; ii) reformulation of a CCP into a deterministic MINLP using an error-embedded (EE)-Wasserstein ambiguity set; iii) an iterative algorithm to solve the deterministic spectrum minimization problem through RBG allocation, MCS selection, and BF design. Simulations showed that Rudra achieves our design objectives, outperforms a customized algorithm based on the state-of-the-art.

REFERENCES

- [1] 3GPP, *5G; Service requirements for the 5G system (3GPP TS 22.261, Version 17.11.0, Release 17, Oct. 2022)*, Available: https://www.etsi.org/deliver/etsi_ts/122200_122299/122261/17.11.00_60/ts_122261v171100p.pdf (Last accessed: Jan. 10, 2024).
- [2] Federal Communications Commission, “Advancing Understanding of Non-Federal Spectrum Usage,” Available: <https://docs.fcc.gov/public/attachments/FCC-23-63A1.pdf> (Last accessed: June 26, 2024).
- [3] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*. Cambridge, U.K.: Cambridge University Press, 2007.
- [4] H. Zhang, N. Prasad, and S. Rangarajan, “MIMO downlink scheduling in LTE systems,” in *Proc. IEEE INFOCOM*, pp. 2936–2940, Orlando, FL, USA, Mar. 25–30, 2012.
- [5] Y. Shi, J. Zhang, and K. B. Letaief, “Optimal Stochastic Coordinated Beamforming for Wireless Cooperative Networks With CSI Uncertainty,” *IEEE Trans. on Signal Processing*, vol. 63, no. 4, pp. 960–973, Feb. 2015.
- [6] M. B. Shenoouda and T. N. Davidson, “Nonlinear and Linear Broadcasting With QoS Requirements: Tractable Approaches for Bounded Channel Uncertainties,” *IEEE Trans. on Signal Processing*, vol. 57, no. 5, pp. 1936–1947, May 2009.
- [7] T. Weber, A. Sklavos, and M. Meurer, “Imperfect channel-state information in MIMO transmission,” *IEEE Trans. on Commun.*, vol. 54, no. 3, pp. 543–552, March 2006.
- [8] K. Ko and J. Lee, “Multiuser MIMO User Selection Based on Chordal Distance,” *IEEE Trans. on Commun.*, vol. 60, no. 3, pp. 649–654, March 2012.
- [9] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, “Massive MIMO Performance With Imperfect Channel Reciprocity and Channel Estimation Error,” *IEEE Trans. on Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.
- [10] M. B. Shenoouda, T. N. Davidson, and L. Lampe, “Outage-based design of robust Tomlinson–Harashima transceivers for the MISO downlink with QoS requirements,” *Signal Processing*, vol. 93, no. 12, pp. 3341–3352, Dec. 2013.
- [11] J. Zhang, M. You, G. Zheng, I. Krikidis, and L. Zhao, “Model-Driven Learning for Generic MIMO Downlink Beamforming With Uplink Channel Information,” *IEEE Trans. on Wireless Commun.*, vol. 21, no. 4, pp. 2368–2382, April 2022.
- [12] 3GPP, *5G; NR; Physical layer procedures for data (3GPP TS 38.214, Version 17.8.0, Release 17, Feb. 2024)*, Available: <https://portal.3gpp.org/desktopmodules/Specifications> (Last accessed: Mar. 26, 2024).
- [13] —, *5G; Management and orchestration; Concepts, use cases and requirements (3GPP TS 28.530, Version 18.0.0, Release 18, May 2024)*, Available: <https://portal.3gpp.org/desktopmodules/Specifications> (Last accessed: August 26, 2024).
- [14] S. Acharya, S. Li, N. Jiang, Y. Wu, Y.T. Hou, W. Lou, and W. Xie, “Mitra: An O-RAN based Real-Time Solution for Coexistence between General and Priority Users in CBRs,” in *Proc. IEEE MASS*, pp. 295–303, Toronto, ON, Canada, Sep. 25–27, 2023.
- [15] N. Jiang and W. Xie, “ALSO-X and ALSO-X+: Better convex approximations for chance constrained programs,” *Operations Research*, vol. 70, no. 6, pp. 3581–3600, Feb. 2022.