

D²BF—Data-Driven Beamforming in MU-MIMO with Channel Estimation Uncertainty

Shaoran Li Nan Jiang Yongce Chen Y. Thomas Hou Wenjing Lou Weijun Xie
Virginia Tech, Blacksburg, VA, 24061

Abstract—Accurate estimation of Channel State Information (CSI) is essential to design MU-MIMO beamforming. However, errors in CSI estimation are inevitable in practice. State-of-the-art works model CSI as random variables and assume certain specific distributions or worst-case boundaries, both of which suffer performance issues when providing performance guarantees to the users. In contrast, this paper proposes a Data-Driven Beamforming (D²BF) that directly handles the available CSI data samples (without assuming any particular distributions). Specifically, we employ chance-constrained programming (CCP) to provide probabilistic data rate guarantees to the users and introduce ∞ -Wasserstein ambiguity set to bridge the unknown CSI distribution with the available (limited) data samples. Through problem decomposition and a novel bilevel formulation for each subproblem, we show that each subproblem can be solved by binary search and convex approximation. We also validate that D²BF offers better performance than the state-of-the-art approach while meeting probabilistic data rate guarantees to the users.

I. INTRODUCTION

Multi-user MIMO (MU-MIMO) beamforming is a key technology component for 5G/next-G [1], [2] and requires Channel State Information (CSI) between the Base Station (BS) and the User Equipments (UEs) [3], [4], [5], [6], [7]. Since CSI is obtained through channel training based on pre-defined signals (e.g., pilots), estimation errors are inevitable due to noise and finite length of training symbols [8], [9]. Further, CSI estimation procedures designed in either Frequency Division Duplex (FDD) systems [10], [11] or Time Division Duplex (TDD) systems [12], [13], [14] introduce errors due to limited feedback or hardware imbalance. So CSI estimation is bound to embed errors [15], and must be carefully addressed when optimizing MU-MIMO beamforming to provide performance guarantees to the UEs.

State-of-the-art approaches to address CSI estimation errors in MU-MIMO beamforming mainly fall into two categories. In the first category, CSI errors are assumed to follow certain simplified distributions, such as Gaussian [16], [17], [18], [19], or uniform distributions [17]. However, such assumed distribution models may be far from accurate due to the complex operating environment. In the second category, CSI errors are assumed to be within some worst-case boundaries, such as norm boundaries [20] or ellipsoid uncertain set [21]. However, it is well known that solutions based on worst-case boundaries are overly conservative.

In this paper, we propose a data-driven approach to design MU-MIMO beamforming in presence of channel estimation

uncertainty. Our approach offers probabilistic data rate guarantees through chance-constrained programming (CCP) based on a limited number of CSI data samples. Since we do not assume any specific channel models and are solely based on CSI data samples, our approach is applicable to a wide range of practical settings [22]. Note that although probabilistic data rate guarantees have been studied in the past (see, e.g., [16], [17], [18]), none of these studies relied only on limited CSI data samples to achieve their performance guarantees. Our main contributions are summarized as follows:

- We study an MU-MIMO beamforming problem through effective use of limited CSI data samples. Our objective is to minimize power consumption at the BS while providing probabilistic data rate guarantees to the UEs. To the best of our knowledge, this is the first work that can offer probabilistic guarantees (through CCP) in MU-MIMO beamforming based on limited CSI data samples without assuming any knowledge of distributions.
- We show that the original problem (involving all resource block groups (RBGs)) can be decomposed into independent and smaller subproblems where each subproblem corresponds to beamforming on one RBG. We show that the optimal solution to the original problem (involving all RBGs) can be recovered exactly by simply combining the optimal solutions to all the subproblems involving individual RBGs. Based on this decomposition, we show that it is sufficient to design a beamforming solution by solving one subproblem (for one RBG).
- To replace the unknown CSI distribution in the original formulation with the available (but limited) CSI data samples, we introduce ∞ -Wasserstein distance that can quantify the differences between empirical and unknown distributions. By designing an upper bound of the ∞ -Wasserstein distance, we construct ∞ -Wasserstein ambiguity set that contains the unknown CSI distribution. We show how to replace the unknown distribution in CCP with empirical distribution (from CSI data samples) and additional constraints based on properties of ∞ -Wasserstein ambiguity set.
- For the new formulation that only involves empirical distribution based on CSI data samples, we propose a solution called Data-Driven Beamforming (D²BF). D²BF breaks up the complex formulation into a bilevel optimization problem with its upper-level problem being a

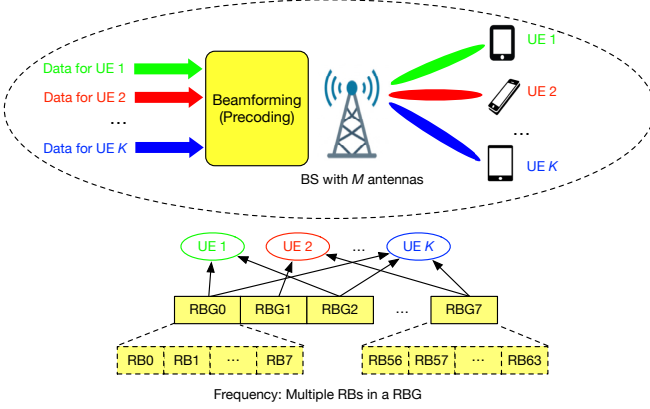


Fig. 1. An example of downlink MU-MIMO beamforming in a 5G cell

simple feasibility check and leaving all its complexity to the lower-level problem (non-convex). We propose a convex approximation for the lower-level problem and show that the overall time complexity of D^2BF remains polynomial.

- Through simulations, we show that D^2BF minimizes the power consumption while successfully providing probabilistic data rate guarantees for the UEs using the available CSI data samples. In particular, it achieves 27% power saving on average compared to the state-of-the-art approach.

II. SYSTEM MODEL AND PROBLEM STATEMENT

A. System Model

Consider a downlink MU-MIMO beamforming problem where a 5G BS needs to transmit different data streams to different UEs simultaneously, as shown in Fig. 1. Without loss of generality, we assume that each UE has one receive antenna and receives one unique data stream from the BS. For MU-MIMO beamforming, we assume that the BS employs the widely used linear precoding scheme and designs a unique precoding vector for each UE.

Following 5G terminology, the time domain is divided into Transmission Time Intervals (TTIs) and the frequency domain is divided into sub-carriers. As defined in 3GPP standards [23], 12 sub-carriers in one TTI is called a Resource Block (RB) and multiple contiguous RBs can be grouped into an RBG for scheduling and beamforming. The number of RBs in an RBG may vary from 2 to 16 and is chosen by the BS. For instance, Fig. 5 shows an example of 64 RBs grouped into 8 RBGs (i.e., 8 RBs per RBG).

The BS collects CSI on all RBs through a channel training procedure based on known signals such as pilots. Then the BS will perform user selection and beamforming for the UEs. User selection schedules RBGs to the UEs while beamforming calculates the downlink precoding vectors. After that, the precoding vectors are used for the next downlink transmission to the UEs. To keep complexity under control, one often decouples these two steps (see, e.g., [24], [25]). Following this decoupled approach, we assume that the user selection

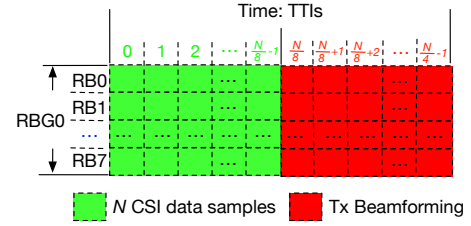


Fig. 2. An example of a sliding window with N CSI data samples

step is completed and the subset of UEs on each RBG are given *a priori* before we design the precoding vectors.

Channel Estimation Uncertainty As discussed in Section I, there are many unknown factors that contribute to the obtained CSI such as channel estimation errors, limited feedback, and hardware imbalance. So existing approaches assuming accurate distribution knowledge of CSI or worst-case boundaries have performance limitations. In contrast, we assume no knowledge of CSI distribution and instead follow a data-driven approach by directly working with the available CSI data samples. Since CSI is highly correlated among contiguous RBs and TTIs, we assume that the CSI for all the RBs within the “super” RBG-TTI window block follows the same distribution (see, e.g., the green and red superblock in Fig. 2) [26], [27].

Based on this assumption, we employ a sliding window to aggregate CSI data samples in the first half of the sliding window and use them to design precoding vectors for the RBs in the second half of the sliding window. Fig. 2 illustrate this idea, where the total number of RBs in the window is $2N$. So the number of CSI data samples is N (green blocks) and we design one set of precoding vectors for the next N RBs (red blocks) as allowed by 3GPP standards [23]. Note that this sliding window is a general form of the widely used “block-fading” model [28], where CSI is assumed to be constant on each block (a group of RBs) but are completely independent on different blocks. The main difference here is that the CSI is a random variable in our setting and we only have limited CSI data samples (no distribution knowledge).

B. Problem Statement

We assume that each UE in the cell has a minimum data rate requirement. Our goal is to design precoding vectors for downlink MU-MIMO so that the power consumption at the BS is minimized. Due to channel estimation uncertainty, it is not always possible to satisfy each UE’s data rate requirement in each time window. So we will offer a probabilistic guarantee that each UE’s data rate requirement is satisfied with at least a target probability (e.g., over 90%).

III. MATHEMATICAL FORMULATION AND ANALYSIS

A. Problem Formulation

Referring to Fig. 1, denote M as the number of antennas at the BS and $\mathcal{K} = \{1, 2, 3, \dots, K\}$ as a set of K UEs served by the BS. Denote $\mathcal{G} = \{1, 2, \dots, G\}$ as the set of G RBGs. For

RBG $g \in \mathcal{G}$, denote $\mathcal{K}^g$ as the subset of UEs that are selected to receive data on RBG g . For the precoding vectors, denote $\mathbf{w}_i^g$ (an $M \times 1$ complex column vector) as the precoding vector for UE i on RBG g .

There are two requirements for feasible precoding vectors: (i) not to exceed the maximum power budget at the BS on all RBGs; and (ii) provide probabilistic data rate guarantees to the UEs. To formulate (i), denote $P^{\max}$ as the maximum power budget at the BS for all RBGs. We have

$$\sum_{g \in \mathcal{G}} \sum_{i \in \mathcal{K}^g} \|\mathbf{w}_i^g\|_2^2 \leq P^{\max}, \quad (1)$$

where $\|\cdot\|_2$ is the L_2 -norm.

As for (ii), we assume each UE has a probabilistic data rate requirement. Per 3GPP standards [23], a UE must use the same Modulation and Coding Scheme (MCS) across all its assigned RBGs in a TTI. Thus, a UE must maintain the same SINR threshold for all its assigned RBGs. This means that each RBG that transmits to UE i contributes the same data rate. Therefore, given the fixed UE selection results, supporting a given data rate is equivalent to maintaining an SINR threshold for UE i . Denote γ_i^{req} as the target SINR threshold, which can be easily calculated based on the selected MCS and Shannon Theorem.

Denote γ_i^g as the actual SINR at UE i on RBG g with the given precoding vectors $\mathbf{w}_i^g$'s. Denote $\mathbf{h}_i^g$ (an $M \times 1$ complex column vector) as the uncertain CSI from the BS to UE i on RBG g . Due to channel estimation uncertainty, we assume $\mathbf{h}_i^g$'s are random variables following some unknown distributions. Then we have

$$\gamma_i^g = \frac{|(\mathbf{w}_i^g)^H \mathbf{h}_i^g|^2}{\sum_{j \in \mathcal{K}^g, j \neq i} |(\mathbf{w}_j^g)^H \mathbf{h}_i^g|^2 + \sigma_i^2} \quad (i \in \mathcal{K}^g, g \in \mathcal{G}), \quad (2)$$

where $(\cdot)^H$ denotes conjugate transpose. σ_i^2 is the power of thermal noise at UE i and is the same for all RBGs at UE i .

In constraints (2), $\mathbf{w}_i^g$'s are deterministic decision variables, σ_i^2 's are deterministic parameters, and $\mathbf{h}_i^g$'s are random variables. Therefore, γ_i^g 's are also random variables. As discussed in Section II, we aim to provide probabilistic SINR guarantees for the UEs, which can be written as

$$\mathbb{P}\{\gamma_i^g \geq \gamma_i^{\text{req}}\} \geq 1 - \epsilon_i \quad (i \in \mathcal{K}^g, g \in \mathcal{G}), \quad (3)$$

where $\mathbb{P}\{\cdot\}$ denotes the probability function, ϵ_i is called *risk level* and is the upper bound of the SINR threshold violation probability for UE i . Constraints (3) mean that the actual SINR γ_i^g on RBG g should be greater or equal than the required SINR threshold γ_i^{req} with a probability at least $1 - \epsilon_i$.

Substituting (2) into (3), we have

$$\mathbb{P}\left\{\frac{|(\mathbf{w}_i^g)^H \mathbf{h}_i^g|^2}{\gamma_i^{\text{req}}} \geq \sum_{j \in \mathcal{K}^g, j \neq i} |(\mathbf{w}_j^g)^H \mathbf{h}_i^g|^2 + \sigma_i^2\right\} \geq 1 - \epsilon_i \quad (i \in \mathcal{K}^g, g \in \mathcal{G}). \quad (4)$$

For the uncertain CSI $\mathbf{h}_i^g$'s, we only have their data samples at the BS, but no knowledge of their distributions. Specifically,

referring to Fig. 2, the BS has N CSI data samples per $\mathbf{h}_i^g$. Denote $\mathbb{P}_{\mathbf{h}_i^g}$ as the probability density function (PDF) of the unknown distribution of $\mathbf{h}_i^g$, i.e., $\mathbf{h}_i^g \sim \mathbb{P}_{\mathbf{h}_i^g}$. So the N data samples of $\mathbf{h}_i^g$ are drawn from the unknown distribution $\mathbb{P}_{\mathbf{h}_i^g}$. Based on the above discussion, our power minimization problem can be stated as follows:

$$\begin{aligned} \text{(P1)} \quad & \min_{\mathbf{w}_i^g} \sum_{g \in \mathcal{G}} \sum_{i \in \mathcal{K}^g} \|\mathbf{w}_i^g\|_2^2 \\ \text{s.t.} \quad & \text{BS power budget (1)}, \\ & \text{Probabilistic SINR guarantees (4)}, \\ & \text{Unknown distribution: } \mathbf{h}_i^g \sim \mathbb{P}_{\mathbf{h}_i^g}, N \text{ } \mathbf{h}_i^g \text{ samples}, \end{aligned}$$

There are two difficulties in P1. First, from the formulation of P1, it seems that the beamforming vectors on all RBGs are coupled together due to the objective function and constraint (1). Second, it is not clear how to calculate the probabilistic SINR guarantees in constraints (4), especially due to uncertain $\mathbf{h}_i^g$ (from unknown distribution $\mathbb{P}_{\mathbf{h}_i^g}$). In the rest of this section, we will address the first issue and leave the second issue to Section IV and Section V.

B. Problem Decomposition

In this subsection, we show that P1 can be decomposed into G independent subproblems and each subproblem corresponds to MU-MIMO beamforming on an RBG.

Preliminaries Note that the RBGs in the objective function of P1 are already independent. So the only issue in P1 is constraint (1), which ties all the transmission powers among the RBGs with a peak sum value. Mathematically, it merely provides an upper bound on the objective function. Consider a new problem, called P2, by ignoring constraint (1) in P1. We have

$$\begin{aligned} \text{(P2)} \quad & \min_{\mathbf{w}_i^g} \sum_{g \in \mathcal{G}} \sum_{i \in \mathcal{K}^g} \|\mathbf{w}_i^g\|_2^2 \\ \text{s.t.} \quad & \text{Probabilistic SINR guarantees (4)}, \\ & \text{Unknown distribution: } \mathbf{h}_i^g \sim \mathbb{P}_{\mathbf{h}_i^g}, N \text{ } \mathbf{h}_i^g \text{ samples}, \end{aligned}$$

Comparing P1 and P2, we have the following lemma:

Lemma 1: Suppose P2 has an optimal solution, then either this solution is an optimal solution to P1 or P1 is infeasible.

The proof is based on the facts that the feasible region of P1 (if exists) falls into that of P2 and that both P1 and P2 share the same minimization objective function. We omit the proof to conserve space. Lemma 1 suggests that we can focus on P2 to derive a solution for P1. After we obtain an optimal solution to P2, we can simply recover an optimal solution to P1 by checking constraint (1) or declare P1 is infeasible.

Decomposition We will show that instead of solving the precoding vectors on all RBGs in P2 jointly, we can first solve the precoding vectors on each RBG independently (and thus in parallel) and then simply combine these precoding vectors. The idea behind this decomposition is that both the objective function and constraints (4) in P2 can be decomposed among the RBGs, as discussed below.

For the objective function of P2, $\sum_{i \in \mathcal{K}^g} \|\mathbf{w}_i^g\|_2^2$ represents the transmission power on RBG g w.r.t. UEs in $\mathcal{K}^g$. Clearly, $\sum_{i \in \mathcal{K}^g} \|\mathbf{w}_i^g\|_2^2$ only depends on RBG g and not other RBGs. Thus, we can rewrite the objective function of P2 as

$$\sum_{g \in \mathcal{G}} \left(\min_{\mathbf{w}_i^g} \sum_{i \in \mathcal{K}^g} \|\mathbf{w}_i^g\|_2^2 \right). \quad (5)$$

This means we can decompose this objective function into G items with the g -th item corresponding to the transmission power on RBG g . For ease of exposition, let us define an $M \times M$ symmetric matrix $\mathbf{W}_i^g = \mathbf{w}_i^g (\mathbf{w}_i^g)^H$, where $\mathbf{W}_i^g$ is positive semidefinite and has rank 1. These properties can be written into the following constraints:

$$\mathbf{W}_i^g \succeq \mathbf{0}, \text{Rank}(\mathbf{W}_i^g) = 1 \quad (i \in \mathcal{K}^g), \quad (6)$$

where $\succeq$ represents positive semidefinite. Then the objective function of P2 can be rewritten as

$$\sum_{g \in \mathcal{G}} \left(\min_{\mathbf{W}_i^g} \sum_{i \in \mathcal{K}^g} \text{Tr}(\mathbf{W}_i^g) \right). \quad (7)$$

Now we consider constraints (4). We can divide them into G groups, where the g -th group corresponds to the probabilistic SINR guarantees on RBG g , i.e.,

$$\mathbb{P} \left\{ \frac{|(\mathbf{w}_i^g)^H \mathbf{h}_i^g|^2}{\gamma_i^{\text{req}}} \geq \sum_{j \in \mathcal{K}^g} |(\mathbf{w}_j^g)^H \mathbf{h}_i^g|^2 + \sigma_i^2 \right\} \geq 1 - \epsilon_i \quad (i \in \mathcal{K}^g). \quad (8)$$

Since $|(\mathbf{w}_i^g)^H \mathbf{h}_i^g|^2 = (\mathbf{h}_i^g)^H \mathbf{W}_i^g \mathbf{h}_i^g$, constraints (8) can be rewritten as

$$\mathbb{P}\{f(\mathbf{W}_i^g, \mathbf{h}_i^g) \leq 0\} \geq 1 - \epsilon_i \quad (i \in \mathcal{K}^g), \quad (9)$$

where $f(\mathbf{W}_i^g, \mathbf{h}_i^g)$ is defined as

$$f(\mathbf{W}_i^g, \mathbf{h}_i^g) = (\mathbf{h}_i^g)^H \cdot \left(\sum_{j \in \mathcal{K}^g} \mathbf{W}_j^g - \frac{\mathbf{W}_i^g}{\gamma_i^{\text{req}}} \right) \cdot \mathbf{h}_i^g + \sigma_i^2. \quad (10)$$

Clearly, constraints (9) represent the probabilistic SINR guarantees for RBG g , which is independent of other RBGs. By decoupling the objective function and constraints (4) for the G RBGs, the subproblem for RBG g is given as

$$\begin{aligned} \text{(P3)} \quad & \min_{\mathbf{W}_i^g} \sum_{i \in \mathcal{K}^g} \text{Tr}(\mathbf{W}_i^g) \\ \text{s.t.} \quad & \text{Probabilistic SINR guarantees for } \mathcal{K}^g, \\ & \text{Unknown distribution } \mathbf{h}_i^g \sim \mathbb{P}_{\mathbf{h}_i^g}, N \text{ samples,} \\ & \mathbf{W}_i^g \succeq \mathbf{0}, \text{Rank}(\mathbf{W}_i^g) = 1 \quad (i \in \mathcal{K}^g). \end{aligned}$$

So we have successfully decomposed P2 into G independent and smaller subproblems (P3) that can be solved in parallel. The optimal solution to P2 is merely a combination of the G optimal solutions to P3. This means that we can focus our study on a P3 instance (one RBG) to design our beamforming solution. For ease of exposition, in the rest of this paper, we will drop the superscript g when there is no confusion.

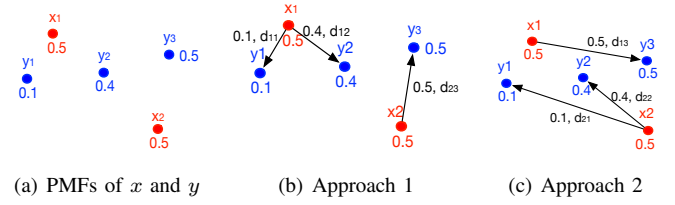


Fig. 3. Moving from distribution x to distribution y .

TABLE I
JOINT DISTRIBUTION FOR
APPROACH 1: $\mathbb{Q}_1$

Probability	x_1	x_2
y_1	0.1	0
y_2	0.4	0
y_3	0	0.5

TABLE II
JOINT DISTRIBUTION FOR
APPROACH 2: $\mathbb{Q}_2$

Probability	x_1	x_2
y_1	0	0.1
y_2	0	0.4
y_3	0.5	0

IV. BRIDGING DATA SAMPLES AND DISTRIBUTIONS

In this section, we show the relationship between data samples and distributions, which will be our basis to address channel estimation uncertainty. Denote $\mathcal{N} = \{1, 2, 3, \dots, N\}$ as the set of N data samples per $\mathbf{h}_i$. Recall that these N data samples are from the first half of the sliding window (see Fig. 2). Denote the N available data samples of $\mathbf{h}_i$ as $\hat{\mathbf{h}}_i^n$, $n \in \mathcal{N}$, where each $\hat{\mathbf{h}}_i^n$ is an $M \times 1$ complex column vector drawn from $\mathbb{P}_{\mathbf{h}_i}$ (the unknown distribution of $\mathbf{h}_i$).

Based on these N data samples, we can construct an empirical distribution for $\mathbf{h}_i$. Denote $\mathbb{P}_{\hat{\mathbf{h}}_i}$ as the probability mass function (PMF) based on the N data samples of $\mathbf{h}_i$ (i.e., $\hat{\mathbf{h}}_i^1, \hat{\mathbf{h}}_i^2, \dots, \hat{\mathbf{h}}_i^N$), given as

$$\mathbb{P}\{\hat{\mathbf{h}}_i = \hat{\mathbf{h}}_i^n\} = \frac{1}{N} \quad (n \in \mathcal{N}). \quad (11)$$

To measure the “distance” between the empirical distribution $\mathbb{P}_{\hat{\mathbf{h}}_i}$ and the true (but unknown) distribution $\mathbb{P}_{\mathbf{h}_i}$, we employ the ∞ -Wasserstein distance [29], [30].

A. ∞ -Wasserstein Distance

The origin of Wasserstein distance traces back to the optimal transport problem that finds the least efforts to transfer a given set of mines to a given set of factories [31]. Wasserstein distance is also called p -Wasserstein distance where $p \in [1, +\infty]$. In this paper, we choose ∞ -Wasserstein distance since it can offer tractable formulations for our problem.

Suppose we have two random variables ξ_1 and ξ_2 with their marginal PMFs (for continuous random variables) or PMFs (for discrete random variables) $\mathbb{P}_{\xi_1}$ and $\mathbb{P}_{\xi_2}$, respectively. To change $\mathbb{P}_{\xi_1}$ to $\mathbb{P}_{\xi_2}$, we need to move each probability mass block over a certain “distance”. Wasserstein distance measures the least effort to complete this move. We use a simple example in Fig. 3 to illustrate this idea.

Example 1. Consider moving a discrete distribution x with PMF: $\mathbb{P}\{x = x_1\} = \mathbb{P}\{x = x_2\} = 0.5$ to another discrete distribution y with PMF: $\mathbb{P}\{y = y_1\} = 0.1$, $\mathbb{P}\{y = y_2\} = 0.4$, and $\mathbb{P}\{y = y_3\} = 0.5$. There are many ways to move distribution x to distribution y and we show two of them in

Fig. 3(b) and Fig. 3(c). In this example, we use Euclidean distance to calculate the moving effort from two points such as d_{11} for moving from x_1 to y_1 . In the definition of p -Wasserstein distance, the effort of moving a probability mass 0.1 from x_1 to y_1 is 0.1 weighted by distance d_{11}^p , i.e., $(0.1 \cdot d_{11}^p)$, where $p \in [1, +\infty)$. Then the total effort of moving all probability mass blocks from x to y is the sum of all the individual efforts. For $p \geq 1$, approach 1 always requires fewer efforts than approach 2. p -Wasserstein distance is defined as the p -th root of the minimum required efforts among all possible approaches.

Mathematically, a moving approach can be mapped to a joint distribution of x and y as illustrated in Table I and Table II for two approaches respectively. So Wasserstein distance corresponds to a specific (optimal) joint distribution. Denote $\mathbb{Q}_1$ as the joint distribution of x and y in approach 1. Under the definition of p -Wasserstein distance, the effort of moving from distribution x to distribution y in approach 1 is

$$(d_{11}^p \cdot 0.1 + d_{12}^p \cdot 0.4 + d_{23}^p \cdot 0.5)^{\frac{1}{p}}. \quad (12)$$

As $p \rightarrow \infty$, (12) becomes:

$$\begin{aligned} & \lim_{p \rightarrow \infty} (d_{11}^p \cdot 0.1 + d_{12}^p \cdot 0.4 + d_{23}^p \cdot 0.5)^{\frac{1}{p}} \\ &= \max\{d_{11}, d_{12}, d_{23}\} = d_{23}, \end{aligned}$$

where the last equality holds due to $d_{23} \geq d_{12} \geq d_{11}$ in Fig. 3(b). In fact, we can show that approach 1 is the most efficient approach in terms of ∞ -Wasserstein distance between x and y , i.e., d_{23} . The physical meaning of ∞ -Wasserstein is the maximum moving distance among all steps in the optimal approach (joint distribution). This interpretation makes ∞ -Wasserstein highly tractable. ■

The formal definition of ∞ -Wasserstein distance is given as follows.

Definition 1: The ∞ -Wasserstein distance of $\mathbb{P}_{\xi_1}$ and $\mathbb{P}_{\xi_2}$ is defined as

$$W_{\infty}(\mathbb{P}_{\xi_1}, \mathbb{P}_{\xi_2}) = \inf_{\mathbb{Q} \in \mathcal{Q}} \{\sup_{\mathbb{Q}} \|\xi_1 - \xi_2\|\}, \quad (13)$$

where $\|\cdot\|$ is any norm, “sup” stands for the supremum,¹ $\mathbb{Q}$ stands for a joint distribution of ξ_1 and ξ_2 , $\mathcal{Q}$ stands for the set of all possible $\mathbb{Q}$ ’s respectively.

Though the definition (13) holds true for any norm $\|\cdot\|$, it is common to choose L_2 -norm due to its attractive computational properties, as in Example 1. It should also be clear that $W_{\infty}(\mathbb{P}_{\xi_1}, \mathbb{P}_{\xi_2}) = 0$ holds if and only if $\mathbb{P}_{\xi_1} = \mathbb{P}_{\xi_2}$ almost surely. Otherwise, we have $W_{\infty}(\mathbb{P}_{\xi_1}, \mathbb{P}_{\xi_2}) > 0$.

B. ∞ -Wasserstein Ambiguity Set

Denote θ_i as a nonnegative number. Denote $\mathcal{P}_{\infty}^W(\theta_i)$ as a set of distributions whose ∞ -Wasserstein distances from $\mathbb{P}_{\hat{\mathbf{h}}_i}$ are upper bounded by θ_i , i.e.,

$$\mathcal{P}_{\infty}^W(\theta_i) = \left\{ \mathbb{P} : W_{\infty}(\mathbb{P}, \mathbb{P}_{\hat{\mathbf{h}}_i}) \leq \theta_i, \mathbb{P} \in \mathcal{P} \right\} \quad (i \in \mathcal{K}),$$

¹The ∞ -Wasserstein distance is also defined in terms of “essential supremum” to avoid some extreme distributions [29], [31]. But such extreme distributions are not encountered in our problem, so we use “sup” instead.

where $\mathcal{P}$ stands for all possible distributions for an $M \times 1$ random vector. $\mathcal{P}_{\infty}^W(\theta_i)$ is called ∞ -Wasserstein ambiguity set [32] and can be viewed as a ball of distributions centered at $\mathbb{P}_{\hat{\mathbf{h}}_i}$ with a radius θ_i .

Suppose that we choose θ_i ’s properly such that

$$\mathbb{P}_{\mathbf{h}_i} \in \mathcal{P}_{\infty}^W(\theta_i) \quad (i \in \mathcal{K}). \quad (14)$$

Then the ∞ -Wasserstein distance between the true distribution $\mathbb{P}_{\mathbf{h}_i}$ and the empirical distribution $\mathbb{P}_{\hat{\mathbf{h}}_i}$ is upper bounded by θ_i . For our problem, θ_i represents the seriousness of CSI estimation errors, which can be set properly when UE i joins the 5G cell and be dynamically changed by tracking UE i ’s SINR violation probability. Note that choosing an overly smaller θ_i may invalidate the probabilistic guarantees for the UEs while choosing an overly larger θ_i may waste BS’s transmission power. Clearly, the first outcome is more detrimental. Nevertheless, for the purpose of designing our beamforming solution, we can consider θ_i ’s are given constants.

We now show how to reformulate P3 based on $\mathcal{P}_{\infty}^W(\theta_i)$. Combining constraints (14) and constraints (9), we have

$$\inf_{\mathbb{P}_{\mathbf{h}_i} \in \mathcal{P}_{\infty}^W(\theta_i)} \mathbb{P}\{f(\mathbf{W}_i, \mathbf{h}_i) \leq 0\} \geq 1 - \epsilon_i \quad (i \in \mathcal{K}). \quad (15)$$

Note that we have dropped superscript g for simplicity when there is no confusion. The “inf” in constraints (15) means that for any distribution $\mathbb{P}_{\mathbf{h}_i}$ from $\mathcal{P}_{\infty}^W(\theta_i)$, the probability SINR guarantees for the UEs should be valid. However, constraints (15) are still challenging due to $\mathbf{h}_i$ ’s unknown distribution $\mathbb{P}_{\mathbf{h}_i}$. Thus, we will reformulate constraints (15) by replacing $\mathbf{h}_i$ with $\hat{\mathbf{h}}_i$ as follows.

Based on the definition of ∞ -Wasserstein ambiguity set, constraints (15) are equivalent to [30]

$$\mathbb{P}\{\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i) \leq 0\} \geq 1 - \epsilon_i \quad (i \in \mathcal{K}), \quad (16)$$

where

$$\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i) = \max_{\mathbf{c}_i} \{f(\mathbf{W}_i, \mathbf{c}_i) : \|\mathbf{c}_i - \hat{\mathbf{h}}_i\|_2 \leq \theta_i\}. \quad (17)$$

Here we introduce an auxiliary variable $\mathbf{c}_i$ (a $M \times 1$ complex column vector). Note that $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i) \leq 0$ in constraints (16) means that given $\hat{\mathbf{h}}_i$, we should have $f(\mathbf{W}_i, \mathbf{c}_i) \leq 0$ holds for any $\mathbf{c}_i$ that satisfies $\|\mathbf{c}_i - \hat{\mathbf{h}}_i\|_2 \leq \theta_i$. We see that the data samples $\hat{\mathbf{h}}_i$ are used in constraints (16). Recall that $\hat{\mathbf{h}}_i$ ’s distribution $\mathbb{P}_{\hat{\mathbf{h}}_i}$ is given in (11) based on N data samples. Plugging in this distribution (11) into (16), we obtain

$$\sum_{n \in \mathcal{N}} \mathbb{I}\{\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n) \leq 0\} \geq N \cdot (1 - \epsilon_i) \quad (i \in \mathcal{K}), \quad (18)$$

where $\mathbb{I}(\cdot)$ is the binary indicator function.

Based on (18), we can rewrite P3 as

$$\begin{aligned} \text{(P4)} \quad & \min_{\mathbf{W}_i} \sum_{i \in \mathcal{K}} \text{Tr}(\mathbf{W}_i) \\ \text{s.t.} \quad & \text{Probabilistic SINR guarantees (18)}, \\ & \mathbf{W}_i \succeq \mathbf{0}, \text{Rank}(\mathbf{W}_i) = 1 \quad (i \in \mathcal{K}). \end{aligned}$$

For the rank constraints “Rank($\mathbf{W}_i$) = 1” in P4, a standard approach is to employ Semi-definite programming (SDP)

relaxation (see, e.g., [33], [34], [35]). In SDP relaxation, we first relax the rank constraints “Rank($\mathbf{W}_i$) = 1” by dropping them. Then we solve the relaxed problem based on the approach proposed in Section V. After we obtain a solution $\mathbf{W}_i$, we check its rank to recover the original $\mathbf{w}_i$ either through Eigendecomposition or Gaussian randomization based on $\mathbf{W}_i$ [36]. So the next step is to find a solution for P4 without the rank constraints, which is still challenging. Even though we have replaced the probability function in (15) with N data samples in (16), it is unclear how to handle the indicator functions and $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n)$ in constraints (18).

V. D²BF—A DATA-DRIVEN BEAMFORMING SOLUTION

In this section, we present D²BF—a Data-Driven BeamForming solution to P4. D²BF is based on a convex approximation of P4, which hinges on a bilevel formulation and a novel reformulation technique called ALSO-X+ [30].

A. Bilevel Formulation

In this section, we present a bilevel formulation of P4 that consists of an upper-level problem and a lower-level problem. This bilevel formulation is an exact reformulation of P4 after dropping its rank constraints. Under this bilevel formulation, we only need to focus on the lower-level problem since the upper-level problem is a simple feasibility check. This bilevel formulation allows us to derive a convex approximation of P4 in Section V-B.

For UE i , denote $z_i(\hat{\mathbf{h}}_i^n)$ as a binary indicator w.r.t. $\hat{\mathbf{h}}_i^n$ as:

$$z_i(\hat{\mathbf{h}}_i^n) = \mathbb{I}(\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n) \leq 0). \quad (19)$$

With $z_i(\hat{\mathbf{h}}_i^n)$, we can rewrite constraints (18) as

$$\sum_{n \in \mathcal{N}} z_i(\hat{\mathbf{h}}_i^n) \geq N \cdot (1 - \epsilon_i) \quad (i \in \mathcal{K}). \quad (20)$$

To put constraints (19) into closed-form constraints, we introduce an auxiliary function $s_i(\hat{\mathbf{h}}_i^n)$ w.r.t. $\hat{\mathbf{h}}_i^n$ such that:

$$s_i(\hat{\mathbf{h}}_i^n) \geq 0 \quad (i \in \mathcal{K}, n \in \mathcal{N}), \quad (21a)$$

$$\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n) \leq s_i(\hat{\mathbf{h}}_i^n) \quad (i \in \mathcal{K}, n \in \mathcal{N}), \quad (21b)$$

$$z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n) = 0 \quad (i \in \mathcal{K}, n \in \mathcal{N}). \quad (21c)$$

The nonnegative $s_i(\hat{\mathbf{h}}_i^n)$ can be considered as a slack function w.r.t. $z_i(\hat{\mathbf{h}}_i^n)$. It is easy to see that constraints (19) can be replaced by constraints (21).

Further, based on (20) and (21c), we have

$$\sum_{n \in \mathcal{N}} \mathbb{I}\{s_i(\hat{\mathbf{h}}_i^n) = 0\} \geq N \cdot (1 - \epsilon_i) \quad (i \in \mathcal{K}). \quad (22)$$

By introducing an auxiliary variable t , we can rewrite the objective function of P4 as “min t ” and add the following constraints:

$$\sum_{i \in \mathcal{K}} \text{Tr}(\mathbf{W}_i) \leq t. \quad (23)$$

Using this new objective function, adding constraints (23), and replacing constraints (18) by constraints (20) and (21), and

also dropping the rank constraints, we obtain a reformulation of P4 as follows:

$$\begin{aligned} \text{(P4-R)} \quad & \min_{\mathbf{W}_i, z_i(\hat{\mathbf{h}}_i^n), s_i(\hat{\mathbf{h}}_i^n)} t \\ \text{s.t.} \quad & \mathbf{W}_i \succeq 0, \text{ Constraints (20), (21), (23)}. \end{aligned}$$

P4-R suggests that we can use binary search to obtain the smallest t such that all constraints of P4-R are feasible. Then we can take this smallest t as the optimal objective value and its corresponding feasible solution as the optimal solution to P4-R. This is the basic idea of a bilevel formulation of P4-R.

Since $z_i^n(\hat{\mathbf{h}}_i) \geq 0$ and $s_i^n(\hat{\mathbf{h}}_i) \geq 0$, constraints (21c) is equivalent to

$$\sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n)\} = 0. \quad (24)$$

Since it is hard to handle the bilinear constraints (21c), we will drop them and use the left-hand side of constraints (24) as the objective function of the lower-level problem, which facilitates our derivation of a convex approximation in Section V-B. By removing constraints (21c), we need to bring constraints (22) back to the problem. We now have a bilevel formulation of P4-R as follows:

$$\begin{aligned} \text{(P5)} \quad & \min_t t \\ & (\mathbf{W}_i^*, z_i^*(\hat{\mathbf{h}}_i^n), s_i^*(\hat{\mathbf{h}}_i^n)) \in \arg \min_{\mathbf{W}_i, z_i(\hat{\mathbf{h}}_i^n), s_i(\hat{\mathbf{h}}_i^n)} \left\{ \right. \\ & \left. \sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n)\} : \mathbf{W}_i \succeq 0, \right. \\ & \left. \text{Constraints (20), (21a), (21b), (23)} \right\}, \\ & \text{Constraints (22)}. \end{aligned}$$

We see the *lower-level problem* preserves most of the constraints of P4-R. The biggest change is that constraints (21c) disappear in P5 while $\sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n)\}$ is used as the objective function of the lower-level problem. As for the *Upper-level problem*, it uses t as the objective function and add constraints (22) to ensure feasibility of the final solution. Constraints (22) have been relocated to the upper-level problem from the lower-level problem, which leads to a more tractable feasible region for the lower-level problem without introducing relaxation errors. It can be shown that P5 is an equivalent reformulation of P4-R [30]. The proof is based on the fact that an optimal solution of P4-R can be constructed based on an optimal solution to P5 with the same $\mathbf{W}_i^*$ and objective value. So is the converse.

The main idea of P5 is that for a given t , we can solve the lower-level problem to obtain an optimal solution $(\mathbf{W}_i^*, z_i^*(\hat{\mathbf{h}}_i^n)$ and $s_i^*(\hat{\mathbf{h}}_i^n))$. If $s_i^*(\hat{\mathbf{h}}_i^n)$'s satisfy constraints (22), then this $\mathbf{W}_i^*$ is a feasible solution with objective function t . Based on this understanding, the minimum t that can derive a feasible $\mathbf{W}_i^*$ is the optimal solution to P5.

Now the question is: Given t , how to find a solution to P5 $(\mathbf{W}_i, z_i(\hat{\mathbf{h}}_i^n)$ and $s_i(\hat{\mathbf{h}}_i^n))$? Since the upper-level problem of P5 is a simple feasibility check with constraints (22), we only

need to focus on the lower-level problem, which is challenging due to $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n)$ in (21b) and its bilinear objective function.

B. Solution to P5: Convex Approximation

In this section, we present an algorithm to derive a solution to P5 for a given t . Our approach is based on convex approximation of the lower-level problem in P5 using $\mathcal{S}$ -lemma [37] and a novel technique called “ALSO-X+” [30], [38].

1) *Reformulation of $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n)$* : Since $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n)$ in constraints (21b) involves a maximization problem (see (17)) which cannot be solved directly, we need to reformulate $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i^n)$. Denote an $M \times 1$ complex column vector $\mathbf{e}_i^n$ as the difference between $\mathbf{c}_i$ and data sample $\hat{\mathbf{h}}_i^n$, i.e., $\mathbf{e}_i^n = \mathbf{c}_i - \hat{\mathbf{h}}_i^n$. Based on $\mathbf{e}_i^n$ and (17), constraints (21b) can be rewritten as

$$\max_{\mathbf{e}_i^n} \{f(\mathbf{W}_i, \mathbf{e}_i^n + \hat{\mathbf{h}}_i^n) : \|\mathbf{e}_i^n\|_2 \leq \theta_i\} \leq s_i(\hat{\mathbf{h}}_i^n) \quad (i \in \mathcal{K}, n \in \mathcal{N}), \quad (25)$$

which means that

$$f(\mathbf{W}_i, \mathbf{e}_i^n + \hat{\mathbf{h}}_i^n) \leq s_i(\hat{\mathbf{h}}_i^n) \quad (i \in \mathcal{K}, n \in \mathcal{N}). \quad (26)$$

holds for any $\mathbf{e}_i^n$ s.t. $\|\mathbf{e}_i^n\|_2 \leq \theta_i$. This can also be viewed as $\|\mathbf{e}_i^n\|_2 \leq \theta_i$ implies constraints (26) hold, i.e.,

$$\|\mathbf{e}_i^n\|_2 \leq \theta_i \implies \text{Constraints (26)} \quad (27)$$

For ease of exposition, let us define an $M \times M$ symmetric matrix $\mathbf{Q}_i$ and a real scalar a_i^n as

$$\mathbf{Q}_i = \sum_{j \in \mathcal{K}} \mathbf{W}_j - \frac{\mathbf{W}_i}{\gamma_i^{\text{req}}}, \quad a_i^n = (\hat{\mathbf{h}}_i^n)^H \mathbf{Q}_i \hat{\mathbf{h}}_i^n + \sigma_i^2. \quad (28)$$

Then constraints (26) can be rewritten as

$$(\mathbf{e}_i^n)^H \mathbf{Q}_i \mathbf{e}_i^n + (\mathbf{e}_i^n)^H \mathbf{Q}_i \hat{\mathbf{h}}_i^n + (\hat{\mathbf{h}}_i^n)^H \mathbf{Q}_i \mathbf{e}_i^n + a_i^n \leq s_i(\hat{\mathbf{h}}_i^n) \quad (i \in \mathcal{K}, n \in \mathcal{N}), \quad (29)$$

Further, $\|\mathbf{e}_i^n\|_2 \leq \theta_i$ can be rewritten as

$$(\mathbf{e}_i^n)^H \mathbf{I}_M \mathbf{e}_i^n - \theta_i^2 \leq 0, \quad (30)$$

where $\mathbf{I}_M$ is the M dimensional identity matrix. Thus, replacing $\|\mathbf{e}_i^n\|_2 \leq \theta_i$ with (30) and replacing constraints (26) with constraints (29), (27) can be rewritten as

$$(\mathbf{e}_i^n)^H \mathbf{I}_M \mathbf{e}_i^n - \theta_i^2 \leq 0 \implies \text{Constraints (29)} \quad (31)$$

To reformulate statement (31), we resort to $\mathcal{S}$ -lemma [37].

Lemma 2: ($\mathcal{S}$ -Lemma) Let $\mathbf{U}$ and $\mathbf{V}$ be $L \times L$ symmetric matrices. Suppose (i) $\mathbf{x}^T \mathbf{U} \mathbf{x} \leq 0$ holds for some $\mathbf{x} \in \mathcal{X}$; and (ii) There exists an $\bar{\mathbf{x}}$ such that $\bar{\mathbf{x}}^T \mathbf{U} \bar{\mathbf{x}} < 0$. Then $\mathbf{x}^T \mathbf{V} \mathbf{x} \leq 0$ holds for $\mathbf{x} \in \mathcal{X}$ if and only if there exists a nonnegative number λ such that $\mathbf{V} \preceq \lambda \mathbf{U}$.

To see statement (31) matches with the standard form of $\mathcal{S}$ -Lemma, we define $\mathbf{x} = [(\mathbf{e}_i^n)^H \ 1]^H$, $\bar{\mathbf{x}} = [\mathbf{0} \ 1]^H$, and

$$\mathbf{U} = \begin{bmatrix} \mathbf{I}_M & \mathbf{0} \\ \mathbf{0} & -\theta_i^2 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{Q}_i & \mathbf{Q}_i \hat{\mathbf{h}}_i^n \\ (\mathbf{Q}_i \hat{\mathbf{h}}_i^n)^H & a_i^n - s_i(\hat{\mathbf{h}}_i^n) \end{bmatrix}.$$

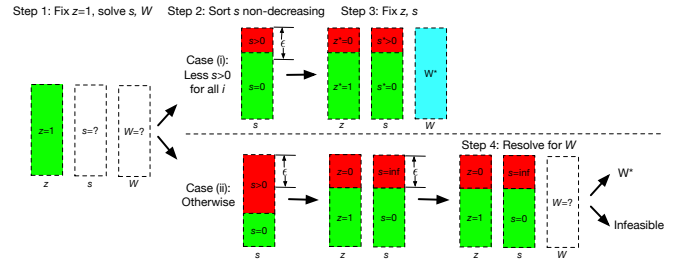


Fig. 4. A solution procedure to P5's lower-level problem for a given t

Then statement (31) holds if and only if

$$\mathbf{V} \preceq \lambda_i^n \mathbf{U}, \quad \lambda_i^n \geq 0 \quad (i \in \mathcal{K}, n \in \mathcal{N}). \quad (32)$$

Replacing statement (31) with constraints (32), the lower-level problem in P5 becomes

$$\min_{\mathbf{W}_i, z_i(\hat{\mathbf{h}}_i^n), s_i(\hat{\mathbf{h}}_i^n)} \left\{ \sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n)\} : \mathbf{W}_i \succeq 0, \right. \\ \left. \text{Constraints (20), (21a), (32), (23)} \right\}.$$

2) *Bilinear Objective Function*: As for the bilinear objective function $\min \sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n)\}$, we employ convex approximation. Our proposed solution is inspired by ALSO-X+ [30], which has been proven to offer better performance comparing to existing reformulation techniques under ∞ -Wasserstein ambiguity set [30], [38]. The main idea behind our convex approximation is to set $z_i(\hat{\mathbf{h}}_i^n)$ and $s_i(\hat{\mathbf{h}}_i^n)$ based on the value of ϵ_i , constraints (20) and (22), and the objective function $\min \sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\hat{\mathbf{h}}_i^n) \cdot s_i(\hat{\mathbf{h}}_i^n)\}$.

As shown in Fig. 4, we start the procedure by setting $z_i(\hat{\mathbf{h}}_i^n) = 1$ and solve for $s_i(\hat{\mathbf{h}}_i^n)$ and $\mathbf{W}_i$ (Step 1). This step is motivated by (20) and the value of ϵ_i (whose value tends to be small). So we would anticipate the majority of $z_i(\hat{\mathbf{h}}_i^n)$'s to be 1. Further, we choose to fix $z_i(\hat{\mathbf{h}}_i^n)$'s first as they only appear in constraints (20) and the bilinear objective function. So their impacts on other constraints are limited. With $z_i(\hat{\mathbf{h}}_i^n) = 1$, constraints (20) hold trivially. Further, the lower-level problem in P5 can be simplified to

$$\min_{\mathbf{W}_i, s_i(\hat{\mathbf{h}}_i^n)} \left\{ \sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} s_i(\hat{\mathbf{h}}_i^n) : \mathbf{W}_i \succeq 0, \right. \\ \left. \text{Constraints (21a), (32), (23)} \right\}. \quad (33)$$

This problem is convex and we can solve its optimal solution $\mathbf{W}_i^*$ and $s_i^*(\hat{\mathbf{h}}_i^n)$. Note that this problem is always feasible as in the special case when $s_i(\hat{\mathbf{h}}_i^n)$'s are sufficiently large, all constraints are trivially satisfied.

After we obtain $s_i^*(\hat{\mathbf{h}}_i^n)$'s, we sort them in non-increasing order for each UE i (Step 2). Specifically, we sort $\{s_i^*(\hat{\mathbf{h}}_i^n), n \in \mathcal{N}\}$ and denote $\mathcal{S}_i^{\text{sort}}$ as the sorted set. Then we count the number of positive numbers in $\mathcal{S}_i^{\text{sort}}$ and divide this number by N . We have two cases:

Case (i): The ratio of positive elements in $\mathcal{S}_i^{\text{sort}}$ over N is no greater than ϵ_i for all $i \in \mathcal{K}$. This is the simple case

as constraints (22) already hold. That is, there are at least $N \cdot (1 - \epsilon_i)$ number of $s_i^*(\mathbf{h}_i^n)$'s with $s_i^*(\mathbf{h}_i^n) = 0$. To minimize the objective function $\sum_{n \in \mathcal{N}} \{z_i(\mathbf{h}_i^n) \cdot s_i(\mathbf{h}_i^n)\}$, we can adjust $z_i(\mathbf{h}_i^n)$ from 1 to 0 corresponding to those $s_i^*(\mathbf{h}_i^n) > 0$, i.e.,

$$z_i^*(\mathbf{h}_i^n) = 1 - \mathbb{I}(s_i^*(\mathbf{h}_i^n) > 0). \quad (34)$$

This adjustment of $z_i(\mathbf{h}_i^n)$ is solely to achieve a minimum objective value $\sum_{n \in \mathcal{N}} \{z_i(\mathbf{h}_i^n) \cdot s_i(\mathbf{h}_i^n)\} = 0$. It has no impacts on $\mathbf{W}_i$ and $s_i(\mathbf{h}_i^n)$ since $z_i(\mathbf{h}_i^n)$'s only appear in constraints (20) and the objective function. Thus, this solution is an optimal solution to P5's lower-level problem for current t .

Case (ii): For some $i \in \mathcal{K}$, the ratio of positive elements in $\mathcal{S}_i^{\text{sort}}$ over N is greater than ϵ_i . In this case, after step 1, constraints (22) do not hold, as there are fewer number of $s_i^*(\mathbf{h}_i^n)$'s with $s_i^*(\mathbf{h}_i^n) = 0$. So we propose to adjust some $z_i(\mathbf{h}_i^n)$'s and $s_i(\mathbf{h}_i^n)$'s so that both constraints (20) and (22) hold and the objective function $\sum_{n \in \mathcal{N}} \{z_i(\mathbf{h}_i^n) \cdot s_i(\mathbf{h}_i^n)\}$ is minimized. To do this, we propose to fix $z_i(\mathbf{h}_i^n)$ and $s_i(\mathbf{h}_i^n)$ based on the sorted set $\mathcal{S}_i^{\text{sort}}$. Specifically, we set $s_i(\mathbf{h}_i^n)$ as

$$s_i(\mathbf{h}_i^n) = \begin{cases} \infty & \text{for the first } \lfloor N \cdot \epsilon_i \rfloor \text{ elements in } \mathcal{S}_i^{\text{sort}}, \\ 0 & \text{otherwise.} \end{cases} \quad (35)$$

Then we adjust $z_i(\mathbf{h}_i^n)$ using (34) and the new values for $s_i(\mathbf{h}_i^n)$ in (35), as shown in Fig 4 (Case (ii) step 3). Note that such setting of $z_i(\mathbf{h}_i^n)$ and $s_i(\mathbf{h}_i^n)$ will ensure the objective function $\sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \{z_i(\mathbf{h}_i^n) \cdot s_i(\mathbf{h}_i^n)\}$ to be 0.

In (35), when $s_i(\mathbf{h}_i^n) = \infty$, the corresponding constraints in (32) trivially hold. Since these constraints have no impacts when solving $\mathbf{W}_i$, they can be safely removed from the lower-level problem in P5 when solving for $\mathbf{W}_i$. Now we have a convex optimization problem (with 0 optimal objective value) and we can solve its optimal solution $\mathbf{W}_i^*$. If $\mathbf{W}_i^*$ can be found then it is the optimal solution to P5's lower-level problem for current t . Otherwise, we decide that the current t is infeasible.

C. Summary and Complexity Analysis

A summary of D²BF is given in Algorithm 1 that combines all the above steps including the binary search for t , solution procedure for a given t , recovering $\mathbf{w}_i^g$ from $\mathbf{W}_i^g$, and recovering solution to P1. Note that in Line 8, we choose $t^{\text{UB}} = \min\{t, \sum_{i \in \mathcal{K}^g} \mathbf{W}_i^g\}$ for faster convergence.

Complexity Analysis The binary search for t in Algorithm 1 consists of $\lceil \log_2(\frac{t^{\text{UB}} - t^{\text{LB}}}{\Delta}) \rceil$ iterations. In each iteration, the complexity is dominated by Line 6—apply the solution procedure in Fig. 4—which consists of at most two convex optimization problems. Both convex problems can be solved efficiently with polynomial complexity using off-the-shelf solvers. So D²BF has polynomial time complexity.

VI. SIMULATION RESULTS

A. Simulation Settings

We consider a 5G cell with a radius of 500 meters and 20 UEs ($K = 20$) served by the BS equipped with 8 transmit antennas ($M = 8$). Fig. 5 shows the topology, where the UEs are randomly distributed inside the 5G cell following a uniform

Algorithm 1 D²BF

```

1: Input:  $\gamma_i^{\text{req}}, P^{\text{max}}, \mathbf{h}_i^n, \theta_i, \Delta$ 
2: Output:  $\mathbf{w}_i^g$  or infeasible
3: for  $g \in \mathcal{G}$  (Subproblems on RBGs) do
4:   Set lower bound  $t^{\text{LB}}$  and upper bound  $t^{\text{UB}}$ 
5:   while  $t^{\text{UB}}/t^{\text{LB}} > 1 + \Delta$  do
6:     Set  $t = (t^{\text{UB}} + t^{\text{LB}})/2$ , apply procedure in Fig. 4
7:     if feasible  $\mathbf{W}_i^g$  found then
8:       Set  $t^{\text{UB}} = \min\{t, \sum_{i \in \mathcal{K}^g} \mathbf{W}_i^g\}$ , save  $\mathbf{W}_i^g$ 
9:     else
10:      Set  $t^{\text{LB}} = t$ 
11:    end if
12:  end while
13:  if  $\text{rank}(\mathbf{W}_i^g) \approx 1$  then
14:    Set  $\mathbf{w}_i^g$  as the eigenvector of  $\mathbf{W}_i^g$ 
15:  else
16:    Gaussian Randomization to generate  $\mathbf{w}_i^g$  using  $\mathbf{W}_i^g$ 
17:  end if
18: end for
19: Check (1) using  $\mathbf{w}_i^g$ : return  $\mathbf{w}_i^g$  or report infeasible

```

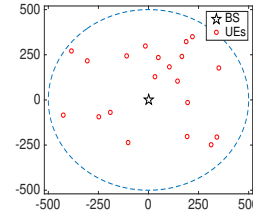


Fig. 5. Network topology: A 5G cell with 20 UEs

distribution. At the BS, we consider there are 8 RBGs ($G = 8$), with each RBG consisting of 8 RBs (see Fig. 1). We use random user selection for each RBG and suppose each RBG supports two UEs ($K^g = 2$). Without loss of generality, all UEs employ QPSK to calculate their required SINR threshold.

The BS has a power budget of $P^{\text{max}} = 46$ dBm across 8 RBGs and the thermal noise is set to -150 dBm/Hz for all UEs [39]. We assume 15 kHz sub-carrier spacing (i.e., numerology 0) as defined in 3GPP standards [26]. Same with Fig. 2, we use the CSI data samples from the first 5 TTIs to perform beamforming for the next 5 TTI. Since each RBG consists of 8 RBs, we have $N = 5 \times 8 = 40$ samples for each $\mathbf{h}_i^g$.

The wireless channel is modeled by path-loss and a truncated Gaussian distribution. The path-loss between UE i and the BS is modeled by $PL_i = 38 + 30 \times \log_{10}(d_i)$ (in dB) where d_i is the distance between UE i and the BS (in meters) [40]. For the truncated Gaussian distribution, we use 0 as mean and 0.1 as variance for the original Gaussian distribution and then truncate it at three times its standard deviation. This setting is similar to that in [14], [16]. Note that the distribution knowledge described here is only used for generating parameters in our simulations. Our proposed solution D²BF only relies on the CSI data samples and is

TABLE III
INSIGHTS OF D²BF

Risk level ϵ	0.1	0.2	0.3	0.4	0.5
Iterations	4.60	4.28	4.08	3.86	3.60
Dominant Eigenvalue (%)	99.92	99.93	99.91	99.90	99.89

blindfolded w.r.t. any knowledge of distribution information.

We use MOSEK 9.2.38 on MATLAB R2017b to run all algorithms and each solution includes beamforming vectors $\mathbf{w}_i^g$ and the objective value for P1. For benchmarking, we also run results from the following two approaches:

- State-of-the-art Gaussian Approximation [16], assumes Gaussian distribution for uncertain CSI to derive convex approximations and provides probabilistic SINR threshold guarantees to the UEs.
- Mean Approximation where the means of $\hat{\mathbf{h}}_i^n$'s are used as the perfect CSI, which leads to a deterministic formulation. Then a classical solution [3] can be employed.

B. A Case Study

Although D²BF can handle different values of ϵ_i 's for the UEs, we use the same value for all UEs (i.e. $\epsilon_i = \epsilon$, $i \in \mathcal{K}$), ranging from 0.1 to 0.5. We choose $\theta_i = 0.035$ to upper bound the ∞ -Wasserstein distance between the true (but unknown) distribution and the empirical distribution based on 40 data samples of $\mathbf{h}_i^n$. The stop criterion in D²BF is set to $\Delta = 0.03$. In D²BF, we initialize $t^{LB} = 0$ and t^{UB} to the smaller value between 46 dBm and the objective value from Gaussian Approximation. We perform 50 runs (i.e., 250 TTIs) under each ϵ and report the averages.

Table III shows the number of iterations in binary search (Line 4-12 of D²BF) and the ranks of the obtained $\mathbf{W}_i$ (Line 13-17). We see the number of iterations in the binary search for t is 4.08 on average and is non-increasing w.r.t. ϵ . This is due to a smaller search range for t under a larger ϵ and a faster convergence due to our updating mechanism of t^{UB} in D²BF (Line 8), where the search range for t can shrink more than half if feasible $\mathbf{W}_i$'s are found.

To check the ranks of $\mathbf{W}_i$, we calculate the ratio between the dominant eigenvalue over the sum of all eigenvalues for different $\mathbf{W}_i$'s. As shown in Table III, the average of this ratio is over 99.9% and the lowest ratio in our case study is 97.08% for a UE under $\epsilon = 0.4$. So we conclude that there is only one dominant eigenvalue for $\mathbf{W}_i$ in our case study, and Line 14 is always used in D²BF, similar to the observations in [16]. This is because SDP solvers such as MOSEK typically exploit low-rank structures when solving for $\mathbf{W}_i$.

We now present the performance of D²BF. Fig. 6(a) shows the actual violation probabilities averaging over all UEs. We see the violation probabilities from D²BF are smaller than the risk level ϵ . We also check the violation probabilities regarding each UE and found that UEs at the cell edge are more likely to experience higher violation probabilities (still below ϵ). This validates that D²BF provides probabilistic SINR threshold guarantees for the UEs. Further, the violation probabilities from Gaussian Approximation are much smaller than that from

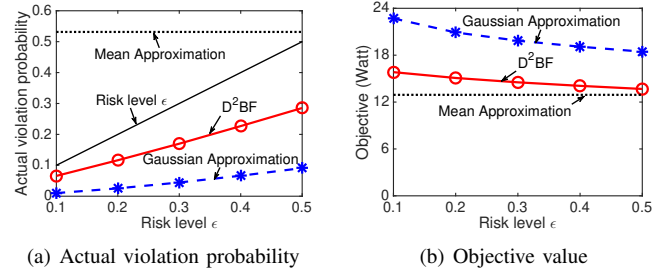


Fig. 6. Performance of D²BF as a function of risk level ϵ .

D²BF, which demonstrates its conservativeness. Finally, we see that the violation probabilities from Mean Approximation are over 50% at all points, which exceeds the target risk level ϵ . Its constant violation probability is because Mean Approximation does not consider any threshold violations.

Fig. 6(b) shows the objective values (BS's power consumption on all RBGs) w.r.t. ϵ . In this figure, we find that the objective values of both D²BF and Gaussian Approximation decrease w.r.t. ϵ . This is because a larger ϵ leads to a higher tolerance of SINR threshold violations and hence the BS can save more power while meeting these loose SINR requirements. Further, D²BF performs better than Gaussian Approximation with 27% power saving on average, which demonstrates that D²BF offers a tighter approximation than Gaussian Approximation. Finally, compared to the lower bound from Mean Approximation (which cannot meet target violation probabilities), the largest gap from D²BF is no more than 23% (for $\epsilon = 0.1$), which demonstrate the closeness of D²BF to the (unknown) optimal objective values.

VII. CONCLUSIONS

We presented a data-driven approach to design downlink MU-MIMO beamforming using a limited number of CSI data samples. Our goal is to minimize the BS's power consumption on all RBGs while meeting probabilistic data rate guarantees for all UEs. We formulated a CCP and showed its decomposition into independent subproblems, each corresponding to MU-MIMO beamforming on an RBG. We introduced ∞ -Wasserstein ambiguity set to incorporate the available CSI data samples and substitute the unknown CSI distribution. For each subproblem, we showed that it can be converted into a bilevel formulation consisting of a simple upper-level problem and a non-convex lower-level problem. Through convex approximation for the lower-level problem and binary search for the objective function, we showed that D²BF can derive a beamforming solution with polynomial time complexity. Simulations confirmed that D²BF can achieve better performance compared to state-of-the-art approach while meeting probabilistic data rate requirements of the UEs.

ACKNOWLEDGMENT

This research was supported in part by NSF under Grants CNS-1617634 and CMMI-2046426, Virginia Commonwealth Cyber Initiative (CCI), and Virginia Tech Institute for Critical Technology and Applied Science (ICTAS).

REFERENCES

- [1] W. Hong, K.-H. Baek, Y. Lee, Y. Kim, and S.-T. Ko, "Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 63–69, Sept. 2014.
- [2] M. Codreanu, A. Tolli, M. Juntti, and M. Latva-Aho, "Joint design of Tx-Rx beamformers in MIMO downlink channel," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4639–4655, Sept. 2007.
- [3] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, Jan. 2004.
- [4] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3060–3097, Fourth Quarter 2018.
- [5] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 172–179, Feb. 2013.
- [6] S. Serbetli and A. Yener, "Transceiver optimization for multiuser MIMO systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 1, pp. 214–226, Jan. 2004.
- [7] Y. Chen, Y. Huang, C. Li, Y. T. Hou, and W. Lou, "Turbo-HB: A novel design and implementation to achieve ultra-fast hybrid beamforming," in *Proc. IEEE INFOCOM 2020*, pp. 1489–1498, Virtual Conference, July 2020.
- [8] Y. Wu, R. H. Louie, and M. R. McKay, "Analysis and design of wireless ad hoc networks with channel estimation errors," *IEEE Transactions on Signal Processing*, vol. 61, no. 6, pp. 1447–1459, Mar. 2013.
- [9] Y. Liu, Z. Tan, H. Hu, L. J. Cimini, and G. Y. Li, "Channel estimation for OFDM," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1891–1908, Fourth Quarter 2014.
- [10] X. Rao and V. K. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3261–3271, June 2014.
- [11] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2868–2882, May 2015.
- [12] X. Jiang and F. Kaltenberger, "Channel reciprocity calibration in TDD hybrid beamforming massive MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 422–431, June 2018.
- [13] X. Jiang, A. Decurninge, K. Gopala, F. Kaltenberger, M. Guillaud, D. Slock, and L. Deneire, "A framework for over-the-air reciprocity calibration for TDD massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5975–5990, Sept. 2018.
- [14] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3734–3749, Sept. 2017.
- [15] Y. Xu, X. Zhao, and Y.-C. Liang, "Robust power control and beamforming in cognitive radio networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 1834–1857, Fourth Quarter 2015.
- [16] K.-Y. Wang, A. M.-C. So, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5690–5705, Nov. 2014.
- [17] M. B. Shenouda, T. N. Davidson, and L. Lampe, "Outage-based design of robust Tomlinson–Harashima transceivers for the MISO downlink with QoS requirements," *Elsevier Signal Processing*, vol. 93, no. 12, pp. 3341–3352, Dec. 2013.
- [18] Y. Shi, J. Zhang, and K. B. Letaief, "Optimal stochastic coordinated beamforming for wireless cooperative networks with CSI uncertainty," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 960–973, Feb. 2014.
- [19] C. Pan, H. Ren, M. ElKashlan, A. Nallanathan, and L. Hanzo, "Robust beamforming design for ultra-dense user-centric C-RAN in the face of realistic pilot contamination and limited feedback," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 780–795, Feb. 2019.
- [20] E. Song, Q. Shi, M. Sanjabi, R.-Y. Sun, and Z.-Q. Luo, "Robust SINR-constrained MISO downlink beamforming: When is semidefinite programming relaxation tight?" *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–11, Aug. 2012.
- [21] M. B. Shenouda and T. N. Davidson, "Nonlinear and linear broadcasting with QoS requirements: Tractable approaches for bounded channel uncertainties," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 1936–1947, May 2009.
- [22] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming*, vol. 158, no. 1, pp. 291–327, July 2016.
- [23] 3GPP, *TR 38.214: 5G; NR; Physical layer procedures for data*, Dec. 2021, version 16.8.0. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216> (Last Accessed: Jan. 2022).
- [24] S. Sur, I. Pefkianakis, X. Zhang, and K.-H. Kim, "Practical MU-MIMO user selection on 802.11ac commodity networks," in *Proc. ACM MobiCom*, pp. 122–134, New York City, NY, Oct. 2016.
- [25] Z. Shen, R. Chen, J. Andrews, R. Heath, and B. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3658–3663, Sept. 2006.
- [26] 3GPP, *TR 38.211: 5G; NR; Physical channels and modulation*, Dec. 2021, version 16.8.0. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213> (Last Accessed: Jan. 2022).
- [27] A. M. Tulino, G. Caire, S. Shamai, and S. Verdú, "Capacity of channels with frequency-selective and time-selective fading," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1187–1215, 2010.
- [28] W. Yang, G. Durisi, and E. Riegler, "On the capacity of large-MIMO block-fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 117–132, Feb. 2013.
- [29] W. Xie, "On distributionally robust chance constrained programs with Wasserstein distance," *Mathematical Programming*, vol. 186, no. 1, pp. 115–155, Mar. 2021.
- [30] N. Jiang and W. Xie, "ALSO-X and ALSO-X+: Better convex approximations for chance constrained programs," *arXiv preprint*, Oct. 2021, Available: <https://arxiv.org/abs/2012.04763.pdf>.
- [31] T. Champion, L. De Pascale, and P. Juutinen, "The ∞ -Wasserstein distance: Local solutions and existence of optimal transport maps," *SIAM Journal on Mathematical Analysis*, vol. 40, no. 1, pp. 1–20, Mar. 2008.
- [32] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, July 2018.
- [33] G. Zheng, K.-K. Wong, and B. Ottersten, "Robust cognitive beamforming with bounded channel uncertainties," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4871–4881, Dec. 2009.
- [34] X. Sun, N. Yang, S. Yan, Z. Ding, D. W. K. Ng, C. Shen, and Z. Zhong, "Joint beamforming and power allocation in downlink NOMA multiuser MIMO networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5367–5381, Aug. 2018.
- [35] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.
- [36] T.-H. Chang, Z.-Q. Luo, and C.-Y. Chi, "Approximation bounds for semidefinite relaxation of max-min-fair multicast transmit beamforming problem," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3932–3943, Aug. 2008.
- [37] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, Chapter 3, pp. 203–205, 2001.
- [38] S. Ahmed, J. Luedtke, Y. Song, and W. Xie, "Nonanticipative duality, relaxations, and formulations for chance-constrained stochastic programs," *Mathematical Programming*, vol. 162, no. 1–2, pp. 51–81, 2017.
- [39] D. Hwang, D. I. Kim, and T.-J. Lee, "Throughput maximization for multiuser MIMO wireless powered communication networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5743–5748, July 2015.
- [40] 3GPP, *TR 36.931: Radio Frequency (RF) requirements for LTE Pico Node B*, July 2021, version 16.0.0. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2589> (Last Accessed: Jan. 2022).