MU-MIMO Beamforming With Limited Channel Data Samples

Shaoran Li, *Member, IEEE*, Nan Jiang, *Graduate Student Member, IEEE*, Yongce Chen, Weijun Xie, *Member, IEEE*, Wenjing Lou, *Fellow, IEEE*, and Y. Thomas Hou, *Fellow, IEEE*

***Abstract*— Channel State Information (CSI) is a critical piece of information for MU-MIMO beamforming. However, CSI estimation errors are inevitable in practice. The random and uncertain nature of CSI estimation errors poses significant challenges to MU-MIMO beamforming. State-of-the-art works addressing such a CSI uncertainty can be categorized into model-based and data-driven works, both of which have limitations when providing a performance guarantee to the users. In contrast, this paper presents Limited Sample-based Beamforming (LSBF)—a novel approach to MU-MIMO beamforming that only uses a limited number of CSI data samples (without assuming any knowledge of channel distributions). Thanks to the use of CSI data samples, LSBF enjoys flexibility similar to data-driven approaches and can provide a theoretical guarantee to the users—a major strength of model-based approaches. To achieve both, LSBF employs chance-constrained programming (CCP) and utilizes the ∞-Wasserstein ambiguity set to bridge the unknown CSI distribution with limited CSI samples. Through problem decomposition and a novel bilevel formulation for each subproblem based on limited CSI data samples, LSBF solves each subproblem with a binary search and convex approximation. We show that LSBF significantly improves the network performance while providing a probabilistic data rate guarantee to the users.**

***Index Terms*— Channel uncertainty, chance-constrained programming, data samples, 5G, MU-MIMO, beamforming.**

I. Introduction

MULTI-USER MIMO (MU-MIMO) beamforming is a key technology component for 5G/NextG networks, which requires Channel State Information (CSI) between a base station (BS) and its connected user equipment (UE) [2], [3], [4]. Most of the existing works assume CSI can be accurately estimated [4], [5], [6], [7], [8]. However, since CSI is obtained through a channel sounding procedure based on pre-defined signals (e.g., pilots), estimation errors are inevitable, due to noise and finite length of training symbols [9], [10]. The estimated CSI is also susceptible to channel aging, meaning that the estimated CSI used for solution derivation differs from the actual CSI during transmission, which may seriously impact the network performance [11], [12]. Further, such a CSI estimation procedure designed in either Frequency Division Duplex (FDD) systems [13], [14] or Time Division Duplex (TDD) systems [15], [16], [17] inherently introduce errors due to limited feedback or hardware imbalance. Thus, CSI estimation is bound to embed errors [18], and must be carefully addressed when optimizing MU-MIMO beamforming.

State-of-the-art approaches to address CSI estimation errors in MU-MIMO beamforming mainly fall into two categories: *model-based* and *data-driven* (i.e., model-free). Model-based works can provide a performance guarantee to the UEs, which include stochastic optimization and worst-case optimization. In stochastic optimization, CSI errors are assumed to follow some well-known distributions, such as Gaussian [19], [20], [21], [22], or uniform [20]. However, such assumed distributions may be far from those in reality due to the discrepancy between the simplified mathematical functions and the complicated operating environment. Consequently, using such a solution may lead to an overly optimistic or pessimistic performance. In worst-case optimization, CSI errors are assumed to stay within some worst-case bounds, such as norm boundaries [23] or ellipsoid uncertain set [24]. However, it is well known that worst-case optimization is very conservative since it only focuses on extreme (unlikely or never) scenarios.

Under the data-driven approach (i.e., model-free), CSI data samples are directly used to derive an MU-MIMO beamforming solution. Since no model is assumed, a data-driven approach can be applied to a wide range of network settings, with the prevailing examples being learning-based solutions (see, e.g., [25], [26], [27], [28]). In these works, a neural network is trained offline based on a large dataset consisting of past CSI data samples and then is used online to derive a beamforming solution using real-time collected CSI samples. However, none of these works can offer any performance guarantee. Neither it is trivial to collect a large number of CSI data samples and perform offline training. Further, when the deployment environment changes, the neural network must be retrained. Although *transfer learning* or *meta-learning* [29], [30], [31], [32] can adapt to new environments without a re-training process, they still cannot offer any theoretical guarantee.

Manuscript received 29 November 2023; revised 28 April 2024; accepted 20 May 2024. Date of publication 22 July 2024; date of current version 18 October 2024. This research was supported in part by NSF under CNS-2312447 and CMMI-2246414, Office of Naval Research (ONR) Multidisciplinary University Research Initiatives (MURI) grant N00014-19-1-2621, Virginia Commonwealth Cyber Initiative (CCI), and Virginia Tech Institute for Critical Technology and Applied Science (ICTAS). An earlier version of this paper was presented at the Proceedings of the IEEE INFOCOM, Virtual Conference, May 2022 [DOI: 10.1109/INFOCOM48880.2022.9796930]. *(Corresponding author: Y. Thomas Hou.)*

Shaoran Li and Yongce Chen are with NVIDIA Corporation, Santa Clara, CA 95051 USA.

Nan Jiang and Weijun Xie are with Georgia Tech, Atlanta, GA 30332 USA.

Wenjing Lou and Y. Thomas Hou are with Virginia Tech, Blacksburg, VA 24061 USA (e-mail: thou@vt.edu).

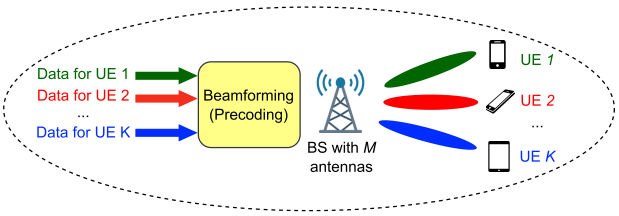
Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2024.3431515.

Digital Object Identifier 10.1109/JSAC.2024.3431515

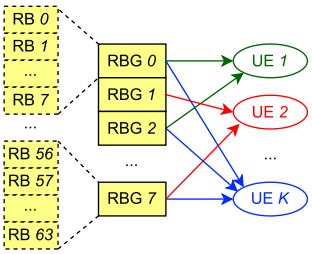
In this paper, we propose a novel approach to design MU-MIMO beamforming called Limited Sample-based Beamforming (LSBF). LSBF is capable of offering a probabilistic guarantee to UE data rates through chance-constrained programming (CCP) and only requires a limited number of CSI data samples. Compared to model-based approaches, LSBF is more adaptive to channel dynamics due to the use of CSI data samples and can provide better performance. Compared to data-driven approaches, LSBF offers a performance guarantee to the UE data rates without training any neural networks or collecting a large dataset. Therefore, LSBF combines the best features of model-based and data-driven works without their pitfalls. Our main contributions are summarized as follows:

- We investigate an MU-MIMO beamforming problem by only using a limited number of CSI data samples. Our objective is to provide a probabilistic guarantee to UE data rates and minimize the BS's power consumption. To the best of our knowledge, this is the first work that offers a probabilistic guarantee to the UE data rates for MU-MIMO beamforming solely based on a limited number of CSI data samples without any knowledge of CSI distributions.
- We decompose the original problem into smaller and independent subproblems. For each subproblem, we propose a novel approach to bridge the true but unknown CSI distribution in the original problem formulation with the limited CSI data samples through the ∞-Wasserstein ambiguity set. We show how to replace the true but unknown distribution in CCP with empirical distribution (based on limited CSI data samples) and additional constraints (based on the properties of ∞-Wasserstein ambiguity set).
- For the new formulation that only involves empirical distribution based on CSI data samples, we propose to break up the complex formulation into a bilevel optimization problem, which has a trivial feasibility check in the upper-level problem and a non-convex lower-level problem. For the lower-level problem, we employ a convex approximation to address its nonlinear objective function and constraints. We show that the entire solution process has a polynomial time complexity.
- Through extensive simulations, we show that LSBF can provide a probabilistic performance guarantee to the UE data rates using limited CSI data samples. In terms of performance, we show that LSBF offers significant power conservation in BS's transmission power compared to state-of-the-art approaches.

We organize the remainder of this paper as follows. In Section II, we describe our system model and state the MU-MIMO beamforming problem. In Section III, we formulate the optimization problem and decompose it into subproblems. In Section IV, we introduce ∞-Wasserstein ambiguity set to reformulate the optimization problem using limited CSI data samples. In Section V, we present the details of LSBF to derive the MU-MIMO beamforming solution. In Section VI, we conduct simulation experiments to evaluate LSBF's performance. Section VII concludes this paper.



(a) Downlink MU-MIMO



(b) RBs to RBGs for scheduling

Fig. 1. An illustration of downlink MU-MIMO beamforming in a 5G cell (a) and RBs to RBG grouping for resource scheduling (b).

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 on the same spectrum, as shown in Fig. 1(a). Without loss of generality, we assume that each UE has one 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 due to its simplicity and effectiveness [33], [34]. In linear precoding, the BS needs to design a unique precoding vector for each UE's data stream, which will subsequently undergo linear multiplication with the downlink symbols before over-the-air transmission.

As defined in 3GPP standards [35], the time domain is divided into Transmission Time Intervals (TTIs), and the frequency domain is divided into sub-carriers. A block of 12 sub-carriers in one TTI is called a Resource Block (RB) and 2$\sim$32 contiguous RBs can be grouped into an RB Group (RBG). The BS uses RBG as the granularity for scheduling and beamforming, meaning that all RBs in an RBG have the same set of serving UEs and precoding vectors. For instance, Fig. 1(b) shows an example of 64 RBs grouped into 8 RBGs (i.e., 8 RBs per RBG). Each RBG can serve multiple UEs and each UE can be served on multiple RBGs.

The BS collects CSI on all RBs through a channel training procedure based on known signals such as pilots. Similar to existing works (see e.g., [4], [5], [19]), we assume the CSI of each UE on each RB is available at the BS. Then the BS will schedule RBGs to the UEs, choose a proper Modulation and Coding Scheme (MCS) for each UE, calculate the downlink precoding vectors, and apply precoding vectors for the upcoming downlink transmission. To keep complexity under control, one often decouples these steps (see, e.g., [36], [37]) into independent problems and solves them in sequence. Following this decoupled approach, we assume that the subset

of UEs on each RBG and each UE's MCS are given *a priori* when we design the precoding vectors.

B. Channel Uncertainty

As discussed in Section I, the obtained CSI is affected by many unknown factors, such as channel estimation errors, limited feedback, and hardware imbalance. In this paper, we address this CSI uncertainty based on a novel approach without assuming knowledge of any distribution information. Specifically, we will *rely on a set of limited CSI data samples collected in recent TTIs to design beamforming in the next TTI.* To show how this can be done, we make the following assumptions in our exposition:

- In the frequency domain, since an RB is typically narrowband (15 kHz to 120 kHz) [38], we assume CSI among the RBs within the same RBG follows the same distribution [39].
- In the time domain, since a TTI is on the order of milliseconds [38], we assume CSI among a small block of contiguous TTIs (i.e., a window) follows the same distribution.

Based on the above two assumptions, the CSI for one UE on all the RBs within the "super" RBG-TTI window block can be different but follows the same distribution. In other words, we only assume the channel to be stationary. Note that this distribution is unknown at the time of deriving an MU-MIMO beamforming solution.

We employ a sliding window mechanism to collect CSI data samples and use them to design precoding vectors for the RBs in the upcoming TTI. Fig. 2 illustrates this idea. Denote S as the number of RBs in an RBG and we have $S = 8$ in Fig. 2. Each window covers $N/S+1$ TTIs, which have a total number of $N+S$ RBs. We will use the N CSI data samples collected in the first N/S TTIs (in green) to design precoding vectors for the S RBs in the upcoming TTI (in red). Then the sliding window will move by one TTI, as shown by "Next Window" in Fig. 2.

In this work, we will show that a small N (i.e., limited CSI data samples) can offer satisfactory performance. So each windows only covers several TTIs and the storage of CSI can be easily done at the BS considering the large volume of memory at the BS. Note that this sliding window is a general form of the widely used "block-fading" model [40], where CSI is assumed to be constant on each block (a group of RBs) but is completely independent on different blocks. The main difference here is that the CSI is a random variable on each RB and we have no knowledge of its distribution.

C. Problem Statement

We assume each UE in the cell has a data rate requirement. It's well-known that UE data rates depend on their Signal-to-Interference-and-Noise Ratios (SINRs), whose calculations are based on the uncertain CSI from the BS to the UEs. Consequently, the SINRs are also random variables. Therefore, it is reasonable to pursue a probabilistic guarantee that each UE's data rate requirement is satisfied with at least a target probability (e.g., over 90%) over all TTIs.

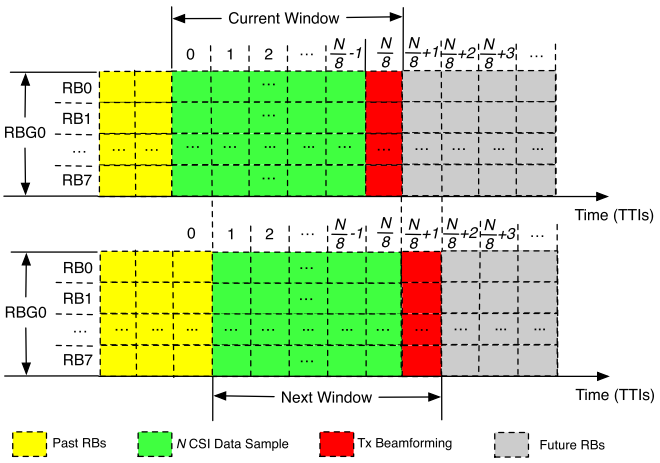


Fig. 2. An illustration of our proposed RBG-TTI window where N CSI data samples (in green) will be used for beamforming (in red).

In this paper, we are interested in designing MU-MIMO beamforming solution to provide a probabilistic guarantee to the UE data rates and minimize the BS's power consumption. We choose power minimization as our objective due to its importance in building a green and sustainable wireless network. From the perspective of network operators, once the pre-defined data rates for UEs are achieved, there is limited incentive to further increase data rates. Instead, the focus shifts towards minimizing radio resources.

There are mainly two challenges in our problem. First, since we only have a limited number of CSI data samples and a lack of CSI distribution knowledge, it is unclear and very challenging how such a guarantee can be achieved. Further, designing precoding vectors requires matrix operations of complex matrices, and hence our problem is mathematically non-trivial.

III. Mathematical Formulation and Analysis

In this section, we formulate our optimization problem and show its decomposition into smaller and independent subproblems to which we will develop solutions later. Table I lists notations used in this paper.

A. Problem Formulation

Referring to Fig. 1(a), denote M as the number of antennas at the BS and $\mathcal{K} = \{1, 2, 3, \cdots, K\}$ as a set of K UEs served by the BS in a 5G cell. Referring to Fig. 1(b), denote $\mathcal{G} = \{1, 2, \cdots, G\}$ 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 (constraints) for feasible precoding vectors: (i) not to exceed the maximum power budget at the BS on all RBGs[1]; and (ii) provide a probabilistic guarantee

[1]In this work, we do not include per-antenna power constraints, as they typically impose less stringent restrictions compared to the total power budget of the BS. This allows us to focus on the probabilistic guarantee to the UE data rates.

TABLE I
NOTATIONS

Symbol	Definition
B	Total transmission bandwidth of one RBG
$\mathbb{C}^{M\times 1}$	The set of all complex $M\times 1$ column vectors
G	Total number of RBGs in set $\mathcal{G}$
$\mathcal{G}$	The set of RBGs, i.e., $\mathcal{G}=\{1,2,\cdots,G\}$
$\mathbf{h}_i^g$	CSI from BS to UE i on RBG g, an $M\times 1$ complex column vector
$\hat{\mathbf{h}}_i^g(n)$	The n-th data sample of $\mathbf{h}_i^g$
K	Total number of UEs in a 5G cell
$\mathcal{K}$	The set of UEs, i.e., $\mathcal{K}=\{1,2,3,\cdots,K\}$
$\mathcal{K}^g$	The subset of UEs from $\mathcal{K}$ scheduled on RBG g
$\lvert\mathcal{K}^g\rvert$	Number of UEs scheduled on RBG g
L_i	Number of RBGs allocated to UE i
M	Number of antennas at the BS
N	Number of data samples for each uncertain CSI $\mathbf{h}_i^g$
$\mathcal{N}$	A set of integers defined as $\mathcal{N}=\{1,2,3,\cdots,N\}$
$P^{\max}$	The maximum power budget of the BS over all RBGs
$\mathbb{P}_{\mathbf{h}_i^g}$	Actual but unknown distribution of CSI $\mathbf{h}_i^g$
$\mathbb{P}_{\hat{\mathbf{h}}_i^g}$	Empirical distribution of $\hat{\mathbf{h}}_i^g$ from N samples of CSI $\mathbf{h}_i^g$
r_i^{req}	Data rate requirement of UE i on all RBGs
S	Number of RBs in an RBG
$\mathcal{P}_{\mathrm{d}_\infty}(\theta_i^g)$	∞-Wasserstein ambiguity set with radius θ_i^g
$\mathbf{w}_i^g$	Beamforming vector for UE i on RBG g, an $M\times 1$ complex column vector, i.e., $\mathbf{w}_i^g\in\mathbb{C}^{M\times 1}$
$\mathbf{W}_i^g$	An $M\times M$ symmetric and positive semidefinite matrix defined as $\mathbf{W}_i^g=\mathbf{w}_i^g(\mathbf{w}_i^g)^H$
ζ_i^g	Actual SINR at UE i on RBG g
ζ_i^{req}	Required SINR threshold at UE i
ϵ_i	Risk level for UE i
θ_i^g	Upper bound of ∞-Wasserstein distance in $\mathcal{P}_{\mathrm{d}_\infty}(\theta_i^g)$
σ_i^2	Power of the thermal noise at UE i (same on all RBGs)

to the UE data rates. To formulate (i), denote $P^{\max}$ as the maximum power budget at the BS for all RBGs. We have

$$\sum_{g\in\mathcal{G}}\sum_{i\in\mathcal{K}^g}||\mathbf{w}_i^g||_2^2\le P^{\max}\,, \tag{1}$$

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

As for (ii), denote r_i^{req} as the data rate requirement of UE i. Per 5G standards [35], a UE must use the same MCS across all its allocated RBGs in a TTI. This means that each RBG that transmits to UE i should contribute the same data rate. For UE i scheduled on an RBG, supporting a data rate r_i^{req} is equivalent to maintaining an SINR threshold (denoted as ζ_i^{req}, same for all its allocated RBGs), which can be calculated based on the bandwidth of an RBG and the Shannon Theorem. Denote L_i as the number of RBGs assigned to UE i in a TTI. To support r_i^{req}, the minimum SINR threshold of UE i must satisfy:

$$r_i^{\text{req}}=L_i\cdot B\cdot\log_2(1+\zeta_i^{\text{req}})\qquad(i\in\mathcal{K})\,, \tag{2}$$

where B is the transmission bandwidth of one RBG. Based on (2), we have

$$\zeta_i^{\text{req}}=2^{\frac{r_i^{\text{req}}}{L_i B}}-1\qquad(i\in\mathcal{K})\,. \tag{3}$$

Denote ζ_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 CSI from the BS to UE i on RBG g. As discussed in Section II, $\mathbf{h}_i^g$ is a random variable with unknown distribution due to channel uncertainty. Then we have

$$\zeta_i^g=\frac{|(\mathbf{w}_i^g)^H\mathbf{h}_i^g|^2}{\sum_{j\in\mathcal{K}^g}^{j\neq i}|(\mathbf{w}_j^g)^H\mathbf{h}_i^g|^2+\sigma_i^2}\qquad(i\in\mathcal{K}^g,g\in\mathcal{G})\,, \tag{4}$$

where $(\cdot)^H$ denotes conjugate transpose. σ_i^2 is the thermal noise power at UE i.

In constraints (4), $\mathbf{w}_i^g$ is a deterministic decision variable, σ_i^2 is a deterministic parameter, and $\mathbf{h}_i^g$ is a random variable. Therefore, ζ_i^g is also a random variable. As discussed in Section II, we aim to provide a probabilistic guarantee to UE data rates (or equivalent SINR thresholds), which can be written as chance constraints:

$$\mathbb{P}\{\zeta_i^g\ge\zeta_i^{\text{req}}\}\ge 1-\epsilon_i\qquad(i\in\mathcal{K}^g,g\in\mathcal{G})\,, \tag{5}$$

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 (5) mean that the actual SINR ζ_i^g on RBG g should be greater than or equal to the required SINR threshold ζ_i^{req} with a probability at least $1-\epsilon_i$.

Substituting (4) into (5), we have

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

which can be rewritten as

$$\mathbb{P}\left\{\frac{|(\mathbf{w}_i^g)^H\mathbf{h}_i^g|^2}{\zeta_i^{\text{req}}}\ge\sum_{j\in\mathcal{K}^g}^{j\neq i}|(\mathbf{w}_j^g)^H\mathbf{h}_i^g|^2+\sigma_i^2\right\}\ge 1-\epsilon_i \qquad(i\in\mathcal{K}^g,g\in\mathcal{G})\,. \tag{6}$$

For the random CSI $\mathbf{h}_i^g$, the BS only has limited data samples without knowledge of its distribution. Recall in Fig. 2, we have N CSI data samples per $\mathbf{h}_i^g$ at the BS. Denote $\mathbb{P}_{\mathbf{h}_i^g}$ as the probability density function (PDF) of the unknown distribution of $\mathbf{h}_i^g$, i.e., $\mathbf{h}_i^g\sim\mathbb{P}_{\mathbf{h}_i^g}$. Then we have the N data samples of $\mathbf{h}_i^g$ drawn from the unknown distribution $\mathbb{P}_{\mathbf{h}_i^g}$. Based on the above discussion, our MU-MIMO beamforming problem can be stated as follows:

$$\begin{aligned}\text{(P1)}\ \min_{\mathbf{w}_i^g}\ &\sum_{g\in\mathcal{G}}\sum_{i\in\mathcal{K}^g}||\mathbf{w}_i^g||_2^2\\ \text{s.t.}\ &\text{BS power budget (1)}\,,\\ &\text{Probabilistic guarantee to UEs' SINRs (6)}\,,\\ &\text{Unknown distribution: }\mathbf{h}_i^g\sim\mathbb{P}_{\mathbf{h}_i^g}\,,N\ \mathbf{h}_i^g\text{ samples}\,,\\ &\mathbf{w}_i^g\in\mathbb{C}^{M\times 1}\,,\end{aligned}$$

where $\mathbb{C}^{M\times 1}$ is the set of all complex $M\times 1$ column vectors.

There are two difficulties in P1. First, from the formulation, it appears that the beamforming vectors on all RBGs are coupled together due to the objective function and constraint (1). Second, it is unclear how to calculate the probabilistic guarantee to UEs' SINRs in constraints (6), especially with only limited CSI data samples from unknown distribution $\mathbb{P}_{\mathbf{h}_i^g}$. In the rest of this section, we address the first issue and leave the second issue to Section IV.

B. Problem Decomposition

In this section, we decompose P1 into G subproblems, where each subproblem corresponds to MU-MIMO beamforming on an RBG and can be solved independently. Note that the RBGs in the objective function of P1 can be easily decoupled and the only constraint that couples beamforming on RBGs is constraint (1), which ties all the BS 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

$$\begin{aligned}(\text{P2})\ \min_{\mathbf{w}_i^g}\ & \sum_{g\in\mathcal{G}}\sum_{i\in\mathcal{K}^g}||\mathbf{w}_i^g||_2^2\\ \text{s.t.}\ & \text{Probabilistic guarantee to UEs' SINRs (6)}\ ,\\ & \text{Unknown distribution: } \mathbf{h}_i^g\sim\mathbb{P}_{\mathbf{h}_i^g}\ , N\ \mathbf{h}_i^g \text{ samples}\ ,\\ & \mathbf{w}_i^g\in\mathbb{C}^{M\times 1}\ .\end{aligned}$$

Comparing the relationship between 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 given in Appendix A. Based on Lemma 1, 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 declaring that P1 is infeasible.

We now show that the objective function and constraints (6) in P2 can be decomposed among the RBGs. 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. its scheduled subset of 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

$$\sum_{g\in\mathcal{G}}\left(\min_{\mathbf{w}_i^g}\sum_{i\in\mathcal{K}^g}||\mathbf{w}_i^g||_2^2\right)\ . \tag{7}$$

This means that we can decompose this objective function into G terms with the g-th term corresponding to the transmission power on RBG g.

Let us define an $M\times M$ 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 as:

$$\mathbf{W}_i^g\succeq\mathbf{0},\ \text{Rank}(\mathbf{W}_i^g)=1\quad(i\in\mathcal{K}^g,g\in\mathcal{G})\ , \tag{8}$$

where $\succeq$ represents positive semidefinite. It is a common technique in beamforming research (see, e.g., [19], [23]) to use the matrix $\mathbf{W}_i^g$ instead of vector $\mathbf{w}_i^g$, which enhances problem solvability, allowing commercial solvers like MOSEK to be employed for deriving the final solution. Using $\mathbf{W}_i^g$, the objective function of P2 can be rewritten as

$$\sum_{g\in\mathcal{G}}\left(\min_{\mathbf{W}_i^g}\sum_{i\in\mathcal{K}^g}\text{Tr}(\mathbf{W}_i^g)\right)\ , \tag{9}$$

where $\text{Tr}(\cdot)$ is the trace of a matrix.

Further, we divide constraints (6) into G groups where the g-th group is the probabilistic SINR guarantee for the UEs on RBG g and is independent of other $G-1$ groups, i.e.,

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

which can be written as

$$\mathbb{P}\left\{\frac{(\mathbf{h}_i^g)^H\mathbf{W}_i^g\mathbf{h}_i^g}{\zeta_i^{\text{req}}}\geq\sum_{j\in\mathcal{K}^g}^{j\neq i}(\mathbf{h}_i^g)^H\mathbf{W}_j\mathbf{h}_i^g+\sigma_i^2\right\}\geq 1-\epsilon_i\quad(i\in\mathcal{K}^g)\ . \tag{10}$$

By decoupling the objective function and constraints (6) for the G RBGs, the subproblem for RBG g is given as

$$\begin{aligned}(\text{P3})\ \min_{\mathbf{W}_i^g}\ & \sum_{i\in\mathcal{K}^g}\text{Tr}(\mathbf{W}_i^g)\\ \text{s.t.}\ & \text{Probabilistic SINR guarantee for } \mathcal{K}^g\ (10)\ ,\\ & \text{Unknown distribution } \mathbf{h}_i^g\sim\mathbb{P}_{\mathbf{h}_i^g}\ , N\ \mathbf{h}_i^g \text{ samples},\\ & \mathbf{W}_i^g\succeq\mathbf{0},\ \text{Rank}(\mathbf{W}_i^g)=1\quad(i\in\mathcal{K}^g)\ .\end{aligned}$$

So we have successfully decomposed P2 into G independent 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 one RBG (i.e., an instance of P3) to design our solution. For ease of exposition, we will drop the superscript g when there is no confusion.

IV. Bridging Data Samples and Distributions

In this section, we show the relationship between data samples and distributions, which will be our novelty to address channel uncertainty. As discussed in Section II, we only have a limited number of samples of $\mathbf{h}_i$ at the BS to design precoding vector $\mathbf{w}_i$. Denote the N available data samples of $\mathbf{h}_i$ as $\hat{\mathbf{h}}_i(n)$, $n\in\mathcal{N}$, where $\mathcal{N}=\{1,2,3,\cdots,N\}$ and each $\hat{\mathbf{h}}_i(n)$ is an $M\times 1$ complex column vector drawn from $\mathbb{P}_{\mathbf{h}_i}$ (the true but unknown distribution of $\mathbf{h}_i$).

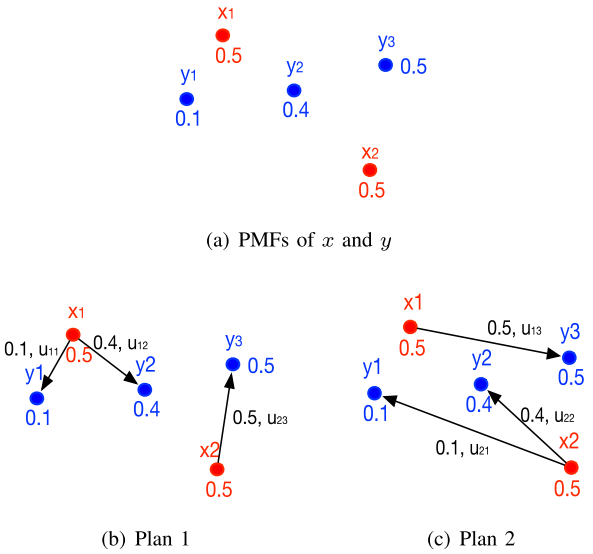
Based on these N data samples, we can construct an empirical distribution for $\mathbf{h}_i$. Denote $\mathbb{P}_{\hat{\mathbf{h}}_i}$ as the probability mass function (PMF) based on the N data samples of $\mathbf{h}_i$ (i.e., $\hat{\mathbf{h}}_i(1),\hat{\mathbf{h}}_i(2),\cdots,\hat{\mathbf{h}}_i(N)$),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

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

Then we have "$\hat{\mathbf{h}}_i\sim\mathbb{P}_{\hat{\mathbf{h}}_i}$". Clearly, $\mathbb{P}_{\hat{\mathbf{h}}_i}$ and $\mathbb{P}_{\mathbf{h}_i}$ are closely related but different. To quantify how "close" they are, we employ the ∞-Wasserstein distance [41], [42].

A. ∞-Wasserstein Distance

The origin of Wasserstein distance traces back to the optimal transport problem that finds the least effort to transfer a given set of mines to a given set of factories [43]. 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 reformulations for our problem.



(a) PMFs of x and y

(b) Plan 1

(c) Plan 2

Fig. 3. Two moving plans from distribution x to distribution y.

TABLE II
JOINT DISTRIBUTION FOR PLAN 1

Probability	x_1	x_2
y_1	0.1	0.0
y_2	0.4	0.0
y_3	0.0	0.5

TABLE III
JOINT DISTRIBUTION FOR PLAN 2

Probability	x_1	x_2
y_1	0.0	0.1
y_2	0.0	0.4
y_3	0.5	0.0

Suppose we have two random variables $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ with their marginal PDFs (for continuous random variables) or PMFs (for discrete random variables) $\mathbb{P}_{\boldsymbol{\xi}_1}$ and $\mathbb{P}_{\boldsymbol{\xi}_2}$, respectively. To change $\mathbb{P}_{\boldsymbol{\xi}_1}$ to $\mathbb{P}_{\boldsymbol{\xi}_2}$, 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, e.g., distance u_{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 the p-th power of distance u_{11}, i.e., $(0.1 \cdot u_{11}^p)$, 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 effort. For $p \geq 1$, Plan 1 always requires less effort than Plan 2. In fact, we can show that Plan 1 is the optimal moving plan with the least effort. p-Wasserstein distance is defined as the p-th root of the minimum required effort among all possible moving plans. So Plan 1 will be used to calculate the p-Wasserstein distance between x and y.

Mathematically, a moving plan, as illustrated in Table II and Table III (for the two respective plans) can be mapped to a joint distribution of x and y. Then Wasserstein distance corresponds to a specific (optimal) joint distribution. Denote $\mathbb{Q}_1$ as the joint distribution of x and y in Plan 1. Under the definition of p-Wasserstein distance, the effort of moving from distribution x to distribution y under Plan 1 is calculated as:

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

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

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

where the first equality holds because the sum of three items only depends on the dominant term as $p \to \infty$; the last equality holds as we have $u_{23} \geq u_{12} \geq u_{11}$ in Fig. 3(b). The physical meaning of ∞-Wasserstein is the maximum moving distance over all steps in the optimal plan (joint distribution). This interpretation makes ∞-Wasserstein highly tractable. ■

We now present the formal definition of ∞-Wasserstein distance as follows.

Definition 1: The ∞-Wasserstein distance of $\mathbb{P}_{\boldsymbol{\xi}_1}$ and $\mathbb{P}_{\boldsymbol{\xi}_2}$ is defined as

$$\mathrm{d}_\infty(\mathbb{P}_{\boldsymbol{\xi}_1}, \mathbb{P}_{\boldsymbol{\xi}_2}) = \inf_{\mathbb{Q}\in\mathcal{Q}} \left\{\sup_{\mathbb{Q}} ||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2||\right\} \ , \tag{13}$$

where $||\cdot||$ is any norm, "$\sup(\cdot)$" stands for the supremum,[2] *$\mathbb{Q}$ stands for a joint distribution of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, $\mathcal{Q}$ stands for the set of all possible $\mathbb{Q}$'s respectively.*

In (13), "$\sup_{\mathbb{Q}}\ ||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2||$" represents the effort under a specific moving plan $\mathbb{Q}$ (i.e., Plan 1 in Fig. 3(b) or Plan 2 in Fig. 3(c)). The "inf over $\mathbb{Q} \in \mathcal{Q}$" represents finding the optimal moving plan with the minimum moving effort. Though the definition (13) holds for any norm $||\cdot||$, it is common to choose L_2-norm due to its attractive computational properties, as in Example 1. Note that $\mathrm{d}_\infty(\mathbb{P}_{\boldsymbol{\xi}_1}, \mathbb{P}_{\boldsymbol{\xi}_2}) = 0$ holds if and only if $P_{\boldsymbol{\xi}_1} = \mathbb{P}_{\boldsymbol{\xi}_2}$ almost surely.[3] Otherwise, we have $\mathrm{d}_\infty(\mathbb{P}_{\boldsymbol{\xi}_1}, \mathbb{P}_{\boldsymbol{\xi}_2}) > 0$.

B. ∞-Wasserstein Ambiguity Set

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

$$\mathcal{P}_{\mathrm{d}_\infty}(\theta_i) = \left\{\mathbb{P}\colon \mathrm{d}_\infty(\mathbb{P}, \mathbb{P}_{\hat{\mathbf{h}}_i}) \leq \theta_i,\ \mathbb{P} \in \mathcal{P}\right\} \quad (i \in \mathcal{K}) \ , \tag{14}$$

[2]The ∞-Wasserstein distance is also defined in terms of "essential supremum" to avoid some extreme distributions [41], [43]. Since such extreme distributions are not encountered in our problem, we use the simplified supermum instead.

[3]An event is said to happen "almost surely" if it happens with probability 1 (or Lebesgue measure 1).

where $\mathcal{P}$ stands for all possible distributions for an $M \times 1$ random vector. $\mathcal{P}_{\mathrm{d}_\infty}(\theta_i)$ is called ∞-*Wasserstein ambiguity set* [44] and can be viewed as a ball of distributions centered at $\mathbb{P}_{\hat{\mathbf{h}}_i}$ with a radius θ_i. In other words, $\mathcal{P}_{\mathrm{d}_\infty}(\theta_i)$ contains the distributions that are "close" to the emprical distribution $\mathbb{P}_{\hat{\mathbf{h}}_i}$.

Suppose that we choose θ_i's properly such that the true (but unknown) distribution $\mathbb{P}_{\mathbf{h}_i}$ falls in the ball, i.e.,

$$\mathbb{P}_{\mathbf{h}_i} \in \mathcal{P}_{\mathrm{d}_\infty}(\theta_i) \quad (i \in \mathcal{K}) \,. \tag{15}$$

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. We will present a simple method to choose θ_i in Section V-D. For the purpose of designing our beamforming solution, we can consider θ_i's as given constants.

We now show how to reformulate P3 based on $\mathcal{P}_{\mathrm{d}_\infty}(\theta_i)$. For ease of exposition, let us rewrite constraints (10) as

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

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

$$\begin{aligned} f(\mathbf{W}_i, \mathbf{h}_i) &= \sum_{j \in \mathcal{K}}^{j \neq i} \mathbf{h}_i^H \mathbf{W}_j \mathbf{h}_i + \sigma_i^2 - \frac{\mathbf{h}_i^H \mathbf{W}_i \mathbf{h}_i}{\zeta_i^{\mathrm{req}}} \\ &= \mathbf{h}_i^H \cdot \left(\sum_{j \in \mathcal{K}}^{j \neq i} \mathbf{W}_j - \frac{\mathbf{W}_i}{\zeta_i^{\mathrm{req}}} \right) \cdot \mathbf{h}_i + \sigma_i^2 \end{aligned} \tag{17}$$

Note that we have dropped superscript g for simplicity when there is no confusion.

Combining constraints (15) and (16), we have

$$\inf_{\mathbb{P}_{\mathbf{h}_i} \in \mathcal{P}_{\mathrm{d}_\infty}(\theta_i)} \mathbb{P}\{f(\mathbf{W}_i, \mathbf{h}_i) \le 0\} \ge 1 - \epsilon_i \quad (i \in \mathcal{K}) \,. \tag{18}$$

The "inf" in constraints (18) means that for any distribution $\mathbb{P}_{\mathbf{h}_i}$ from $\mathcal{P}_{\mathrm{d}_\infty}(\theta_i)$, the probabilistic SINR threshold guarantee for the UEs should be valid.

Based on the definition of ∞-Wasserstein ambiguity set, constraints (18) can be equivalently reformulated into [42]

$$\mathbb{P}\{\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i) \le 0\} \ge 1 - \epsilon_i \quad (i \in \mathcal{K}) \,, \tag{19}$$

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

$$\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i) = \max_{\mathbf{c}_i} \{ f(\mathbf{W}_i, \mathbf{c}_i) \colon ||\mathbf{c}_i - \hat{\mathbf{h}}_i||_2 \le \theta_i \} \,. \tag{20}$$

We see that the uncertain CSI $\mathbf{h}_i$ in (18) disappears. Instead, the estimated CSI $\hat{\mathbf{h}}_i$ is included in (19). Here we introduce an auxiliary variable $\mathbf{c}_i \in \mathbb{C}^{M \times 1}$. Note that $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i) \le 0$ in constraints (19) means that given $\hat{\mathbf{h}}_i$, we should have $f(\mathbf{W}_i, \mathbf{c}_i) \le 0$ holds for any $\mathbf{c}_i$ that satisfies $||\mathbf{c}_i - \hat{\mathbf{h}}_i||_2 \le \theta_i$.

It's worth noting that our approach is largely different from the worst-case optimization. First, we use CSI data samples, the empirical distribution, and an upper bound of its distance to the true but unknown distribution. In contrast, worst-case optimization requires a conservative boundary of uncertain parameters (min and max CSI estimation errors). Second, in our model, the uncertain CSI $\mathbf{h}_i$ can be within θ_i distance from "any" CSI data sample $\hat{\mathbf{h}}_i(n)$ out of the N available CSI data samples. But in the worst-case optimization, the uncertain CSI $\mathbf{h}_i$ must be within predefined boundaries. Last but not least, our proposed approach offers controllable occasional SINR threshold violations through (19) and better performance. In comparison, worst-case optimization does not allow any threshold violation and is known to be overly conservative.

Recall that $\hat{\mathbf{h}}_i$'s closed-form distribution $\mathbb{P}_{\hat{\mathbf{h}}_i}$ is given in (11) based on N CSI data samples. Thus, we plug in this distribution knowledge (11) into (19) and obtain:

$$\sum_{n \in \mathcal{N}} \mathbb{I}\{\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)) \le 0\} \ge N \cdot (1 - \epsilon_i) \quad (i \in \mathcal{K}) \,, \tag{21}$$

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

Based on (21), we can rewrite P3 as

$$\begin{aligned} \text{(P4)} \quad \min_{\mathbf{W}_i} \ & \sum_{i \in \mathcal{K}} \mathrm{Tr}(\mathbf{W}_i) \\ \text{s.t.} \ & \text{Probabilistic data rate guarantee (21)} \,. \\ & \mathbf{W}_i \succeq \mathbf{0}, \ \mathrm{Rank}(\mathbf{W}_i) = 1 \quad (i \in \mathcal{K}) \,. \end{aligned}$$

For the rank constraints "$\mathrm{Rank}(\mathbf{W}_i) = 1$" in P4, a widely used technique is to employ semi-definite programming (SDP) relaxation (see, e.g., [45], [46], [47]). In SDP relaxation, we first relax the rank constraints "$\mathrm{Rank}(\mathbf{W}_i) = 1$" 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 $\mathbf{W}_i$ [48]. So the next step is to find a solution for P4 without the rank constraints.

However, after dropping the rank constraints, we still have constraints (21). Even though we have replaced the unknown PDF $\mathbb{P}_{\mathbf{h}_i}$ in (18) with N CSI data samples in (19), it is still unclear how to handle $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i(n))$ defined in (20) and the indicator functions.

V. LSBF—Limited Sample-Based Beamforming Solution

In this section, we present LSBF—A Limited Sample-based BeamForming solution to P4 (and P1). The design of LSBF is based on a convex approximation of P4, which hinges on a bilevel formulation and a novel reformulation technique called ALSO-X+ [42].

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 follows:

$$z_i(\hat{\mathbf{h}}_i(n)) = \begin{cases} 1, & \text{if } \hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)) \le 0 \,, \\ 0, & \text{otherwise} \,. \end{cases} \tag{22}$$

With $z_i(\hat{\mathbf{h}}_i(n))$, we can rewrite (21) as

$$\sum_{n \in \mathcal{N}} z_i(\hat{\mathbf{h}}_i(n)) \ge N \cdot (1 - \epsilon_i) \qquad (i \in \mathcal{K}) \,. \tag{23}$$

To put constraints (22) into closed-form constraints, we introduce an auxiliary variable $s_i\big(\hat{\mathbf{h}}_i(n)\big)$ w.r.t. $\hat{\mathbf{h}}_i(n)$ such that:

$$s_i\big(\hat{\mathbf{h}}_i(n)\big) \geq 0 \qquad (i \in \mathcal{K}, n \in \mathcal{N}) , \quad (24a)$$

$$\hat{f}\big(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)\big) \leq s_i\big(\hat{\mathbf{h}}_i(n)\big) \qquad (i \in \mathcal{K}, n \in \mathcal{N}) , \quad (24b)$$

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

The nonnegative $s\big(\hat{\mathbf{h}}_i(n)\big)$ can be considered as a slack function w.r.t. $z\big(\hat{\mathbf{h}}_i(n)\big)$. It is easy to see that constraints (22) can be replaced by constraints (24).

Using an auxiliary variable t, we can rewrite the objective function of P4 as "$\min\ t$" and add the following constraints:

$$\sum_{i \in \mathcal{K}} \mathrm{Tr}(\mathbf{W}_i) \leq t . \quad (25)$$

Using this new objective function, adding constraints (25), replacing constraints (21) by constraints (23) and (24), and also dropping the rank constraints, we obtain a reformulation of P4 as follows:

$$\text{(P4-R)} \quad \min_{\mathbf{W}_i, z_i\left(\hat{\mathbf{h}}_i(n)\right), s_i\left(\hat{\mathbf{h}}_i(n)\right)} \quad t$$
$$\text{s.t.} \quad \mathbf{W}_i \succeq 0 \ , \ \text{Constraints (23),(24),(25)} .$$

P4-R suggests that we can use a 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.

Based on (23) and (24c), we have

$$\sum_{n \in \mathcal{N}} \mathbb{I}\big\{ s_i\big(\hat{\mathbf{h}}_i(n)\big) = 0 \big\} \geq N \cdot (1 - \epsilon_i) \quad (i \in \mathcal{K}) . \quad (26)$$

Further, since $z_i(\hat{\mathbf{h}}_i(n)) \geq 0$ and $s_i(\hat{\mathbf{h}}_i(n)) \geq 0$, constraints (24c) is equivalent to:

$$\sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \big\{ z_i\big(\hat{\mathbf{h}}_i(n)\big) \cdot s_i\big(\hat{\mathbf{h}}_i(n)\big) \big\} = 0 . \quad (27)$$

To handle the bilinear constraints (24c), we will drop them and use the left-hand side of constraints (27) as the objective function of the lower-level problem, which facilitates our derivation of a convex approximation in Section V-B. By removing constraints (24c), we need to bring constraints (26) to the problem. We now have a bilevel formulation of P4-R as follows:

$$\text{(P5)} \quad \min_{t} \quad t$$
$$\Big(\mathbf{W}_i^*, z_i^*\big(\hat{\mathbf{h}}_i(n)\big), s_i^*\big(\hat{\mathbf{h}}_i(n)\big)\Big) \in \arg \min_{\mathbf{W}_i, z_i\left(\hat{\mathbf{h}}_i(n)\right), s_i\left(\hat{\mathbf{h}}_i(n)\right)} \Bigg\{ \sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \big\{ z_i\big(\hat{\mathbf{h}}_i(n)\big) \cdot s_i\big(\hat{\mathbf{h}}_i(n)\big) \big\} : \mathbf{W}_i \succeq 0 \ ,$$
$$z_i\big(\hat{\mathbf{h}}_i(n)\big) \in \{0, 1\}, \ \text{Constraints (23),(24a),(24b),(25)} \Bigg\} ,$$
$$\text{Constraints (26)} .$$

We offer some insights into deriving this bilevel formulation:

- *Lower-level problem*: We see the lower-level problem preserves most of the constraints of P4-R. The biggest change is that constraints (24c) disappear in P5 while a new term $\sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \big\{ z_i\big(\hat{\mathbf{h}}_i(n)\big) \cdot s_i\big(\hat{\mathbf{h}}_i(n)\big) \big\}$ is used as the objective function of the lower-level problem.
- *Upper-level problem*: We see t is used as the objective function and constraints (26) are added to ensure the feasibility of the final solution. Note that constraints (26) have been relocated to the upper-level problem from the lower-level problem.
- It can be shown that P5 is an equivalent reformulation of P4-R [42]. The proof is based on the fact that an optimal solution to P4-R can be constructed based on an optimal solution to P5 with the same $\mathbf{W}_i^*$ and the objective value. So is the converse. Placing constraints (26) into the upper-level problem leads to a more tractable feasible region for the lower-level problem without introducing any relaxation errors.

The main idea of P5 is that for a given t, we can solve the lower-level problem to obtain an optimal $\mathbf{W}_i^*$, $z_i^*\big(\hat{\mathbf{h}}_i(n)\big)$ and $s_i^*\big(\hat{\mathbf{h}}_i(n)\big)$. If $s_i^*\big(\hat{\mathbf{h}}_i(n)\big)$ can satisfy the $1 - \epsilon_i$ guarantee, i.e., constraints (26), 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, which can be found through a binary search with a few iterations.

Now the question is: For a given t, how to find a solution to P5 in the form of $\mathbf{W}_i$, $z_i\big(\hat{\mathbf{h}}_i(n)\big)$, and $s_i\big(\hat{\mathbf{h}}_i(n)\big)$? Since the upper-level problem of P5 is a simple feasibility check with constraints (26), we only need to focus on the lower-level problem. This lower-level problem is not trivial due to $\hat{f}\big(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)\big)$ in constraints (24b) and its bilinear objective function $\sum_{i \in \mathcal{K}} \sum_{n \in \mathcal{N}} \big\{ z_i\big(\hat{\mathbf{h}}_i(n)\big) \cdot s_i\big(\hat{\mathbf{h}}_i(n)\big) \big\}$.

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 [49] and a novel technique called "ALSO-X+" [42], [50].

1) Reformulation of $\hat{f}\big(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)\big)$: Since $\hat{f}\big(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)\big)$ in constraints (24b) involves a maximization problem (see (20)) which cannot be solved directly, we need to reformulate $\hat{f}\big(\mathbf{W}_i, \hat{\mathbf{h}}_i(n)\big)$.

Denote $\mathbf{e}_i^n \in \mathbb{C}^{M \times 1}$ as the difference between $\mathbf{c}_i$ and data sample $\hat{\mathbf{h}}_i(n)$, i.e.,

$$\mathbf{e}_i^n = \mathbf{c}_i - \hat{\mathbf{h}}_i(n) \quad (i \in \mathcal{K} \ , \ n \in \mathcal{N}) . \quad (28)$$

Based on the definition of $\hat{f}(\mathbf{W}_i, \hat{\mathbf{h}}_i)$ in (20) and the definition of $\mathbf{e}_i$ in (28), constraints (24b) can be rewritten as:

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

which means that

$$f\big(\mathbf{W}_i, \mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big) \le s_i\big(\hat{\mathbf{h}}_i(n)\big) \quad (i \in \mathcal{K}\ ,\ n \in \mathcal{N}) \tag{29}$$

holds for any $\mathbf{e}_i^n$ that satisfies $||\mathbf{e}_i^n||_2 \le \theta_i$. That is, $||\mathbf{e}_i^n||_2 \le \theta_i$ implies that constraints (29) hold. We have:

$$||\mathbf{e}_i^n||_2 \le \theta_i \Longrightarrow \text{Constraints (29)}\ . \tag{30}$$

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

$$\mathbf{Q}_i = \sum_{j\in\mathcal{K}}^{j\neq i} \mathbf{W}_j - \frac{\mathbf{W}_i}{\zeta_i^{\text{req}}}\ , \tag{31a}$$

$$a_i^n = \big(\hat{\mathbf{h}}_i(n)\big)^H \mathbf{Q}_i \hat{\mathbf{h}}_i(n) + \sigma_i^2\ . \tag{31b}$$

Since $\mathbf{W}_i$ is a Hermitian matrix, $\mathbf{Q}_i$ is also a Hermitian matrix, and consequently, a_i^n is a real number.

Based on the definition of $f\big(\mathbf{W}_i, \mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big)$ in (17), we have

$$\begin{aligned}
&f\big(\mathbf{W}_i, \mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big)\\
&= \big(\mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big)^H \left(\sum_{j\in\mathcal{K}}^{j\neq i} \mathbf{W}_j - \frac{\mathbf{W}_i}{\zeta_i^{\text{req}}}\right) \big(\mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big) + \sigma_i^2\\
&= \big(\mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big)^H \mathbf{Q}_i \big(\mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big) + \sigma_i^2\\
&= (\mathbf{e}_i^n)^H \mathbf{Q}_i \mathbf{e}_i^n + (\mathbf{e}_i^n)^H \mathbf{Q}_i \hat{\mathbf{h}}_i(n) + \big(\hat{\mathbf{h}}_i(n)\big)^H \mathbf{Q}_i \mathbf{e}_i^n +\\
&\quad \big(\hat{\mathbf{h}}_i(n)\big)^H \mathbf{Q}_i \hat{\mathbf{h}}_i(n) + \sigma_i^2\\
&= (\mathbf{e}_i^n)^H \mathbf{Q}_i \mathbf{e}_i^n + (\mathbf{e}_i^n)^H \mathbf{Q}_i \hat{\mathbf{h}}_i(n) + \big(\hat{\mathbf{h}}_i(n)\big)^H \mathbf{Q}_i \mathbf{e}_i^n + a_i^n.
\end{aligned}$$

Plugging in the above results for $f\big(\mathbf{W}_i, \mathbf{e}_i^n + \hat{\mathbf{h}}_i(n)\big)$ into constraints (29), we have

$$\begin{aligned}
(\mathbf{e}_i^n)^H \mathbf{Q}_i \mathbf{e}_i^n + (\mathbf{e}_i^n)^H \mathbf{Q}_i \hat{\mathbf{h}}_i(n) + \big(\hat{\mathbf{h}}_i(n)\big)^H \mathbf{Q}_i \mathbf{e}_i^n + a_i^n\\
\le s_i\big(\hat{\mathbf{h}}_i(n)\big) \qquad (i \in \mathcal{K}, n \in \mathcal{N})\ .
\end{aligned} \tag{32}$$

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

$$(\mathbf{e}_i^n)^H \mathbf{I}_M \mathbf{e}_i^n - \theta_i^2 \le 0\ , \tag{33}$$

where $\mathbf{I}_M$ is the M-dimension identity matrix. Thus, by replacing $||\mathbf{e}_i^n||_2 \le \theta_i$ with (33) and replacing constraints (29) with constraints (32), statement (30) can be rewritten as

$$(\mathbf{e}_i^n)^H \mathbf{I}_M \mathbf{e}_i^n - \theta_i^2 \le 0 \Longrightarrow \text{Constraints (32)}. \tag{34}$$

To derive a closed-form for (34), we resort to $\mathcal{S}$-lemma [49]. For the sake of completeness, we reiterate $\mathcal{S}$-lemma as follows.

Lemma 2: ($\mathcal{S}$-Lemma) *Let $\mathbf{V}_1$ and $\mathbf{V}_2$ be $M \times M$ Hermitian matrices. Suppose (i) $\mathbf{x}^H\mathbf{V}_1\mathbf{x} \le 0$ holds for some $\mathbf{x} \in \mathcal{X}$ and $\mathcal{X} \subseteq \mathbb{C}^{M\times 1}$; and (ii) There exists an $\bar{\mathbf{x}}$ such that $\bar{\mathbf{x}}^H\mathbf{V}_1\bar{\mathbf{x}} < 0$. Then $\mathbf{x}^H\mathbf{V}_2\mathbf{x} \le 0$ holds for $\mathbf{x} \in \mathcal{X}$ if and only if there exists a nonnegative number λ such that $\mathbf{V}_2 \preceq \lambda\mathbf{V}_1$.*

In $\mathcal{S}$-Lemma, "$\mathbf{V}_2 \preceq \lambda\mathbf{V}_1$" means that $\lambda\mathbf{V}_1 - \mathbf{V}_2$ is a positive semidefinite matrix. $\mathcal{S}$-Lemma can convert a statement like (34) into a closed-form constraint with an auxiliary variable λ. Then this closed-form constraint can be directly handled by commercial solvers. To apply $\mathcal{S}$-Lemma to statement (34), we first rewrite it as

$$\begin{aligned}
&\begin{bmatrix}\mathbf{e}_i^n\\ 1\end{bmatrix}^H \begin{bmatrix}\mathbf{I}_M & \mathbf{0}\\ \mathbf{0} & -\theta_i^2\end{bmatrix} \begin{bmatrix}\mathbf{e}_i^n\\ 1\end{bmatrix} \le 0 \Longrightarrow\\
&\begin{bmatrix}\mathbf{e}_i^n\\ 1\end{bmatrix}^H \begin{bmatrix}\mathbf{Q}_i & \mathbf{Q}_i\hat{\mathbf{h}}_i(n)\\ \big(\mathbf{Q}_i\hat{\mathbf{h}}_i(n)\big)^H & a_i^n - s_i\big(\hat{\mathbf{h}}_i(n)\big)\end{bmatrix} \begin{bmatrix}\mathbf{e}_i^n\\ 1\end{bmatrix} \le 0,
\end{aligned}$$

which matches the standard form in $\mathcal{S}$-Lemma with

$$\mathbf{V}_1 = \begin{bmatrix}\mathbf{I}_M & \mathbf{0}\\ \mathbf{0} & -\theta_i^2\end{bmatrix},\ \mathbf{x} = \begin{bmatrix}\mathbf{e}_i^n\\ 1\end{bmatrix},\ \bar{\mathbf{x}} = \begin{bmatrix}\mathbf{0}\\ 1\end{bmatrix},$$

$$\mathbf{V}_2 = \begin{bmatrix}\mathbf{Q}_i & \mathbf{Q}_i\hat{\mathbf{h}}_i(n)\\ \big(\mathbf{Q}_i\hat{\mathbf{h}}_i(n)\big)^H & a_i^n - s_i\big(\hat{\mathbf{h}}_i(n)\big)\end{bmatrix}.$$

Therefore, we have (i) $\mathbf{x}^H\mathbf{V}_1\mathbf{x} = (\mathbf{e}_i^n)^H\mathbf{I}_M\mathbf{e}_i^n - \theta_i^2 \le 0$ for some $\mathbf{x} \in \mathcal{X}$ and $\mathcal{X} \subseteq \mathbb{C}^{M\times 1}$; and (ii) $\bar{\mathbf{x}}^H\mathbf{V}_1\bar{\mathbf{x}} = -\theta_i^2 < 0$ hold. Based on $\mathcal{S}$-Lemma, statement (34) holds if and only if

$$\begin{aligned}
&\begin{bmatrix}\mathbf{Q}_i & \mathbf{Q}_i\hat{\mathbf{h}}_i(n)\\ \big(\mathbf{Q}_i\hat{\mathbf{h}}_i(n)\big)^H & a_i^n - s_i\big(\hat{\mathbf{h}}_i(n)\big)\end{bmatrix} \preceq \lambda_i^n \begin{bmatrix}\mathbf{I}_M & \mathbf{0}\\ \mathbf{0} & -\theta_i^2\end{bmatrix}, \lambda_i^n \ge 0\\
&\qquad\qquad (i \in \mathcal{K}, n \in \mathcal{N})\ .
\end{aligned} \tag{35}$$

Replacing statement (34) with constraints (35), the lower-level problem of P5 becomes:

$$\begin{aligned}
\min_{\mathbf{W}_i, z_i\left(\hat{\mathbf{h}}_i(n)\right), s_i\left(\hat{\mathbf{h}}_i(n)\right)} \Bigg\{ &\sum_{i\in\mathcal{K}}\sum_{n\in\mathcal{N}} \big\{z_i\big(\hat{\mathbf{h}}_i(n)\big) \cdot s_i\big(\hat{\mathbf{h}}_i(n)\big)\big\} :\\
&\mathbf{W}_i \succeq 0\ , z_i\big(\hat{\mathbf{h}}_i(n)\big) \in \{0,1\}\ ,\\
&\text{Constraints (23),(24a),(35),(25)} \Bigg\}\ .
\end{aligned}$$

2) Bilinear Objective Function: As for the bilinear objective function $\sum_{i\in\mathcal{K}}\sum_{n\in\mathcal{N}}\big\{z_i\big(\hat{\mathbf{h}}_i(n)\big) \cdot s_i\big(\hat{\mathbf{h}}_i(n)\big)\big\}$, we will employ convex approximation. Our proposed solution is inspired by ALSO-X+ [42], which has been proven to offer better performance compared to existing reformulation techniques for problems such as P5 [42], [50]. Fig. 4 shows the main idea of our convex approximation where we solve $z_i\big(\hat{\mathbf{h}}_i(n)\big)$, $s_i\big(\hat{\mathbf{h}}_i(n)\big)$, and $\mathbf{W}_i$ in sequence.

As shown in Fig. 4, we start the procedure by setting all $z_i\big(\hat{\mathbf{h}}_i(n)\big) = 1$ and solve for $s_i\big(\hat{\mathbf{h}}_i(n)\big)$ and $\mathbf{W}_i$ (Step 1). This step is motivated by (23) and the value of ϵ_i (whose value tends to be small). So we would anticipate the majority of $z_i\big(\hat{\mathbf{h}}_i(n)\big)$ to be 1. Further, we choose to fix $z_i\big(\hat{\mathbf{h}}_i(n)\big)$ first as they only appear in constraints (23) and the bilinear objective function. So their impacts on other constraints are limited.

With $z_i\big(\hat{\mathbf{h}}_i(n)\big) = 1$, constraints (23) hold trivially. Further, the lower-level problem in P5 can be simplified to

$$\begin{aligned}
\min_{\mathbf{W}_i, s_i\left(\hat{\mathbf{h}}_i(n)\right)} \Bigg\{ &\sum_{i\in\mathcal{K}}\sum_{n\in\mathcal{N}} s_i\big(\hat{\mathbf{h}}_i(n)\big) : \mathbf{W}_i \succeq 0\ ,\\
&\text{Constraints (24a),(35),(25)} \Bigg\}\ .
\end{aligned} \tag{36}$$

This problem is convex and we can solve its optimal solution $\mathbf{W}_i^*$ and $s_i^*\big(\hat{\mathbf{h}}_i(n)\big)$. Note that problem (36) is always

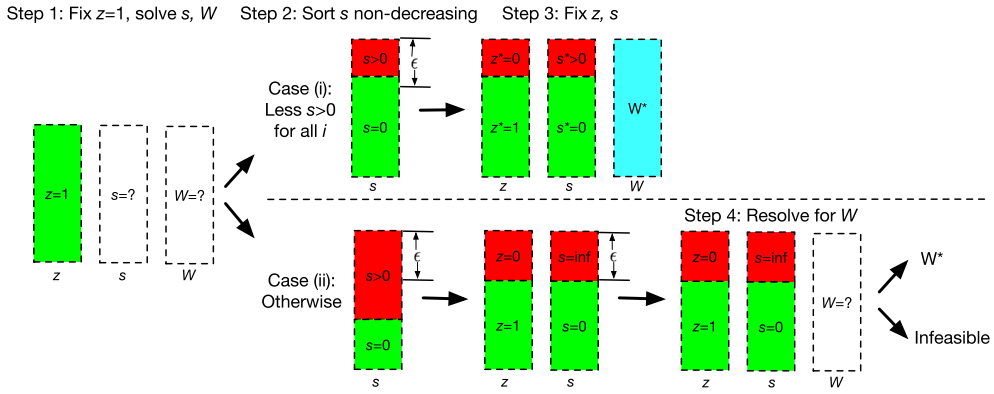


Fig. 4. A solution procedure to the lower-level problem for a given t.

feasible—there is at least a feasible solution with sufficiently large $s_i(\hat{\mathbf{h}}_i(n))$ and $\mathbf{W}_i = \mathbf{0}$.[4]

After we obtain $s_i^*(\hat{\mathbf{h}}_i(n))$, 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 have the following two cases:

- Case (i) The number of positive elements in $\mathcal{S}_i^{\text{sort}}$ divided by N is no greater than ϵ_i for all $i \in \mathcal{K}$.
- Case (ii) Otherwise, i.e., at least for some $i \in \mathcal{K}$, the number of positive elements in $\mathcal{S}_i^{\text{sort}}$ divided by N is greater than ϵ_i.

We discuss how LSBF works for each case as follows.

Case (i): This is the simple case as constraints (26) already hold. To minimize the objective function $\sum_{n\in\mathcal{N}}\{z_i(\hat{\mathbf{h}}_i(n)) \cdot s_i(\hat{\mathbf{h}}_i(n))\}$, we can adjust $z_i(\hat{\mathbf{h}}_i(n))$ from 1 to 0 corresponding to those $s_i^*(\hat{\mathbf{h}}_i(n)) > 0$, i.e.,

$$z_i^*(\hat{\mathbf{h}}_i(n)) = \begin{cases} 0, & \text{if } s_i^*(\hat{\mathbf{h}}_i(n)) > 0\ , \\ 1, & \text{otherwise}\ . \end{cases} \tag{37}$$

This adjustment of $z_i(\hat{\mathbf{h}}_i(n))$ is solely to achieve a minimum objective value $\sum_{n\in\mathcal{N}}\{z_i(\hat{\mathbf{h}}_i(n)) \cdot s_i(\hat{\mathbf{h}}_i(n))\} = 0$. Further, it has no impact on $\mathbf{W}_i$ and $s_i(\hat{\mathbf{h}}_i(n))$ since $z_i(\hat{\mathbf{h}}_i(n))$ only appears in constraints (23) and the objective function. Therefore, this solution is the optimal solution to P5's lower-level problem under the current t.

Case (ii): In this case, after Step 1, constraints (26) do not hold, as there is a fewer number of $s_i^*(\hat{\mathbf{h}}_i(n))$ with $s_i^*(\hat{\mathbf{h}}_i(n)) = 0$. So we propose to first adjust $z_i(\hat{\mathbf{h}}_i(n))$ and $s_i(\hat{\mathbf{h}}_i(n))$ so that both constraints (23) and (26) hold and the objective function $\sum_{n\in\mathcal{N}}\{z_i(\hat{\mathbf{h}}_i(n)) \cdot s_i(\hat{\mathbf{h}}_i(n))\}$ is minimized.

To do this, we propose to fix $z_i(\hat{\mathbf{h}}_i(n))$ and $s_i(\hat{\mathbf{h}}_i(n))$ based on the sorted set $\mathcal{S}_i^{\text{sort}}$. For each UE $i \in \mathcal{K}$, we adjust $s_i(\hat{\mathbf{h}}_i(n))$ with

$$s_i(\hat{\mathbf{h}}_i(n)) = \begin{cases} \infty, & \text{for the first } \lfloor N \cdot \epsilon_i \rfloor \text{ elements in } \mathcal{S}_i^{\text{sort}}\ , \\ 0, & \text{otherwise}\ . \end{cases} \tag{38}$$

[4]This trivial solution is not a feasible solution to P5 due to constraints (26) in its upper-level problem.

Then we adjust $z_i(\hat{\mathbf{h}}_i(n))$ using (37) and the new values for $s_i(\hat{\mathbf{h}}_i(n))$ in (38), as shown in Fig. 4 (Case (ii) Step 3). Note that such setting of $z_i(\hat{\mathbf{h}}_i(n))$ and $s_i(\hat{\mathbf{h}}_i(n))$ will ensure the objective function $\sum_{i\in\mathcal{K}}\sum_{n\in\mathcal{N}}\{z_i(\hat{\mathbf{h}}_i(n)) \cdot s_i(\hat{\mathbf{h}}_i(n))\}$ to be the minimum value 0.

In (38), by setting $s_i(\hat{\mathbf{h}}_i(n)) = \infty$, the corresponding constraints in (35) hold trivially. Since these constraints have no impact when solving $\mathbf{W}_i$, they can be safely removed from the lower-level problem of P5 when solving for $\mathbf{W}_i$ with a lower computation complexity.

Now we have a convex optimization problem (with 0 as the 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 declare that the current t is infeasible.

C. Summary of LSBF

A summary of our proposed solution LSBF is given in Algorithm 1. It 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 the output of LSBF is an MU-MIMO beamforming solution on all RBGs, so we bring back the superscript g. We see Line 5–12 is the binary search to find the minimum t, which runs iteratively until a stop criterion is reached. In each iteration, we set $t = (t^{\text{UB}} + t^{\text{LB}})/2$, apply the procedure in Fig. 4, and updates the upper bound and lower bound for t. Note that in Line 8, we choose $t^{\text{UB}} = \min\{t,\ \sum_{i\in\mathcal{K}^g}\mathbf{W}_i^g\}$ for a faster convergence.

Due to the complicated mathematical structure of the bilevel optimization problem P5 and the convex approximation in Line 6 involving fixing binary variables, we are not able to quantify the approximation errors from LSBF theoretically (as in many existing works using CCP). We will show that LSBF offers much better performance compared to state-of-the-art approaches through simulation experiments in Section VI.

Warm Start To reduce the number of iterations in the binary search, we need a warm-start to initialize t^{LB} and t^{UB} in Line 4. We initialize t^{UB} as the objective value of Gaussian Approximation [21] since we find its objective value is always greater than that from LSBF. In case Gaussian Approximation fails to find a feasible solution, we will simply

Algorithm 1 LSBF

1: **Input**: ζ_i^{req}, $P^{\max}$, $\hat{\mathbf{h}}_i(n)$, θ_i^g, Δ
2: **Output**: $\mathbf{w}_i^g$ or infeasible
3: **parfor** $g \in \mathcal{G}$ (subproblems on RBGs) **do**
4: Set upper bound t^{UB} and lower bound t^{LB}
5: **while** $t^{\text{UB}} - t^{\text{LB}} > \Delta$ **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 obtain $\mathbf{w}_i^g$ from $\mathbf{W}_i^g$
17: **end if**
18: **end parfor**
19: **if** $\mathbf{w}_i^g$'s meet the power budget of the BS (1) **then**
20: Return $\mathbf{w}_i^g$ as the final solution
21: **else**
22: Return infeasible
23: **end if**

set $t^{\text{UB}} = P^{\max}$ due to constraints (1) (BS's power budget on all RBGs).

As for the initial lower bound t^{LB}, we need to design an algorithm for t^{LB} with low complexity. Specifically, we initialize t^{LB} by ignoring the inter-user interference terms when calculating the SINRs. This means that the MU-MIMO system is approximated by $|\mathcal{K}^g|$ MISO transmissions where each UE has a dedicated transmission from the BS. Since we ignore the interference among UEs, the BS will use less power for beamforming and the objective value can serve as the initial t^{LB}. We propose to use Maximum Ratio Transmission (MRT) in the MISO system to derive t^{LB} but it requires perfect CSI $\mathbf{h}_i$, which is not available in our system model. Therefore, we simply use the measured (inaccurate) mean of $\mathbf{h}_i^g$ to derive the MRT solution. Therefore, t^{LB} is initialized as

$$t^{\text{LB}} = \sum_{i\in\mathcal{K}^g} \frac{\zeta_i^{\text{req}}\sigma_i^2}{||\frac{1}{N}\cdot\sum_{n\in\mathcal{N}}\hat{\mathbf{h}}_i^g(n)||_2^2}\,. \tag{39}$$

Recover $\mathbf{w}_i^g$ from $\mathbf{W}_i^g$ Regarding Lines 13–17, although we use "if-else" to consider two cases, in all of our tested cases, we found that we only had the case under "if", i.e., $\text{rank}(\mathbf{W}_i^g) \approx 1$. This is consistent with the observations in [21] and can be explained by the fact that commercial SDP solvers (e.g., MOSEK) typically exploit low-rank structures when solving SDP for matrix solutions such as $\mathbf{W}_i^g$. Therefore, we can just drop the rank constraints $\mathbf{W}_i^g = 1$ in LSBF and use Line 14 directly to recover $\mathbf{w}_i^g$.

Complexity Analysis The binary search for t in Algorithm 1 consists of at most $\lceil \log_2(\frac{t^{\text{UB}}-t^{\text{LB}}}{\Delta})\rceil$ 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 LSBF has polynomial time complexity. While additional optimizations could reduce its actual running time, such enhancements are beyond the scope of this paper.

D. Choosing θ_i^g

In the above discussion, we assume θ_i^g is a given constant. Now we discuss how to choose a suitable θ_i^g for ∞-Wasserstein ambiguity set $\mathcal{P}_{\text{d}_\infty}(\theta_i^g)$ such that $\mathbb{P}_{\mathbf{h}_i^g} \in \mathcal{P}_{\text{d}_\infty}(\theta_i^g)$ holds for $i \in \mathcal{K}^g,\ g \in \mathcal{G}$. If θ_i^g is chosen too small (overly optimistic), then the true (but unknown) distribution $\mathbb{P}_{\mathbf{h}_i^g}$ may fall out of $\mathcal{P}_{\text{d}_\infty}(\theta_i^g)$, and the probabilistic performance guarantee for the UEs in a solution may not hold. If θ_i^g is chosen too large (overly conservative), then we will use more transmission powers of the BS than what's necessary. Comparing these two cases, it is clear that the first consequence is more detrimental. So we can choose a sufficiently large θ_i^g that can provide a probabilistic performance guarantee to the UE data rates but may incur a slightly higher BS power consumption.

We propose to calculate θ_i^g based on fast heuristics for each sliding window before executing LSBF. For each sliding window and a UE $i \in \mathcal{K}^g,\ g \in \mathcal{G}$, we have N CSI data samples $\hat{\mathbf{h}}_i^g(n)$. Then we choose θ_i^g based on a constant factor and the estimated variance from the N CSI data samples, i.e.,

$$\theta_i^g = \frac{\rho}{N}\cdot\sqrt{\frac{1}{N-1}\sum_{n\in\mathcal{N}}\left(||\hat{\mathbf{h}}_i^g(n) - \frac{\sum_{n\in\mathcal{N}}\hat{\mathbf{h}}_i^g(n)}{N}||_2^2\right)} \quad (i \in \mathcal{K}^g,\ g \in \mathcal{G})\,. \tag{40}$$

In (40), ρ is the constant factor we need to choose and the term inside the square root is the unbiased sample variance [51]. The use of ρ/N is because θ_i^g is related to the neighboring region of each CSI data sample. Once ρ is given, θ_i^g can be easily calculated based on the N CSI data samples in the current window.

We propose to choose ρ based on a data-driven approach. For a specific setting, the actual threshold violation probabilities are non-increasing w.r.t. ρ and the objective values are non-decreasing w.r.t. ρ (similar to θ_i^g). Therefore, we can perform multiple runs for a range of ρ and calculate the actual SINR threshold violation probabilities and the achieved objective value under each ρ. Then we can pick ρ such that the probabilistic SINR threshold is guaranteed and possibly with a minor performance increase in the objective value. Clearly, this approach can be easily applied to general network settings. In practice, ρ (or θ_i^g) can be dynamically tuned during run-time by keeping track of the actual SINR threshold violation probabilities at the UEs.

VI. Simulation Results

In this section, we evaluate the performance of LSBF. We will focus on the actual SINR threshold violation probability and achieved objective value.

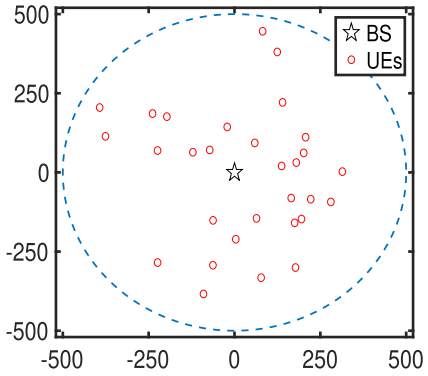


Fig. 5. Topology of a 5G cell with 30 UEs.

A. Simulation Settings

We consider a 5G cell with a 500-meter radius and the network topology is shown in Fig. 5. There are $K = 30$ UEs randomly distributed inside the cell. The number of antennas M at the BS and the number of UEs per RBG $|\mathcal{K}^g|$ will be specified for each simulation below.

We assume the BS has $G = 8$ RBGs and each RBG consists of $S = 8$ RBs. We consider 5G numerology 0 where the sub-carrier spacing is 15 KHz [38]. Therefore, each RBG covers 1 millisecond in the time domain and has a bandwidth of 12×15 KHz $\times\ 8 = 1.44$ MHz. The BS has a power budget $P^{\max} = 46$ dBm for all 8 RBGs and the thermal noise σ_i^2 is set to -150 dBm/Hz for all UEs. For the required SINR threshold ζ_i^{req}, we set it according to Shannon Theorem, $(500/d_i) = \log_2(1 + \zeta_i^{\text{req}})$, where 500 is the cell radius and d_i is the distance between UE i and the BS (both in meters) [52]. This simulates the lower data rate for the edge UEs in a practical 5G cell.

For the wireless channel, we consider the path-loss model and Rician fading. The path-loss between UE i and the BS is modeled by $PL_i = 38 + 30 \times \log_{10}(d_i)$ (in dB) [53]. We employ Rician fading with a 10 dB Rician factor [54], which is a common model for correlated RBs. In addition to the channel variation, we also need to include the CSI estimation errors during the collection of CSI data samples. Therefore, we employ a truncated Gaussian distribution to simulate the CSI estimation errors [17], [19]. Specifically, we use 0 as the mean and 0.1 as the variance for the original Gaussian distribution and then truncate it at three times its standard deviation. Note that the channel model described here is used only for generating parameters in our numerical studies. LSBF only relies on the CSI data samples and is blindfolded with respect to any knowledge of distribution information.

We use MOSEK 9.2.38 on MATLAB R2017b to run all algorithms and each solution includes beamforming vectors $\mathbf{w}_i^g$ on all RBGs (the output of LSBF) For benchmarking, we run results from the following two approaches:

- State-of-the-art Gaussian Approximation [19], where the uncertain CSI is assumed to follow Gaussian distribution with estimated mean and covariance. This approach provides probabilistic SINR threshold guarantee to the UEs.
- Mean Approximation where the means of $\hat{\mathbf{h}}_i^g(n)$ is used as the perfect CSI. This approach gives us a deterministic formulation with constants CSI. Then a classical solution with perfect CSI (e.g., [4]) can be employed.

For each setting below, we perform 50 runs and all results shown represent the average. In each run, we randomly pick $|\mathcal{K}^g|$ UEs from the 30 UEs for each RBG. Since we have 8 RBGs, there can be at most $8|\mathcal{K}^g|$ active UEs at the same time. Given that one UE can be served by multiple RBGs, the number of active UEs may be lower than $8|\mathcal{K}^g|$.

B. A Case Study

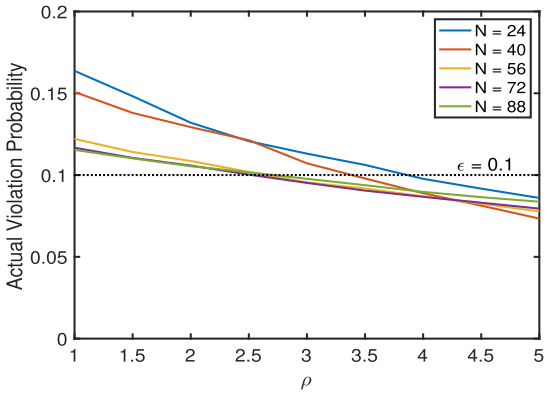
In this subsection, we use a case study to evaluate LSBF and understand its behavior. Although LSBF can handle different values of ϵ_i for the UEs, we use the same value for all UEs (i.e., $\epsilon_i = \epsilon,\ i \in \mathcal{K}$). This allows us to evaluate LSBF under varying risk levels for all UEs, ranging from 0.1 to 0.5. The stop criterion in LSBF is set to $\Delta = 0.03$, i.e., iteration will stop when $t^{\text{UB}} - t^{\text{LB}} \leq \Delta$.

1) Choose N and θ_i^g: We first validate the approach of choosing θ_i^g under a given N presented in Section V-D and also answer one interesting question regarding N in LSBF: *How many CSI data samples do we need?* Intuitively, we would like to choose a smaller N to reduce the computation complexity while achieving a satisfactory MU-MIMO beamforming performance. Thus, we propose to run the simulation for a series of N and find the minimum N such that the probabilistic SINR guarantee of the UEs is met and the BS transmission power is minimized. For each N, we employ different θ_i^g and show how to choose a proper θ_i^g using the approach in Section V-D. Recall that our method to choose θ_i^g is based on the unbiased sample variance of the collected N CSI data samples and a long-term parameter ρ, as given in (40). Therefore, we need to run LSBF under different combinations of N and ρ. Specifically, we set $\epsilon_i = 0.1$ for all $i \in \mathcal{K}$ and run LSBF under $24 \leq N \leq 88$ and $1 \leq \rho \leq 5$. The actual violation probabilities and the achieved objectives are given in Fig. 6.

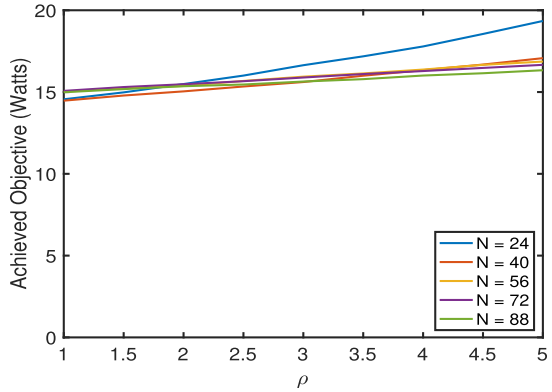
As shown in Fig. 6(a), the actual threshold violation probabilities are decreasing w.r.t. ρ, which shows that the actual but unknown distribution is more likely to be included in ∞-*Wasserstein ambiguity set* $\mathcal{P}_{\mathrm{d}_\infty}(\theta_i)$. As for the achieved objectives shown in Fig. 6(b), they only increase slightly w.r.t. ρ. Further, there is a notable gap in the objectives achieved by $N = 24$ and $N = 40$. Taking both Fig. 6(a) and Fig. 6(b) into account, we conclude that it is prudent to choose $N = 40$ and $\rho = 4$ when calculating θ_i^g in (40). This setting is what we have shown in Fig. 2 where a sliding window covers 6 TTIs and we use 40 CSI data samples from the first 5 TTIs to design beamforming vectors for the 6th TTI.

2) Performance: We now present the performance of LSBF. For a fair comparison, all three algorithms' actual violation probability and achieved objective value will be evaluated under the same target risk level ϵ. This is because the theoretical guarantee of UE data rates can only be claimed using the target risk level ϵ.

Fig. 7 shows the actual violation probabilities and the achieved objective values. As shown in Fig. 7(a), the violation

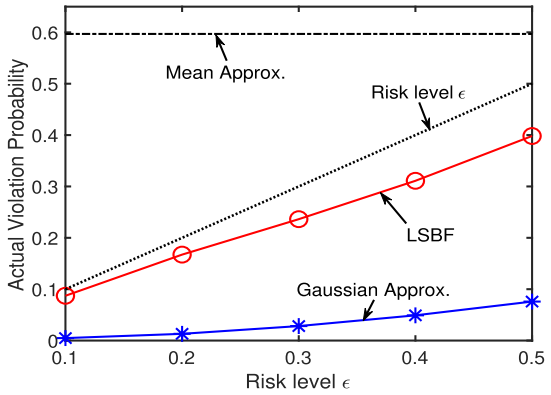


(a) Actual violation probability

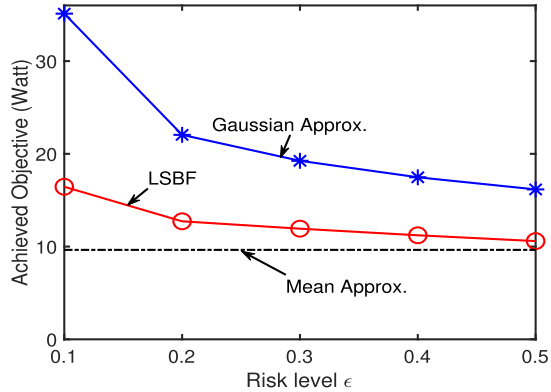


(b) Objective value

Fig. 6. Performance of LSBF as a function of ρ when $\epsilon = 0.1$.



(a) Actual violation probability



(b) Objective value

Fig. 7. Performance of LSBF as a function of risk level ϵ.

probabilities from LSBF are below the risk level ϵ at each point, which affirms the probabilistic performance guarantee for the UEs. Further, the violation probabilities from Gaussian Approximation are much smaller than that from LSBF, which demonstrates the conservativeness of this approach. Finally, we see that the violation probabilities from Mean Approximation are around 60%, which exceeds the target risk level ϵ. The violation probability under Mean Approximation is constant because it does not consider any threshold violations, i.e., independent of ϵ.

Fig. 7(b) shows the achieved objective values (BS's power consumption on all RBGs) w.r.t. ϵ. In this figure, we find that the objective values of both LSBF 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 to meet the loose SINR threshold requirement. Second, compared to the maximum power budget (46 dBm), LSBF can save 59% transmission power even with $\epsilon = 0.1$, which is significant. Third, LSBF performs better than Gaussian Approximation with 53% power saving when $\epsilon = 0.1$. which demonstrates that LSBF offers a tighter approximation than Gaussian Approximation.

Finally, the (unknown) optimal objective value should appear between LSBF and the lower bound from the Mean Approximation (which is overly aggressive and cannot meet target violation probabilities). So our results in Fig. 7(b) suggest that the achieved objectives by LSBF are very close to the (unknown) optimal objective values.

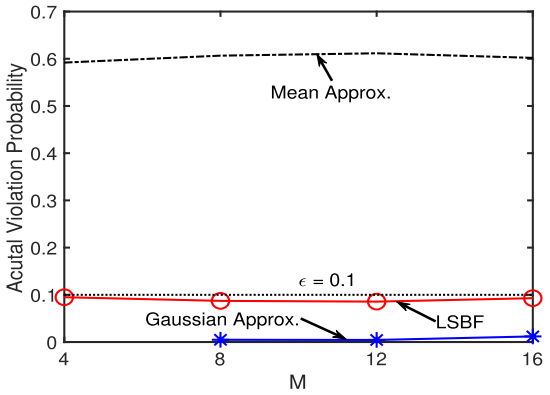
C. Varying Parameters

In this section, we evaluate the performance of LSBF under varying number of BS antennas M and number of UEs per RBG $|\mathcal{K}^g|$. To focus on the impact of varying parameters, we fix the risk level $\epsilon = 0.1$ (i.e., 90% probabilistic guarantee of UE data rates) and all other settings are the same with Section VI-B.

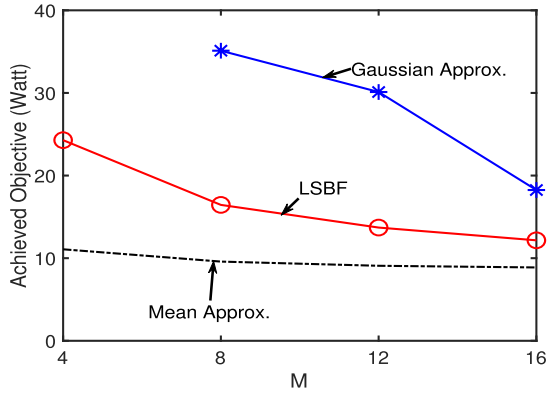
1) Varying Number of BS Antennas M: We fix the number of UEs per RBG $|\mathcal{K}^g| = 2$ and change the number of transmit antennas at the BS from $M = 4$ to $M = 16$. Since Gaussian Approximation fails to find a feasible solution in many runs when $M = 4$ due to its conservativeness, we only include this benchmark when $M = 8 \sim 16$. The threshold violation probabilities and the achieved objectives are shown in Fig. 8.

As shown in Fig. 8(a), the actual threshold violation probabilities from LSBF are all smaller than $\epsilon = 0.1$, which means that it provides a probabilistic guarantee with a varying number of transmit antennas at the BS. As for the objective values shown in Fig. 8(b), LSBF saves 47% power consumption on average compared to that from Gaussian Approximation. Further, the objective values monotonically decrease w.r.t. M due to diversity gain from an increasing number of BS transmit antennas.

2) Varying Number of UEs per RBG $|\mathcal{K}^g|$: We fix the number of antennas $M = 16$ and vary the number of UEs per RBG $|\mathcal{K}^g|$ from 1 to 4. The threshold violation probability and objective values are shown in Fig. 9. Similar to the case of

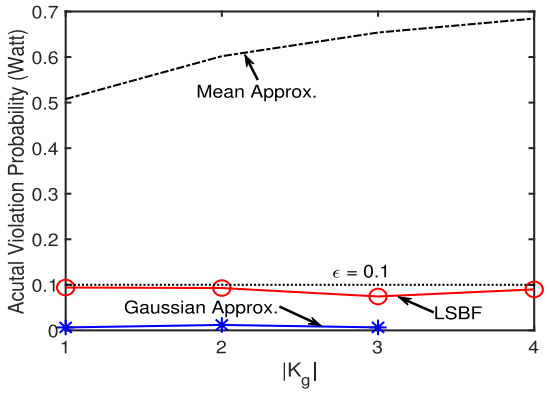


(a) Actual violation probability

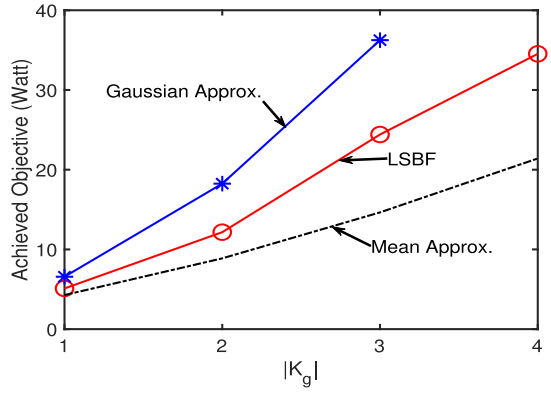


(b) Objective value

Fig. 8. Performance of LSBF under varying M when $\epsilon = 0.1$.



(a) Actual violation probability



(b) Objective value

Fig. 9. Performance of LSBF under varying $|\mathcal{K}^g|$ when $\epsilon = 0.1$.

varying M, we only include Gaussian Approximation when $|\mathcal{K}^g|$ is between 1 and 3.

As shown in Fig. 9(a), the actual threshold violation probabilities from LSBF are all smaller than $\epsilon = 0.1$, which means that it provides probabilistic performance guarantee with varying UEs per RBG. As for the objective values shown in Fig. 9(a), they are monotonically increasing w.r.t. $|\mathcal{K}^g|$. This is expected since the BS consumes more transmit power to serve more UEs on a fixed RBG. All other conclusions are the same with the case of varying M.

VII. Conclusion

We presented a novel approach called Limited Sample-based Beamforming (LSBF) for MU-MIMO beamforming that only requires a limited number of CSI data samples (without any distribution knowledge). LSBF combines the flexibility of data-driven approaches and the ability to provide performance guarantees from model-based approaches. Our goal is to design an MU-MIMO beamforming solution that provides a probabilistic guarantee to UE data rates and minimizes the BS's power consumption. We formulated a chance-constrained problem (CCP) and decomposed it into independent subproblems across RBGs. For each subproblem, we introduced the ∞-Wasserstein ambiguity set to incorporate the limited CSI data samples and substitute the true but unknown CSI distribution. Through the development of a novel bilevel formulation and convex approximation of its lower-level problem, we demonstrated that LSBF can effectively derive an MU-MIMO beamforming solution with polynomial time complexity. Simulation experiments confirmed that LSBF can achieve better performance comparing to the state-of-the-art approaches while meeting the probabilistic data rate requirements of the UEs.

Appendix A
Proof of Lemma 1

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.

Suppose P2 has an optimal solution $\boldsymbol{\pi}_2^*$. Then we check whether or not $\boldsymbol{\pi}_2^*$ satisfies constraint (1), which leads to two cases.

Case (i): If constraint (1) is satisfied, then $\boldsymbol{\pi}_2^*$ is also feasible to P1. Since the feasible region of P2 contains that of P1 and P1 and P2 have the same objective (minimize BS transmission power), $\boldsymbol{\pi}_2^*$ must be an optimal solution to P1.

Case (ii): If constraint (1) is not satisfied, we now show that P1 is infeasible. We prove this statement by contradiction. Suppose P1 is feasible and has an optimal solution $\boldsymbol{\pi}_1^*$. Then its objective value must be smaller than or equal to $P^{\max}$ due to constraint (1). On the other hand, Case (ii) assumes constraint (1) is not satisfied by $\boldsymbol{\pi}_2^*$, i.e., $\boldsymbol{\pi}_2^*$'s objective value is greater than $P^{\max}$. But this contradicts to the fact that $\boldsymbol{\pi}_2^*$ is the optimal solution to P2 since $\boldsymbol{\pi}_1^*$ has a lower objective than

$\boldsymbol{\pi}_2^*$ (both P1 and P2 share the same minimization objective function). Therefore, P1 must be infeasible.

Combining both cases, the proof is complete. ■

REFERENCES

[1] S. Li, N. Jiang, Y. Chen, Y. T. Hou, W. Lou, and W. Xie, "D^{2}BF—Data-driven beamforming in MU-MIMO with channel estimation uncertainty," in *Proc. IEEE INFOCOM*, May 2022, pp. 120–129.

[2] 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 Commun. Mag.*, vol. 52, no. 9, pp. 63–69, Sep. 2014.

[3] M. Codreanu, A. Tolli, M. Juntti, and M. Latva-aho, "Joint design of Tx-Rx beamformers in MIMO downlink channel," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4639–4655, Sep. 2007.

[4] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[5] I. Ahmed et al., "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.

[6] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 172–179, Feb. 2013.

[7] S. Serbetli and A. Yener, "Transceiver optimization for multiuser MIMO systems," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 214–226, Jan. 2004.

[8] Y. Chen, Y. Huang, C. Li, Y. Thomas Hou, and W. Lou, "Turbo-HB: A novel design and implementation to achieve ultra-fast hybrid beamforming," in *Proc. IEEE INFOCOM*, Jul. 2020, pp. 1489–1498.

[9] Y. Wu, R. H. Y. Louie, and M. R. McKay, "Analysis and design of wireless ad hoc networks with channel estimation errors," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1447–1459, Mar. 2013.

[10] Y. Liu, Z. Tan, H. Hu, L. J. Cimini, and G. Y. Li, "Channel estimation for OFDM," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1891–1908, 4th Quart., 2014.

[11] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, 2013.

[12] C. Kong, C. Zhong, A. K. Papazafeiropoulos, M. Matthaiou, and Z. Zhang, "Sum-rate and power scaling of massive MIMO systems with channel aging," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4879–4893, Dec. 2015.

[13] X. Rao and V. K. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.

[14] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2868–2882, May 2015.

[15] X. Jiang and F. Kaltenberger, "Channel reciprocity calibration in TDD hybrid beamforming massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 422–431, Jun. 2018.

[16] X. Jiang et al., "A framework for over-the-air reciprocity calibration for TDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5975–5990, Sep. 2018.

[17] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.

[18] Y. Xu, X. Zhao, and Y.-C. Liang, "Robust power control and beamforming in cognitive radio networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1834–1857, 4th Quart., 2015.

[19] K. Wang, A. M. So, T. Chang, W. Ma, and C. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5690–5705, Nov. 2014.

[20] M. Botros Shenouda, T. N. Davidson, and L. Lampe, "Outage-based design of robust Tomlinson–Harashima transceivers for the MISO downlink with QoS requirements," *Signal Process.*, vol. 93, no. 12, pp. 3341–3352, Dec. 2013.

[21] Y. Shi, J. Zhang, and K. B. Letaief, "Optimal stochastic coordinated beamforming for wireless cooperative networks with CSI uncertainty," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 960–973, Feb. 2015.

[22] 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 Trans. Wireless Commun.*, vol. 18, no. 2, pp. 780–795, Dec. 2019.

[23] 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 J. Wireless Commun. Netw.*, vol. 2012, no. 1, pp. 1–11, Aug. 2012.

[24] M. B. Shenouda and T. N. Davidson, "Nonlinear and linear broadcasting with QoS requirements: Tractable approaches for bounded channel uncertainties," *IEEE Trans. Signal Process.*, vol. 57, no. 5, pp. 1936–1947, May 2009.

[25] A. Rico-Alvariño and R. W. Heath, "Learning-based adaptive transmission for limited feedback multiuser MIMO-OFDM," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3806–3820, Jul. 2014.

[26] 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. Wireless Commun.*, vol. 21, no. 4, pp. 2368–2382, Apr. 2022.

[27] P. Mertikopoulos and A. L. Moustakas, "Learning in an uncertain world: MIMO covariance matrix optimization with imperfect feedback," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 5–18, Jan. 2016.

[28] M. Alrabeiah, Y. Zhang, and A. Alkhateeb, "Neural networks based beam codebooks: Learning mmWave massive MIMO beams that adapt to deployment and hardware," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3818–3833, Jun. 2022.

[29] Y. Yuan, G. Zheng, K.-K. Wong, B. Ottersten, and Z.-Q. Luo, "Transfer learning and Meta learning-based fast downlink beamforming adaptation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1742–1755, Mar. 2021.

[30] J. Xia and D. Gunduz, "Meta-learning based beamforming design for MISO downlink," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Melbourne, VIC, Australia, Jul. 2021, pp. 2954–2959.

[31] J. Zhang, Y. Yuan, G. Zheng, I. Krikidis, and K.-K. Wong, "Embedding model-based fast meta learning for downlink beamforming adaptation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 149–162, Jan. 2022.

[32] Y. Long and S. Murphy, "Few-shot learning based hybrid beamforming under birth-death process of scattering paths," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1687–1691, May 2021.

[33] M. Vu and A. Paulraj, "MIMO wireless linear precoding," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 86–105, Sep. 2007.

[34] A. H. Mehana and A. Nosratinia, "Diversity of MIMO linear precoding," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1019–1038, Feb. 2014.

[35] *5G; NR; Physical Layer Procedures for Data*, Version 18.2.0, document TS 38.214, 3GPP, May 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216

[36] S. Sur, I. Pefkianakis, X. Zhang, and K.-H. Kim, "Practical MU-MIMO user selection on 802.11ac commodity networks," in *Proc. ACM MobiCom*, New York City, NY, USA, Oct. 2016, pp. 122–134.

[37] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658–3663, Sep. 2006.

[38] *5G; NR; Physical Channels Modulation*, Version 18.2.0, document TS 38.211, 3GPP, May 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213

[39] A. M. Tulino, G. Caire, S. Shamai, and S. Verdu, "Capacity of channels with frequency-selective and time-selective fading," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1187–1215, Mar. 2010.

[40] W. Yang, G. Durisi, and E. Riegler, "On the capacity of large-MIMO block-fading channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 117–132, Feb. 2013.

[41] W. Xie, "On distributionally robust chance constrained programs with Wasserstein distance," *Math. Program.*, vol. 186, no. 1, pp. 115–155, Mar. 2021.

[42] N. Jiang and W. Xie, "ALSO-X and ALSO-X+: Better convex approximations for chance constrained programs," *Operations Res.*, vol. 70, no. 6, pp. 3581–3600, Nov. 2022.

[43] T. Champion, L. De Pascale, and P. Juutinen, "The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps," *SIAM J. Math. Anal.*, vol. 40, no. 1, pp. 1–20, Jan. 2008.

[44] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, no. 1, pp. 115–166, 2018.
[45] G. Zheng, K.-K. Wong, and B. Ottersten, "Robust cognitive beamforming with bounded channel uncertainties," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4871–4881, Dec. 2009.
[46] X. Sun et al., "Joint beamforming and power allocation in downlink NOMA multiuser MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5367–5381, Aug. 2018.
[47] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max–min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.
[48] T.-H. Chang, Z.-Q. Luo, and C.-Y. Chi, "Approximation bounds for semidefinite relaxation of max-min-fair multicast transmit beamforming problem," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3932–3943, Aug. 2008.
[49] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Philadelphia, PA, USA: SIAM, 2001, ch. 3, pp. 203–205.
[50] S. Ahmed, J. Luedtke, Y. Song, and W. Xie, "Nonanticipative duality, relaxations, and formulations for chance-constrained stochastic programs," *Math. Program.*, vol. 162, nos. 1–2, pp. 51–81, Mar. 2017.
[51] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. Hoboken, NJ, USA: Wiley, 2010, ch. 7, pp. 224–226.
[52] J. Wang et al., "Spectral efficiency improvement with 5G technologies: Results from field tests," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1867–1875, Aug. 2017.
[53] *Radio Frequency (RF) Requirements for LTE Pico Node B*, Version 18.0.0, document TR 36.931, 3GPP, Apr. 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2589
[54] X. Li, S. Jin, H. A. Suraweera, J. Hou, and X. Gao, "Statistical 3-D beamforming for large-scale MIMO downlink systems over Rician fading channels," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1529–1543, Apr. 2016.

Shaoran Li (Member, IEEE) received the B.S. degree from Southeast University, Nanjing, China, in 2014, the M.S. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2017, and the Ph.D. degree in electrical engineering from Virginia Tech, in 2022. He is currently a Senior System Software Engineer with NVIDIA Corporation, Santa Clara, CA, USA. His research interests include algorithm design and implementation for wireless networks, with a focus on network uncertainty. He was a recipient of the 2019 Fred W. Ellersick MILCOM Award for the Best Paper in the unclassified technical program.

Nan Jiang (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in operations research with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Tech, Atlanta, GA, USA. Before attending Georgia Tech, he was a Ph.D. student at Virginia Tech, Blacksburg, VA, USA. His research interests include developing new methods or algorithms for decision-making under uncertainty.

Yongce Chen received the B.S. and M.S. degrees in electrical engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2013 and 2016, respectively, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2021. He is currently a Senior System Software Engineer with NVIDIA Corporation, Santa Clara, CA, USA. His current research interests include optimization, MIMO techniques, and GPU-based real-time design and implementation. During his Ph.D. study at Virginia Tech, he was awarded the VT Wireless Fellowship in 2016 and the Pratt Fellowship in 2021. He was a recipient of the Best Paper Award at IEEE INFOCOM 2021.

Weijun Xie (Member, IEEE) received the Ph.D. degree in operations research from Georgia Institute of Technology in August 2017. He is currently an Assistant Professor with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Tech. His research interests include theory and applications of stochastic, discrete, and convex optimization. His works have received multiple awards, such as 2021 NSF CAREER Award, Winner of 2020 INFORMS Young Researchers Paper Prize, Runner-up of Dupacova-Prekopa Best Student Paper Prize in Stochastic Programming at ICSP 2019, Honorable Mention in George Nicholson Student Paper Competition at INFORMS 2017.

Wenjing Lou (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Florida. She is currently the W.C. English Endowed Professor of computer science with Virginia Tech, Arlington, VA, USA. Her research interests cover many topics in the cybersecurity field, with her current research interest focusing on wireless networks, privacy protection in machine learning systems, and security and privacy problems in the Internet of Things (IoT) systems. She is a highly cited researcher by the Web of Science Group. She received the Virginia Tech Alumni Award for Research Excellence in 2018, the highest university level faculty research award. She was a recipient of INFOCOM Test-of-Time paper award in 2020. She was a TPC chair for IEEE INFOCOM 2019 and ACM WiSec 2020. She was the Steering Committee Chair for IEEE CNS conference from 2013 to 2020. She is currently a steering committee member of IEEE INFOCOM. She served as a program director at US National Science Foundation (NSF) from 2014 to 2017.

Y. Thomas Hou (Fellow, IEEE) received the Ph.D. degree from the NYU Tandon School of Engineering in 1998. He was a Member of Research Staff with Fujitsu Laboratories of America, Sunnyvale, CA, USA, from 1997 to 2002. He is currently a Bradley Distinguished Professor of electrical and computer engineering with Virginia Tech, Blacksburg, VA, USA, which he joined, in 2002. He has published over 350 papers in IEEE/ACM journals and conferences. He holds six U.S. patents. He has authored/co-authored two graduate textbooks: *Applied Optimization Methods for Wireless Networks* (Cambridge University Press, 2014) and *Cognitive Radio Communications and Networks: Principles and Practices* (Academic Press/Elsevier, 2009). His current research interests include developing real-time optimal solutions to complex science and engineering problems arising from wireless and mobile networks. He is also interested in wireless security. He was named an IEEE Fellow for contributions to modeling and optimization of wireless networks. His papers were recognized by ten best paper awards from IEEE and ACM, including an IEEE INFOCOM Test of Time Paper Award in 2023. He was/is on the editorial boards of a number of IEEE and ACM transactions and journals. He was the Steering Committee Chair of IEEE INFOCOM Conference and a member of the IEEE Communications Society Board of Governors. He was also a Distinguished Lecturer of the IEEE Communications Society.