Energy-Efficient Processing and Robust Wireless Cooperative Transmission for Edge Inference

Kai Yang, *Student Member, IEEE*, Yuanming Shi, *Member, IEEE*, Wei Yu, *Fellow, IEEE*, and Zhi Ding, *Fellow, IEEE*

***Abstract*—Edge machine learning can deliver low-latency and private artificial intelligent (AI) services for mobile devices by leveraging computation and storage resources at the network edge. This article presents an energy-efficient edge processing framework to execute deep learning inference tasks at the edge computing nodes whose wireless connections to mobile devices are prone to channel uncertainties. Aimed at minimizing the sum of computation and transmission power consumption with probabilistic Quality-of-Service (QoS) constraints, we formulate the joint inference tasking and the downlink beamforming problem that is characterized by a group sparse objective function. We provide a statistical learning-based robust optimization approach to approximate the highly intractable probabilistic-QoS constraints by nonconvex quadratic constraints, which are further reformulated as matrix inequalities with a rank-one constraint via matrix lifting. We design a reweighted power minimization approach by iteratively reweighted ℓ_1 minimization with difference-of-convex-functions (DC) regularization and updating weights, where the reweighted approach is adopted for enhancing group sparsity whereas the DC regularization is designed for inducing rank-one solutions. The numerical results demonstrate that the proposed approach outperforms other state-of-the-art approaches.**

***Index Terms*—Difference-of-convex-functions (DC), edge intelligence, energy efficiency, group sparse beamforming, robust communication, robust optimization.**

Manuscript received December 14, 2019; revised February 28, 2020; accepted March 1, 2020. Date of publication March 10, 2020; date of current version October 9, 2020. The work of Kai Yang and Yuanming Shi was supported in part by the National Nature Science Foundation of China under Grant 61601290. The work of Wei Yu was supported by the Natural Sciences and Engineering Research Council of Canada. The work of Zhi Ding was supported by the National Science Foundation under Grant ECCS-1711823. *(Corresponding author: Yuanming Shi.)*

Kai Yang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, also with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yangkai@shanghaitech.edu.cn).

Yuanming Shi is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China and also with Yoke, Shanghai 201414, China. (e-mail: shiym@shanghaitech.edu.cn).

Wei Yu is with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: weiyu@comm.utoronto.ca).

Zhi Ding is with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616 USA (e-mail: zding@ucdavis.edu).

Digital Object Identifier 10.1109/JIOT.2020.2979523

I. Introduction

MACHINE learning has transformed many aspects of our daily lives by taking advantage of abundant data and computing power in the cloud center. In particular, the strong capability of capturing the representations of data for detection or classification using deep neural networks (DNNs) [1] has made impressive gains in face recognition, natural language processing tasks, etc. With the explosion of mobile data and the increasing edge computing capability, there is an emerging trend of *edge intelligence* [2]–[4], followed by the evolution of future mobile networks from "connected things" to "connected intelligence" [5]. Instead of uploading all data collected by mobile devices to the remote cloud data center, edge intelligence emphasizes the use of the computation and storage resources at network edges to provide low-latency and reliable artificial intelligent (AI) services [6], [7] for privacy/security sensitive devices, such as wearable devices, augmented reality, smart vehicles, and drones. However, since mobile devices are usually equipped with limited computation power, storage, and energy [2], it is usually infeasible to deploy deep learning models, i.e., DNNs, at resource-constrained mobile devices, and execute inference tasks locally. A promising solution is to enable processing at the mobile network access points (APs) to facilitate deep learning inference, which is called *edge inference* [8], [9].

In this article, we shall present the edge processing framework for edge inference (as illustrated in Fig. 1) that the input (e.g., a piece of rough doodle) of each mobile user (MU) is uploaded to wireless APs (e.g., base stations) served as edge computing nodes, each task is performed with pretrained deep learning model (e.g., Nvidia's AI system GauGAN [10] for turning rough doodles into photorealistic landscapes) at multiple edge computing nodes, and the output results (e.g., landscape images) are transmitted to MUs via coordinated beamforming among multiple APs. In such a system, the provisioning of wireless transmissions in both the uplink and the downlink are important design considerations. In addition to the low-latency requirement, improving the energy efficiency [11] is also critical due to the high computational complexity of processing DNNs, for which a number of works focusing on model compression methods [12], [13].

There is a communication and computation tradeoff for the edge inference system in the downlink. In particular, performing an inference task at more edge computing nodes can achieve a higher Quality of Service (QoS) through cooperative

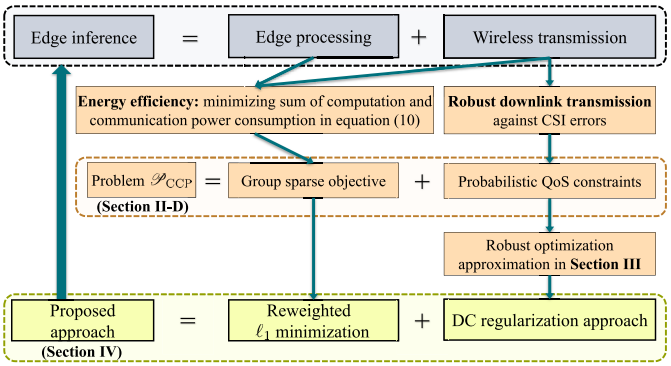


Fig. 1. Illustration of our energy-efficient processing and robust wireless cooperative transmission framework for edge inference.

downlink transmission for delivering the output results to MUs. This, however, results in more computation power consumption for executing the deep learning models. We thus propose to jointly decide on the task allocation strategy at edge nodes and design downlink beamforming vectors by minimizing the sum of transmission power consumption and computation power consumption. In particular, the power consumption of deep learning inference tasks can be determined through the estimated energy [14] and computation time. We observe that there is an intrinsic connection between the group sparse structure [15], [16] of the downlink aggregative beamforming vector and the combinatorial variable, i.e., the set of tasks performed at edge nodes. The cooperative transmission strategies require global channel state information (CSI), while uncertainty in CSI acquisition is inevitable in practice due to the training-based channel estimation [17], limited feedback [18], partial CSI acquisition [19], and CSI acquisition delays [20]. We thus formulate the joint task selection and the downlink beamforming problem for energy-efficient processing and robust transmission against CSI errors in edge inference system as a group sparse beamforming problem with probabilistic-QoS constraints [19].

The joint chance constraints make the formulated probabilistic group sparse beamforming problem highly intractable since it has no closed-form expression generally. To address the chance-constrained programs, a number of works focus on finding computationally tractable approximations based on the collected samples of the random variables. A recognized scenario generation (SG) approach [21] is proposed that uses a collection of sampled constraints to approximate the original chance constraints. However, the SG is overconservative since the volume of the feasible region decreases by increasing the sample size, which leads to the deterioration of its performance. In addition, given the prespecified probability $1-\epsilon$ and the confidence level $1-\delta$ for the probabilistic-QoS constraints, the required samples size of SG should satisfy $\sum_{i=1}^{NKL-1}\binom{T}{i}\epsilon^i(1-\epsilon)^{T-i} \leq \delta$, which increases roughly linearly with $1/\epsilon$. In [19], a stochastic optimization approach is provided to address the overconservativeness of SG. However, its computational cost grows linearly with the sample size, which is not scalable for obtaining high-robustness solutions. Moreover, its statistical guarantee under a finite sample size is still not available. To overcome the limitations of existing methods, we present a robust optimization approximation approach for the joint chance constraints by enforcing the QoS constraints for any element within a high probability region. The high probability region is further determined by adopting a statistical learning [22] approach. This approach enjoys the benefits that the minimum required sample size is only $\log\delta/\log(1-\epsilon)$, and the computational cost is independent of the sample size.

With the statistical learning-based robust optimization approximation approach, the resulting robust group sparse beamforming problem has nonconvex quadratic constraints and a nonconvex group sparse objective function. We find that the nonconvex quadratic constraints can be convexified by matrix lifting and semidefinite relaxation (SDR) [23]. Specifically, the nonconvex quadratic robust QoS constraints can be lifted as convex constraints in terms of a rank-one positive semidefinite (PSD) matrix variable, which is then convexified by simply dropping the rank-one constraint. However, the SDR approach cannot guarantee that the obtained solution is feasible with respect to the original nonconvex quadratic constraints. The mixed ℓ_1/ℓ_2-norm [24] is a well-known convex group sparsity inducing norm, which has been successfully applied in green cloud radio access networks [15] and cooperative wireless cellular network [25]. However, the SDR approach requires a quadratic form of the objective function, which makes the mixed ℓ_1/ℓ_2-norm minimization approach inapplicable. To overcome this problem, a quadratic variational form of weighted mixed ℓ_1/ℓ_2-norm is proposed in [26] to induce group sparsity. Note that Shi *et al.* [26] also considers a group sparse beamforming problem with nonconvex quadratic constraints. However, the performance of a quadratic variational form of weighted mixed ℓ_1/ℓ_2-norm minimization with SDR is still not satisfactory.

To address the limitations of existing approaches, we propose a reweighted power minimization approach to enhance the group sparsity as well as improve the feasibility of nonconvex quadratic constraints. Specifically, we first adopt the iteratively reweighted ℓ_1 minimization approach for enhancing group sparsity [27], [28]. To further guarantee the feasibility of the original nonconvex quadratic constraints, we exploit the matrix lifting technique to recast the nonconvex quadratic constraints as the convex constraints with respect to a rank-one PSD matrix and propose a novel difference-of-convex-functions (DC) regularization approach to induce rank-one solutions. Numerical results demonstrate that the proposed approach improves the probability of feasibility by avoiding the overconservativeness of SG. Benefiting from both the reweighted ℓ_1 minimization and the DC regularization, the proposed approach achieves a much lower total power consumption than the algorithm proposed in [26] and has a better capability of inducing group sparsity with nonconvex quadratic constraints.

A. Contributions

In this article, we consider an edge computing system to execute deep learning inference tasks for resource-constrained

mobile devices. In order to provide energy-efficient processing and robust wireless cooperative transmission service for edge inference, we propose to jointly design the downlink beamforming vector and the set of inference tasks performed at each edge computing nodes under probabilistic-QoS constraints. We provide a statistical learning-based robust optimization approximation for the highly intractable joint chance constraints, which guarantees that the probabilistic-QoS constraints are feasible with a certain confidence level. The resulting problem turns out to be a group sparse beamforming problem with nonconvex quadratic constraints. We propose a reweighted power minimization approach based on the principles of iteratively reweighted ℓ_1 minimization for group sparsity inducing, matrix lifting technique, and a novel DC representation for rank-one PSD matrices. The proposed approach can enhance group sparsity and induce rank-one solutions.

We summarize the major contributions of this article as follows.

1) We propose an energy-efficient processing and robust transmission approach for executing deep learning inference tasks at possibly multiple edge computing enabled wireless APs. The selection of an optimal set of APs for each task is formulated as a group sparse beamforming problem with joint chance constraints.
2) We provide a robust optimization counterpart to approximate the joint chance constraints followed by a statistical learning approach to learn the parameters from data samples of the random channel coefficients. It turns out a nonconvex group sparse beamforming problem with nonconvex quadratic constraints.
3) We show that the nonconvex quadratic constraints can be reformulated as convex constraints with a rank-one constraint, where the rank-one constraint can be reformulated with a novel DC representation. To enhance the group sparsity and inducing rank-one solutions, we propose a reweighted power minimization approach by iteratively reweighted ℓ_1 minimization with DC regularization and updating weights.
4) We conduct extensive numerical experiments to demonstrate the advantages of the proposed approach in providing energy efficient and robust transmission service for edge inference.

B. Organization and Notations

The remainder of this article is organized as follows. In Section II, we introduce the system model and the power consumption model of edge inference and formulate the energy-efficient processing and robust cooperative transmission problem as a group sparse beamforming problem with joint chance constraints. Section III provides a statistical learning-based robust optimization approach to approximate the joint chance constraints. In Section IV, we design a reweighted power minimization approach for solving the robust group sparse beamforming problem. The simulation results are illustrated in Section V to demonstrate the superiority of the proposed approach over other state-of-the-art approaches. Finally, we conclude this article in Section VI.

TABLE I
NOTATIONS

Notation	Explanation
N, K, L	the number of APs, MUs, and AP's antennas, respectively
$[K]$	the set of $\{1, \cdots, K\}$
$\mathcal{A}$	task allocation of APs
P_{nk}^c	power consumption of performing the k-th user's task at the n-th AP
P_n^{Tx}	maximum transmit power of the n-th AP
η_n	power amplifier efficiency
P^c	total computation power consumption at APs
P	total power consumption
$\boldsymbol{v}_{nk}, \boldsymbol{v}_k, \boldsymbol{v}$	beamforming vectors at the APs
$\boldsymbol{V}_{ij}[s,t], \boldsymbol{V}_{ij}, \boldsymbol{V}$	lifted matrices of beamforming vectors
$\boldsymbol{h}_{kn}, \boldsymbol{h}_k, \boldsymbol{h}$	downlink channel coefficient vectors between APs and MUs
$\hat{\boldsymbol{h}}_{kn}, \hat{\boldsymbol{h}}_k, \hat{\boldsymbol{h}}$	estimated channel coefficient vectors
$\boldsymbol{e}_{kn}, \boldsymbol{e}$	random errors of CSI
γ_k, ζ	the target QoS and its target tolerance level
ϵ, δ	the tolerance level and its confidence level
$\mathcal{U}_k$	high probability region of $\boldsymbol{h}_k$
$\mathcal{D}$	the data set consisting of D i.i.d. samples of $\boldsymbol{h}$
$\mathcal{D}^1, \mathcal{D}^2$	the partitioned two parts of the data set $\mathcal{D}$ with size D_1 and $D_2 = D - D_1$, respectively
$\tilde{\boldsymbol{h}}^{(j)}$	the j-th data sample
$q_{1-\epsilon}$	$(1-\epsilon)$-quantile

Throughout this article, we use lowercase bold letters (e.g., $\boldsymbol{v}$) to denote column vectors and letters with one subscript to denote their subvectors (e.g., $\boldsymbol{v}_k$). We further use lowercase bold letters with two subscripts to denote the subvectors of subvectors (e.g., $\boldsymbol{v}_{nk}$ is a subvector of $\boldsymbol{v}_k$). We denote scalars with lowercase letters, matrices with capital letters (e.g., $\boldsymbol{V}$), and sets with calligraphic letters (e.g., $\mathcal{A}$). The conjugate transpose of a vector or matrix, ℓ_2-norm of a vector, and spectral norm of a matrix are denoted as $(\cdot)^{\mathsf{H}}$, $\|\cdot\|_2$, and $\|\cdot\|$, respectively. Table I summarizes the notations used in this article.

II. System Model and Problem Formulation

This section provides the system model and power consumption model of edge inference for DNNs, followed by the proposal of the energy-efficient edge processing under probabilistic-QoS constraints.

A. System Model

Consider the edge processing network consisting of N L-antenna edge computing-enabled wireless APs and K single-antenna MUs, as shown in Fig. 2. Each MU k has a deep learning inference task $\phi_k(d_k)$ with input d_k. Instead of relying on a cloud data center, we execute deep learning tasks at the APs to address latency and privacy concerns for high-stake applications, such as drones and smart vehicles [2]. In this article, we propose to store the trained DNN models ϕ_ks to APs in advance. Each AP collects all inputs $\{d_k\}_{k=1}^K$ from each MU in the first phase. In the second phase, each AP will selectively execute some inference tasks and transmit the output results to the MUs through cooperative downlink transmission, thereby providing low-latency intelligent services for MUs. The point is that the same inference task can be executed at multiple APs so that the multiple APs can jointly transmit the result

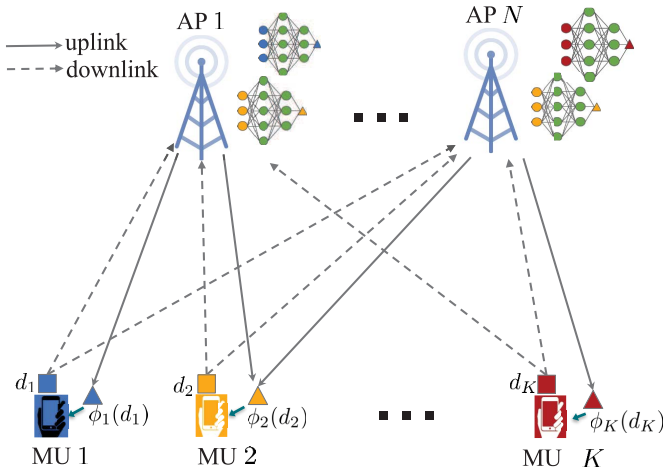


Fig. 2. System model of edge inference for DNNs. This article focuses on the computing and downlink transmission phase.

to the MUs through beamforming, thus improving the downlink transmission efficiency (at the expense of larger energy consumption due to executing the same task at multiple APs). This article focuses on the joint task selection and downlink transmit beamforming problem in the second phase.

Let $\phi_k(d_k)$ be the requested output for MU k, $s_k \in \mathbb{C}$ be the encoded scalar to be transmitted, and $\boldsymbol{v}_{nk} \in \mathbb{C}^L$ be the beamforming vector for message $\phi_k(d_k)$ at the nth AP. We consider the downlink communication scenario, where all inputs d_k's have already been collected at APs. Then, the received signal at MU l is given by

$$y_k = \sum_{n=1}^{N} \sum_{l=1}^{K} \boldsymbol{h}_{kn}^{\mathsf{H}} \boldsymbol{v}_{nl} s_l + z_k \tag{1}$$

where $\boldsymbol{h}_{kn} \in \mathbb{C}^L$ is the channel coefficient vector between the nth AP and the kth MU, and $z_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the additive isotropic white Gaussian noise. Suppose all data symbols s_k,s are mutually independent with unit power, i.e., $\mathbb{E}[|s_k|^2] = 1$, and also independent with the noise. Denote $[K]$ as the set $\{1, \ldots, K\}$. Let $\mathcal{A} \subseteq \{(n,k) : n \in [N], k \in [K]\}$ denote a feasible allocation for the inference tasks on APs, i.e., computational task ϕ_k shall be performed at the nth AP for $(n,k) \in \mathcal{A}$. In terms of the group sparsity structure of the aggregative beamforming vector

$$\boldsymbol{v} = \left[\boldsymbol{v}_{11}^{\mathsf{H}}, \ldots, \boldsymbol{v}_{N1}^{\mathsf{H}}, \ldots, \boldsymbol{v}_{NK}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}^{NKL} \tag{2}$$

we have that if the inference task k will not be performed at AP n, i.e., $(n,k) \notin \mathcal{A}$, the beamforming vector $\boldsymbol{v}_{nk}$ will be set as zero. Let $\mathcal{T}(\boldsymbol{v})$ be the group sparsity pattern of $\boldsymbol{v}$ given as

$$\mathcal{T}(\boldsymbol{v}) = \{(n,k) | \boldsymbol{v}_{nk} \neq \boldsymbol{0}\}. \tag{3}$$

The signal-to-interference-plus-noise ratio (SINR) for mobile device k is given by

$$\mathrm{SINR}_k(\boldsymbol{v}; \boldsymbol{h}_k) = \frac{|\boldsymbol{h}_k^{\mathsf{H}} \boldsymbol{v}_k|^2}{\sum_{l \neq k} |\boldsymbol{h}_k^{\mathsf{H}} \boldsymbol{v}_l|^2 + \sigma_k^2} \tag{4}$$

where $\boldsymbol{h}_k$ and $\boldsymbol{v}_k$ are given by

$$\boldsymbol{h}_k = \left[\boldsymbol{h}_{k1}^{\mathsf{H}}, \ldots, \boldsymbol{h}_{kN}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}^{NL} \tag{5}$$

$$\boldsymbol{v}_k = \left[\boldsymbol{v}_{1k}^{\mathsf{H}} \ \cdots \ \boldsymbol{v}_{Nk}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}^{NL} \tag{6}$$

and the aggregative channel coefficient vector is denoted as

$$\boldsymbol{h} = \left[\boldsymbol{h}_1^{\mathsf{H}}, \ldots, \boldsymbol{h}_K^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}^{NKL}. \tag{7}$$

The transmit power constraint at the nth AP is given by

$$\mathbb{E}\left[\sum_{l=1}^{K} \|\boldsymbol{v}_{nl} s_l\|_2^2\right] = \sum_{l=1}^{K} \|\boldsymbol{v}_{nl}\|_2^2 \leq P_n^{\mathrm{Tx}}, \qquad n \in [N] \tag{8}$$

where P_n^{Tx} is the maximum transmit power.

B. Power Consumption Model

Although widespread applications of deep learning present numerous opportunities for intelligent systems, energy consumption becomes one of the main concerns [8]. Indeed, the energy consumption of performing DNN inference is dominated by memory access. As pointed out in [12], memory access of 32-b dynamic random access memory (DRAM) consumes 640 pJ, while a cache access of 32-b static random access memory (SRAM) consumes 5 pJ and a 32-b floating point add operation consumes 0.9 pJ. Large DNN models probably cannot fit in the storage of mobile devices which requires more costly DRAM memory accesses. Therefore, small models can be directly deployed on mobile devices but large models are preferably executed at the powerful edge nodes. Let the power consumption of computing task ϕ_k at the nth edge computing node be P_{nk}^{c}. The total computation power consumption for all edge computing nodes is thus given by

$$P^{\mathrm{c}} = \sum_{n,k} P_{nk}^{\mathrm{c}} I_{(n,k) \in \mathcal{T}(\boldsymbol{v})} \tag{9}$$

where the indicator function I is 1 if $(n,k) \in \mathcal{T}(\boldsymbol{v})$ and 0 otherwise. Therefore, the total power consumption consists of transmission power consumption for output results delivery and computation power consumption for deep learning tasks execution, which is given by

$$P = \sum_{n,k} \frac{1}{\eta_n} \|\boldsymbol{v}_{nk}\|_2^2 + \sum_{n,k} P_{nk}^{\mathrm{c}} I_{(n,k) \in \mathcal{T}(\boldsymbol{v})} \tag{10}$$

where η_n is the power amplifier efficiency.

DNNs, especially deep convolutional neural networks (CNNs), become an indispensable and the state-of-the-art paradigm for the real-world intelligent services. Its high energy cost has attracted much interest in designing energy-efficient structures of neural networks [12]. Estimating the energy consumption of a neural network is thus critical for inference at the edge, for which an estimation tool is developed in [29]. The energy consumption of performing an inference task consists of the computation part and the data movement part [14]. The computation energy consumption can be calculated by counting the number of multiply-and-accumulate (MACs) in the layer and weighing it with the energy consumption of each

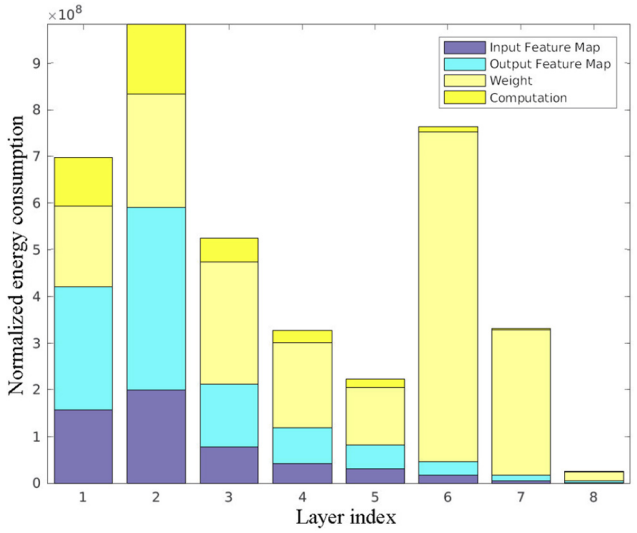


Fig. 3. Energy consumption breakdown of the AlexNet [30]. The unit of energy is normalized by the energy for one MAC operation (i.e., 10^2 = energy of 100 MACs).

MAC operation in the computation core. The energy consumption of data movement is calculated by counting the number of accessing memory at each level of the memory hierarchy in the corresponding hardware and weighing it with the energy consumption of accessing the memory in the corresponding level.

Here, we illustrate how to estimate the computation power consumption of performing image classification tasks using the classic CNN (i.e., AlexNet consisting of five convolutional layers and three fully connected layers) on the Eyeriss chip. The energy estimation tool takes network configuration as input and outputs the estimated energy breakdown of each layer in terms of computation part and the data movement part of three data types (weight, input feature map, and output feature map). Fig. 3 demonstrates the estimated energy of each layer running on Eyeriss chip, and the overall energy consumption is the sum of four parts. The unit of energy is normalized by the energy for one MAC. Based on the total energy consumption, the computation power consumption can be further determined by dividing the energy consumption by the computation time.

C. Channel Uncertainty Model

For high-stake intelligent applications, such as autonomous driving and automation, robustness is a critical requirement. In practice, inevitably there is uncertainty in the available CSI $\boldsymbol{h}$, which is taken into consideration to provide robust transmission in this article. It may originate from training-based channel estimation [17], limited precision of feedback [18], partial CSI acquisition [19], and delays in CSI acquicition [20]. In this article, we adopt the additive error model [31], [32] of the channel imperfection, i.e.,

$$\boldsymbol{h} = \hat{\boldsymbol{h}} + \boldsymbol{e} \tag{11}$$

where $\hat{\boldsymbol{h}} \in \mathbb{C}^{NKL}$ is the estimated aggregative channel vector and $\boldsymbol{e} \in \mathbb{C}^{NKL}$ is the random errors of the CSI with unknown distribution and expectation as $\mathbf{0}$. We apply the probabilistic-QoS constraints [19] to characterize the robustness of delivering the inference results to MUs

$$\Pr(\mathrm{SINR}_k(\boldsymbol{v}; \boldsymbol{h}_k) \geq \gamma_k) \geq 1 - \zeta \quad \forall k \in [K]. \tag{12}$$

Here, ζ is the tolerance level and "$\mathrm{SINR}_k \geq \gamma_k$" is called safe condition.

D. Problem Formulation

In the proposed edge processing framework for deep learning inference tasks, there is a fundamental tradeoff between computation and communication. Specifically, executing the same inference task at multiple edge nodes will require higher computation power consumption, while the downlink transmission power consumption shall be reduced due to the cooperative transmission gains. In this article, we propose an energy-efficient processing and robust transmission approach to minimize the total network power consumption, while satisfying the probabilistic-QoS constraints and transmit power constraints. It is formulated as the following probabilistic group sparse beamforming problem:

$$\mathscr{P}_{\mathrm{CCP}}: \min_{\boldsymbol{v} \in \mathbb{C}^{NKL}} \sum_{n,k} \frac{1}{\eta_n} \|\boldsymbol{v}_{nk}\|_2^2 + \sum_{n,k} P_{nk}^{\mathrm{c}} I_{(n,k) \in \mathcal{T}(\boldsymbol{v})}$$
$$\text{s.t.} \quad \Pr(\mathrm{SINR}_k(\boldsymbol{v}; \boldsymbol{h}_k) \geq \gamma_k) \geq 1 - \zeta, \qquad k \in [K] \tag{13}$$

$$\sum_{k=1}^{K} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_n^{\mathrm{Tx}}, \qquad n \in [N]. \tag{14}$$

Remark 1: In the edge inference, data privacy is another main concern for high-stake applications, such as smart vehicles and drones. MUs in these applications may be reluctant to send their raw data to APs. To avoid the exposure of raw data, a hierarchical distributed structure has been studied in the literature, such as [33], by determining a partition point of a DNN model and deploying the partitioned model across the mobile device and the edge-computing-enabled AP. The data privacy is protected since only the output of the layers before the partition point is uploaded to APs. Note that our proposed framework is also applicable to the privacy-preserving hierarchical distributed structure. In this case, the input d_k becomes the locally computed output of the layers before the partition point. The computation task ϕ_k becomes the task of computing the inference result with the layers after the partition point.

To achieve the robustness of QoS against CSI errors, we shall collect D independent and identically distributed (i.i.d.) samples of the imperfect CSI as the data set $\mathcal{D} = \{\tilde{\boldsymbol{h}}^{(1)}, \ldots, \tilde{\boldsymbol{h}}^{(D)}\}$ to learn the uncertainty model of CSI before providing edge inference service. Based on the data set $\mathcal{D}$, we aim to design a beamforming vector $\boldsymbol{v}$ such that the safe condition is satisfied with probability at least $1-\zeta$. However, since we do not know the prior distribution of random errors, the statistical guarantee of a given approach is usually expressed as certain confidence level $1-\delta$ for certain tolerance level $1-\epsilon$, e.g., the SG approach [21]. That is, the confidence level of

$$\Pr(\mathrm{SINR}_k(\boldsymbol{v}; \boldsymbol{h}_k) \geq \gamma_k) \geq 1 - \epsilon \tag{15}$$

is no less than $1-\delta$ for some $\boldsymbol{v}$, D, $0<\epsilon<1$, and $0<\delta<1$. Thus, the violation probability of the safe condition is upper bounded by

$$\Pr(\mathrm{SINR}_k(\boldsymbol{v};\boldsymbol{h}_k)<\gamma_k)<\delta+\epsilon(1-\delta). \tag{16}$$

By setting ϵ and δ such that $\zeta > \delta+\epsilon(1-\delta)$, the safe condition (12) is guaranteed to be met.

We consider the block fading channel where the channel distribution is assumed invariant [34] within T_s blocks and the channel coefficient vector remains unchanged within each block. Note that the training by collecting D channel samples within each block will result in high signaling overhead. We will show that our proposed approach for addressing the probabilistic-QoS constraints can be integrated with a cost-effective channel sampling strategy in Section III-D.

E. Problem Analysis

Directly solving the joint chance constraints (13) is usually a highly intractable task [21], especially when there is no exact knowledge about the uncertainty. In this article, we shall propose a general framework for edge inference without assuming the prior distribution of random errors. A natural idea is to find a computationally tractable approximation for the probabilistic-QoS constraints (13).

1) Scenario Generation: SG [21] is a well-known approach by obtaining D independent samples of the random channel coefficient vector $\boldsymbol{h}$ and imposing the target QoS constraints $\mathrm{SINR}_k \geq \gamma_k, k \in [K]$ for each sample. However, because it ensures robustness in the minimax sense, it is too conservative when a large number of samples are drawn, since the volume of the feasible region will decrease, which may result in the infeasibility of problem $\mathscr{P}_{\mathrm{CCP}}$. In addition, the sample size D should be chosen such that $\sum_{i=1}^{NKL-1}\binom{D}{i}\epsilon^i(1-\epsilon)^{D-i}\leq\delta$, where $1-\delta$ gives the confidence level for the probabilistic-QoS constraints defined in (12). Therefore, the SG approach has a scalability issue since the required minimum sample size D increases roughly linearly with $1/\epsilon$ for small ϵ and also with NKL.

2) Stochastic Programming: To address this overconservativeness issue of the SG approach, a stochastic programming approach is further provided in [19] by finding a DC approximation for the chance constraints. The resulting DC-constrained stochastic program can be solved by successive convex approximation with the Monte Carlo approach at each iteration. However, its computation cost grows linearly with the number of samples D which is not scalable for obtaining high-robustness solutions, and the statistical guarantee is not available for the joint chance constraints under finite sample size.

To address the limitations of the existing works, we shall present a robust optimization approach in Section III to approximate the chance constraint via a statistical learning approach [22]. This approach enjoys the main advantages that the minimum required number of observations is only $\log\delta/\log(1-\epsilon)$ and the computational cost is independent of the sample size.

III. Learning-Based Robust Optimization Approximation for Joint Chance Constraints

In this section, we provide a robust optimization approximation for the joint chance constraints in problem $\mathscr{P}_{\mathrm{CCP}}$, followed by a statistical learning approach to learn the shape and size of the high probability region.

A. Approximating Joint Chance Constraints via Robust Optimization

Robust optimization [22] uses safe approximation and imposes that the safe conditions are always satisfied when the random variables lie in some geometric set. Specifically, the robust optimization approximation of the joint chance constraints (13) is given by

$$\mathrm{SINR}_k(\boldsymbol{v};\boldsymbol{h}_k)\geq\gamma_k, \boldsymbol{h}_k\in\mathcal{U}_k \quad \forall k\in[K] \tag{17}$$

where $\mathcal{U}_k$ is the high probability region that $\boldsymbol{h}_k$ lies in. The robust optimization approximation for the joint chance constraints should yield a solution such that the probabilistic-QoS constraint is satisfied with high confidence. The robust optimization approximation approach is realized by constructing a high probability region $\mathcal{U}_k$ from the data set $\mathcal{D}$ such that $\mathcal{U}_k$ covers a $1-\epsilon$ content of $\boldsymbol{h}_k$, i.e.,

$$\Pr(\boldsymbol{h}_k\in\mathcal{U}_k)\geq 1-\epsilon \tag{18}$$

with confidence level at least $1-\delta$. More precisely, since $\mathcal{U}_k$ is generated from data and therefore is random, we require that the proportion of time (18) is satisfied to be at least $1-\delta$ in the repeated application of the data generation and high probability region construction procedure. By imposing the QoS constraints for element in the high probability region as presented in (17), the confidence level for the probabilistic-QoS constraints (15) will be at least $1-\delta$. We thus obtain the robust optimization approximation for problem $\mathscr{P}_{\mathrm{CCP}}$ as

$$\begin{aligned}\mathscr{P}_{\mathrm{RO}}:\ \underset{\boldsymbol{v},\boldsymbol{h}}{\text{minimize}}\ & \sum_{n,k}\frac{1}{\eta_n}\|\boldsymbol{v}_{nk}\|_2^2+\sum_{n,l}P_{nk}^{\mathrm{c}}I_{(n,k)\in\mathcal{T}(\boldsymbol{v})}\\ \text{subject to}\ & \mathrm{SINR}_k(\boldsymbol{v};\boldsymbol{h}_k)\geq\gamma_k, \qquad \boldsymbol{h}_k\in\mathcal{U}_k;\, k\in[K]\\ & \sum_{k=1}^{K}\|\boldsymbol{v}_{nk}\|_2^2\leq P_n^{\mathrm{Tx}}, \qquad n\in[N].\end{aligned} \tag{19}$$

The choice of the geometric shape of the uncertainty set $\mathcal{U}_k$ is critical to the performance and the tracatability of the robust optimization approximation. Motivated by the tractability of robust optimization, ellipsoids and polytopes are commonly chosen as the basic uncertainty sets. The uncertainty set can be further augmented as the unions or intersection of these basic sets. In this article, we choose ellipsoidal uncertainty set to model the uncertainty of each group of channel coefficient vector $\boldsymbol{h}_k$ for its wide use in modeling CSI uncertainties [26], [35], as well as its tractability as shown in Section III-C. The high probability region $\mathcal{U}_k$ is parameterized as

$$\mathcal{U}_k=\left\{\boldsymbol{h}_k:\boldsymbol{h}_k=\hat{\boldsymbol{h}}_k+\boldsymbol{B}_k\boldsymbol{u}_k, \boldsymbol{u}_k^{\mathsf{H}}\boldsymbol{u}_k\leq 1\right\}. \tag{20}$$

Here, the parameters $\boldsymbol{B}_k\in\mathbb{C}^{NL\times NL}$ and $\hat{\boldsymbol{h}}_k\in\mathbb{C}^{NL}$ shall be learned from the data set $\mathcal{D}$, which will be presented in

Section III-B. We will then present the tractable reformulation of the robust optimization counterpart problem $\mathscr{P}_{\mathrm{RO}}$ in Section III-C.

B. Learning the High Probability Region From Data Samples

Note that (17) only gives a feasibility guarantee for the joint chance constraints with statistical confidence of at least $1-\delta$, but its conservativeness is still a challenging problem. Generally speaking, problem $\mathscr{P}_{\mathrm{RO}}$ is a less conservative approximation for problem $\mathscr{P}_{\mathrm{CCP}}$ if it has a larger feasible region. Therefore, we prefer a smaller volume of the high probability region $\mathcal{U}$ which provides a larger feasible region. In our problem formulation, we set the volume of the high probability region such that the statistical confidence for the probabilistic-QoS constraints is close to $1-\delta$.

In this article, we propose to use a statistical learning approach [22] for the parameters of the high probability region $\mathcal{U}$ which consists of a shape learning procedure and a size calibration procedure via the quantile estimation. First, we split the samples in data set $\mathcal{D}$ into two parts, i.e., $\mathcal{D}^1 = \{\tilde{\boldsymbol{h}}^{(1)}, \ldots, \tilde{\boldsymbol{h}}^{(D_1)}\}$ and $\mathcal{D}^2 = \{\tilde{\boldsymbol{h}}^{(D_1+1)}, \ldots, \tilde{\boldsymbol{h}}^{(D)}\}$, each for one procedure.

1) Shape Learning: Each ellipsoid set $\mathcal{U}_k$ can be reparameterized as

$$\mathcal{U}_k = \left\{ \boldsymbol{h}_k : \left(\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k\right) \le s_k \right\} \tag{21}$$

where $\hat{\boldsymbol{h}}_k$ and $\boldsymbol{\Sigma}_k$ are shape parameters of the ellipsoid $\mathcal{U}_k$, $s_k > 0$ determines its size, and $\boldsymbol{\Sigma}_k / s_k = \boldsymbol{B}_k \boldsymbol{B}_k^{\mathsf{H}}$. Suppose the observations of $\boldsymbol{h}_k$ are given by $\mathcal{D}_k = \mathcal{D}_k^1 \cup \mathcal{D}_k^2 = \{\tilde{\boldsymbol{h}}_k^{(j)}\}_{j=1}^{D}$. The shape parameter $\hat{\boldsymbol{h}}_k$ can be chosen as the sample mean, i.e.,

$$\hat{\boldsymbol{h}}_k = \frac{1}{D_1} \sum_{j=1}^{D_1} \tilde{\boldsymbol{h}}_k^{(j)}. \tag{22}$$

To reduce the complexity of the ellipsoid, we omit the correlation between each $\{\boldsymbol{h}_{kn}\}$ and choose $\boldsymbol{\Sigma}_k$ as the block diagonal matrix where each diagonal element is the sample covariance of the first part of the data set for $\boldsymbol{h}_{kn}$, i.e.,

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{k1} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_{kN} \end{bmatrix}$$

where

$$\boldsymbol{\Sigma}_{kn} = \frac{1}{D_1 - 1} \sum_{j=1}^{D_1} \left(\tilde{\boldsymbol{h}}_{kn}^{(j)} - \hat{\boldsymbol{h}}_{kn}\right)\left(\tilde{\boldsymbol{h}}_{kn}^{(j)} - \hat{\boldsymbol{h}}\boldsymbol{n}_k\right)^{\mathsf{H}}. \tag{23}$$

2) Size Calibration via Quantile Estimation: We then use the second part of data set $\mathcal{D}_k^2$ for calibrating the ellipsoid size s_k. The key idea is to estimate a $1-\epsilon$ quantile with $1-\delta$ confidence of a transformation of the data samples in $\mathcal{D}_k^2$. Let

$$\mathcal{G}(\xi) = \left(\xi - \hat{\boldsymbol{h}}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\xi - \hat{\boldsymbol{h}}_k\right) \tag{24}$$

be the map from the random space that $\boldsymbol{h}_k$ lies into $\mathbb{R}$. The size parameter s_k will be chosen as an estimated $(1-\epsilon)$-quantile of the underlying distribution of $\mathcal{G}(\xi)$ based on the data samples in $\mathcal{D}_{nk}^2$, where the $(1-\epsilon)$-quantile $q_{1-\epsilon}$ is defined from

$$\Pr(\mathcal{G}(\xi) \le q_{1-\epsilon}) = 1 - \epsilon. \tag{25}$$

Specifically, by computing the function values of $\mathcal{G}$ on each sample of $\mathcal{D}_k^2$, we can obtain the observations $G_1, \ldots, G_{D-D_1}$ where $G_j = \mathcal{G}(\boldsymbol{h}_k^{(D_1+j)})$. Then, the $t^\star$th value of the ranked observations $G_{(1)} \le \cdots \le G_{(D-D_1)}$ in ascending order, denoted as $G_{(j^\star)}$, can be an upper bound of the $(1-\epsilon)$-quantile of the underlying distribution of $G(\xi)$ based on the following proposition.

Proposition 1: s_k is an upper bound of the $(1-\epsilon)$-quantile of the underlying distribution with $1-\delta$ confidence, i.e.,

$$\Pr(s_k \ge q_{1-\epsilon}) \ge 1 - \delta \tag{26}$$

if s_k is set as

$$s_k = G_{(j^\star)}, \quad \text{where } j^\star \text{ is given by}$$
$$\min_{1 \le j \le D-D_1} \left\{ j : \sum_{k=0}^{j-1} \binom{D-D_1}{k} (1-\epsilon)^k \epsilon^{D-D_1-k} \ge 1 - \delta \right\}. \tag{27}$$

Proof: According to the definition of the quantile $q_{1-\epsilon}$, we have

$$\begin{aligned} &\Pr\left(G_{(j)} \ge q_{1-\epsilon}\right) \\ &= \Pr\left(G_{(k)} < q_{1-\epsilon}, k = 0, \ldots, j-1\right) \\ &= \sum_{k=0}^{j-1} \binom{D-D_1}{k} (1-\epsilon)^k \epsilon^{D-D_1-k}. \end{aligned} \tag{28}$$

Therefore, $G_{(j^\star)}$ is the smallest one among all upper bounds of the $(1-\epsilon)$-quantile of the underlying distribution with $1-\delta$ confidence. ■

Using the presented two procedures, we learn a high probability region $\mathcal{U}$ of the random channel coefficient vector h_ks. The statistical guarantee of this statistical learning-based robust optimization approximation approach is given by the following proposition.

Proposition 2: Suppose the data samples in the data set D_k are i.i.d. and chosen from a continuous distribution for any k. The data set is split into two independent parts $\mathcal{D}_k^1$ and $\mathcal{D}_k^2$. Each uncertainty set is chosen as $\mathcal{U}_k = \{\boldsymbol{h}_k : (\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k) \le s_k\}$. Their parameters $\hat{\boldsymbol{h}}_k$, $\boldsymbol{\Sigma}_k$, and s_k are determined following (22), (23), and (27), respectively. Thus, any feasible solution to problem $\mathscr{P}_{RO}$ guarantees that the probabilistic-QoS constraints (15) are satisfied with confidence at least $1-\delta$.

Proof: Since $\mathcal{G}$ depends only on $\mathcal{D}_k^1$, we have

$$\Pr_{\mathcal{D}_k^2}(\boldsymbol{v} \in \mathcal{V}) = \Pr_{\mathcal{D}_k^2}\left(G_{(t^\star)} \ge q_{1-\epsilon}\right) \ge 1 - \delta. \tag{29}$$

Therefore, it is readily obtained that $\Pr(\mathrm{SINR}_k \ge \gamma_k) \ge 1-\epsilon$ satisfies with confidence at least $1-\delta$. ■

Note that $j^\star$ exists only if

$$\sum_{k=0}^{D-D_1-1} \binom{D-D_1}{k} (1-\epsilon)^k \epsilon^{D-D_1-k} \ge 1 - \delta \tag{30}$$

Algorithm 1: Statistical Learning-Based Approach for the High Probability Region $\mathcal{U}_k$

Input: the data set $\mathcal{D} = \{\tilde{\boldsymbol{h}}^{(1)}, \cdots, \tilde{\boldsymbol{h}}^{(D)}\}$.
for *each* $k = 1, \cdots K$ **do**
  Data splitting: Randomly split the samples of $\boldsymbol{h}_k$, namely $\mathcal{D}_k$, into two parts $\mathcal{D}_k^1$ and $\mathcal{D}_k^2$.
  Shape learning: Set the shape parameters $\hat{\boldsymbol{h}}_k$ and $\boldsymbol{\Sigma}_k$ as (22) and (23) based on $\mathcal{D}_k^1$.
  Size calibration: Set the size parameter s_k as $G_{(j^\star)}$ by computing the values of function $\mathcal{G}$ on $\mathcal{D}_k^2$, where $\mathcal{G}$ is defined in (24) and $j^\star$ is chosen as (27).
  Compute $\boldsymbol{B}_k = \sqrt{s_k}\boldsymbol{\Delta}_k$ through Cholesky decomposition $\boldsymbol{\Sigma}_k = \boldsymbol{\Delta}_k\boldsymbol{\Delta}_k^{\mathsf{H}}$.
end
Output: $\hat{\boldsymbol{h}}_k$, $\boldsymbol{B}_k$ for all k.

which implies that $1-(1-\epsilon)^{D-D_1} \geq 1-\delta$. In other words, the required minimum number of samples is $D > D - D_1 \geq \log\delta/\log(1-\epsilon)$ to achieve the $1-\delta$ confidence of the probabilistic-QoS constraint (13). Matrix $\boldsymbol{B}_k$ can be computed as

$$\boldsymbol{B}_k = \sqrt{s_k}\boldsymbol{\Delta}_k \tag{31}$$

where $\boldsymbol{\Delta}_k$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_k$, i.e., $\boldsymbol{\Sigma}_k = \boldsymbol{\Delta}_k\boldsymbol{\Delta}_k^{\mathsf{H}}$. We summarize the whole procedure for learning the high probability region $\mathcal{U}$ from data set $\mathcal{D}$ in Algorithm 1.

C. Tractable Reformulations for Robust Optimization Problem

According to the ellipsoidal uncertainty model (20), the robust optimization approximation (17) can be rewritten as

$$\boldsymbol{h}_k^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{h}_k \geq \sigma_k^2 \tag{32}$$

$$\boldsymbol{h}_k = \hat{\boldsymbol{h}}_k + \boldsymbol{B}_k\boldsymbol{u}_k, \qquad \boldsymbol{u}_k^{\mathsf{H}}\boldsymbol{u}_k \leq 1 \tag{33}$$

where $\boldsymbol{u}_{nk} \in \mathbb{C}^L$. By defining matrices

$$\boldsymbol{H}_k = \begin{bmatrix}\hat{\boldsymbol{h}}_k & \boldsymbol{B}_k\end{bmatrix} \in \mathbb{C}^{NL\times(NL+1)} \tag{34}$$

and using the S-procedure [36], we obtain the following equivalent tractable reformulation for (32) and (33):

$$\boldsymbol{H}_k^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{H}_k \succeq \boldsymbol{Q}_k \tag{35}$$

$$\lambda_k \geq 0 \tag{36}$$

where $\boldsymbol{\lambda} = [\lambda_k] \in \mathbb{R}_+^K$ and $\boldsymbol{Q}_k$ is given by

$$\boldsymbol{Q}_k = \begin{bmatrix}\lambda_k + \sigma_k^2 & \boldsymbol{0} \\ \boldsymbol{0} & -\lambda_k\boldsymbol{I}_{NL}\end{bmatrix} \in \mathbb{C}^{(NL+1)\times(NL+1)}. \tag{37}$$

The derivation details of (35) and (36) from (32) and (33) are relegated to the Appendix.

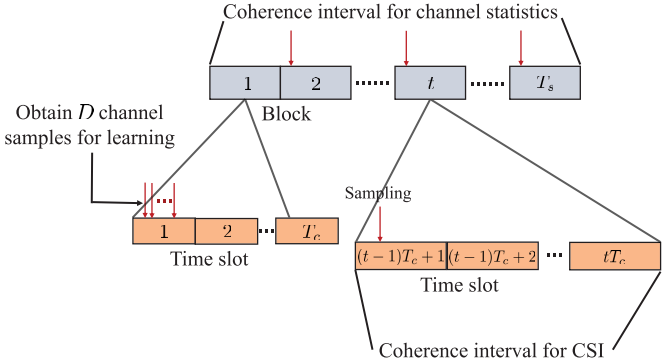


Fig. 4. Timeline of a cost-effective channel sampling strategy.

Thus, the proposed robust optimization approximation for problem $\mathscr{P}_{\text{CCP}}$ is given by the following group sparse beamforming problem with nonconvex quadratic constraints:

$$\mathscr{P}_{\text{RGS}}: \underset{\boldsymbol{v}\in\mathbb{C}^{NKL},\boldsymbol{\lambda}\in\mathbb{R}^K}{\text{minimize}} \sum_{n,l}\frac{1}{\eta_n}\|\boldsymbol{v}_{nl}\|_2^2 + \sum_{n,l}P_{nl}^{\text{c}}I_{(n,l)\in\mathcal{T}(\boldsymbol{v})}$$

$$\text{subject to} \quad (35), \lambda_k \geq 0 \quad \forall k \in [K] \tag{38}$$

$$\sum_{l=1}^K \|\boldsymbol{v}_{nl}\|_2^2 \leq P_n^{\text{Tx}} \quad \forall n \in [N]. \tag{39}$$

Its computational complexity of solving problem $\mathscr{P}_{\text{RGS}}$ is independent of the sample size D. An effective approach for obtaining approximate solution of nonconvex quadratic-constrained quadratic program is to lift the aggregative beamforming vector as a rank-one PSD matrix $\boldsymbol{V} = \boldsymbol{v}\boldsymbol{v}^{\mathsf{H}}$ and simply drop the rank-one constraint, which is termed as the SDR technique [23]. The obtained solution however may be infeasible for the original nonconvex quadratic constraints. To induce the group sparsity with nonconvex quadratic constraints, a quadratic variational form of weighted mixed ℓ_1/ℓ_2-norm is adopted in [26]. In this article, we will adopt an iteratively reweighted minimization approach which has demonstrated its effectiveness in cloud radio access network [27], [28] to further enhance the group sparsity of the aggregative beamforming vector. In addition, to improve the feasibility for the nonconvex quadratic constraint for each subproblem of the reweighted approach, we shall provide a novel DC approach for inducing rank-one solution. It should be mentioned that the uplink–downlink duality is not applicable to efficiently address the robust QoS constraints (35) due to the CSI uncertainty.

D. Integrating the Robust Optimization Approximation With a Cost-Effective Sampling Strategy

Consider the block fading channel where the channel distribution is assumed invariant [34] within the *coherence interval for channel statistics*. The coherence interval for channel statistics consists of T_s blocks, where each block is called a *coherence interval for CSI* and the channel coefficient vector remains unchanged within each block. However, collecting D channel samples within each block leads to high signaling overhead. To address this issue, we provide a cost-effective sampling strategy for enabling robust transmission, whose timeline is illustrated in Fig. 4.

At the beginning of the coherence interval for channel statistics, we collect D i.i.d. channel samples as $\mathcal{D}$. Based on the data set $\mathcal{D}$, we can learn the estimated channel coefficient vector $\hat{\boldsymbol{h}}_k$ from (22) and the estimated high probability region of the error $\boldsymbol{e}_k$ as $\boldsymbol{B}_k$ from (31). For the transmission in the first block, we can obtain $\{\boldsymbol{H}_k\}$ by combining these two parts following (34) and solve the resulting problem $\mathscr{P}_{\mathrm{RGS}}$. For any other block $t > 1$, we can obtain the estimated channel coefficient $\hat{\boldsymbol{h}}[t]$ as the sample mean by collecting as few as one sample of the channel coefficient vector. By replacing the estimated channel coefficient $\hat{\boldsymbol{h}}$ and keeping the error information $\{\boldsymbol{B}_k : k \in [K]\}$, we can construct the parameter $\{\boldsymbol{H}_k[t]\}$ at the tth block as

$$\boldsymbol{H}_k[t] = \begin{bmatrix} \hat{\boldsymbol{h}}_k[t] & \boldsymbol{B}_k \end{bmatrix} \quad \forall k \in [K] \tag{40}$$

and design the transmitter beamformer by solving problem $\mathscr{P}_{\mathrm{RGS}}(\{\boldsymbol{H}_k[t]\})$, which significantly reduces the signaling overhead for channel sampling. The effectiveness of this cost-effective scheme will be demonstrated in Section V-A numerically.

IV. Reweighted Power Minimization for Group Sparse Beamforming With Nonconvex Quadratic Constraints

This section presents a reweighted power minimization approach to induce the group sparsity structure for problem $\mathscr{P}_{\mathrm{RGS}}$. We further demonstrate that the nonconvex quadratic constraints can be reformulated as convex constraints with respect to a rank-one PSD matrix using a matrix lifting technique, followed by proposing a DC approach to induce rank-one solutions.

A. Matrix Lifting for Nonconvex Quadratic Constraints

We observe that constraints (35) are convex with respect to $\boldsymbol{v}\boldsymbol{v}^{\mathsf{H}}$ despite its nonconvexity with respect to $\boldsymbol{v}$. This motivates us to adopt the matrix lifting technique [23] to address the nonconvex quadratic constraints in problem $\mathscr{P}_{\mathrm{RGS}}$ by denoting

$$\boldsymbol{V}_{ij}[s,t] = \boldsymbol{v}_{si}\boldsymbol{v}_{tj}^{\mathsf{H}} \in \mathbb{C}^{L\times L} \tag{41}$$

$$\boldsymbol{V}_{ij} = \begin{bmatrix} \boldsymbol{V}_{ij}[1,1] & \cdots & \boldsymbol{V}_{ij}[1,N] \\ \vdots & \ddots & \vdots \\ \boldsymbol{V}_{ij}[N,1] & \cdots & \boldsymbol{V}_{ij}[N,N] \end{bmatrix} = \boldsymbol{v}_i\boldsymbol{v}_j^{\mathsf{H}} \in \mathbb{C}^{NL\times NL} \tag{42}$$

$$\boldsymbol{V} = \boldsymbol{v}\boldsymbol{v}^{\mathsf{H}} = \begin{bmatrix} \boldsymbol{V}_{11} & \cdots & \boldsymbol{V}_{1K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{V}_{K1} & \cdots & \boldsymbol{V}_{KK} \end{bmatrix} \in \mathbb{S}_+^{NKL} \tag{43}$$

where $\mathbb{S}_+^{NKL}$ denotes the set of Hermitian PSD matrices. The aggregative beamforming vector $\boldsymbol{v}$ is thus lifted as a rank-one PSD matrix $\boldsymbol{V}$. The constraint $\mathcal{C}_k$ of problem $\mathscr{P}_{\mathrm{RGS}}$, which given by (35), can be equivalently rewritten as the following PSD constraint:

$$\boldsymbol{H}_k^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{V}_{kk} - \sum_{l\neq k}\boldsymbol{V}_{ll}\right)\boldsymbol{H}_k \succeq \boldsymbol{Q}_k \tag{44}$$

and the transmit power constraint (39) can be equivalently rewritten as

$$\sum_{l=1}^{K}\|\boldsymbol{v}_{nl}\|_2^2 = \sum_{l=1}^{K}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \le P_n^{\mathrm{Tx}} \quad \forall n = 1,\ldots,N. \tag{45}$$

Therefore, using the matrix lifting technique, we obtain an equivalent reformulation for problem $\mathscr{P}_{\mathrm{RGS}}$ as

$$\begin{aligned} \mathscr{P}\colon \underset{\boldsymbol{V},\boldsymbol{\lambda}}{\text{minimize}} \quad & \sum_{n,l}\left(\frac{1}{\eta_n}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) + P_{nl}^{\mathrm{c}} I_{\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n])\neq 0}\right) \\ \text{subject to} \quad & (44),\ \lambda_k \ge 0 \quad \forall k \in [K] && (46) \\ & \sum_{l=1}^{K}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \le P_n^{\mathrm{Tx}} \quad \forall n \in [N] && (47) \\ & \boldsymbol{V} \succeq \boldsymbol{0}, \mathrm{rank}(\boldsymbol{V}) = 1. && (48) \end{aligned}$$

Note that the constraints are still nonconvex due to the nonconvexity of the rank-one constraint.

B. DC Representations for Rank-One Constraint

For a PSD matrix $\boldsymbol{V} \in \mathbb{S}_+^{NKL}$, its rank is one if and only if it has only one nonzero singular value, i.e.,

$$\sigma_i(\boldsymbol{V}) = 0, \qquad i = 2,\ldots,NKL \tag{49}$$

where $\sigma_i(\boldsymbol{V})$ is the ith largest singular value of $\boldsymbol{V}$. The trace norm and spectral norm of the PSD matrix $\boldsymbol{V}$ are, respectively, given as

$$\mathrm{Tr}(\boldsymbol{V}) = \sum_{i=1}^{NKL}\sigma_i(\boldsymbol{V}), \qquad \|\boldsymbol{V}\| = \sigma_1(\boldsymbol{V}). \tag{50}$$

Thus, we obtain an equivalent DC representation for the rank-one constraint of $\boldsymbol{V}$

$$\mathcal{R}(\boldsymbol{V}) = \mathrm{Tr}(\boldsymbol{V}) - \|\boldsymbol{V}\| = 0. \tag{51}$$

$\mathcal{R}$ is a DC function of $\boldsymbol{V}$ since both the trace norm and the spectral norm are convex.

C. Reweighted ℓ_1 Minimization for Inducing Group Sparsity

The reweighted ℓ_1 minimization approach has shown its advantages in enhancing group sparsity for improving the energy efficiency of cloud radio access networks [27], [28]. ℓ_1-norm is a well-recognized convex surrogate for the ℓ_0-norm. In order to further enhance the sparsity, reweighted ℓ_1 minimization is proposed to iteratively minimize a weighted ℓ_1-norm and update the weights. For the objective function of problem $\mathscr{P}$, we observe that the indicator function $I_{\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n])\neq 0}$ can be interpreted as the ℓ_0-norm of $\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n])$. We can thus use the reweighed ℓ_1 minimization technique via approximating $I_{\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n])\neq 0}$ by $w_{nl}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n])$, which consists of alternatively minimizing the approximated objective function and updating the weight as

$$w_{nl} = \frac{c}{\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) + \tau} \tag{52}$$

where $\tau > 0$ is a constant regularization factor and $c > 0$ is a constant. If $\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n])$ is small, the reweighted

ℓ_1 minimization approach will put larger weight on the transceiver pair (n, l) which prompts that the inference task l is not preferred to be executed at the nth edge node.

D. Proposed Reweighted Power Minimization Approach

In this section, we provide a reweighted power minimization approach by combining the matrix lifting, DC representation, and reweighted ℓ_1 minimization techniques. In the jth step, we shall update $\boldsymbol{V}^{[j+1]}$ via solving

$$\begin{aligned} \underset{\boldsymbol{V},\boldsymbol{\lambda}}{\text{minimize}} \quad & \sum_{n,l}\left(\frac{1}{\eta_n}+w_{nl}^{[j]}P_{nl}^{\mathrm{c}}\right)\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \\ \text{subject to} \quad & (44),\ \lambda_k \geq 0 \quad \forall k \in [K] \\ & \sum_{l=1}^{K}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\mathrm{Tx}} \quad \forall n \in [N] \\ & \boldsymbol{V} \succeq \boldsymbol{0},\ \mathrm{rank}(\boldsymbol{V}) = 1 \end{aligned} \tag{53}$$

and the weights $\{w_{nl}^{[j]}\}$ are updated following (52) which are initialized as 1 at the beginning.

To solve problem (53) with nonconvex rank-one constraint, we propose to use the DC representation (51) by solving the following DC program:

$$\begin{aligned} \mathscr{P}_{\mathrm{DC}}: \ \underset{\boldsymbol{V},\boldsymbol{\lambda}}{\text{minimize}} \quad & \sum_{n,l}\left(\frac{1}{\eta_n}+w_{nl}^{[j]}P_{nl}^{\mathrm{c}}\right)\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) + \mu\mathcal{R}(\boldsymbol{V}) \\ \text{subject to} \quad & (44),\ \lambda_k \geq 0 \quad \forall k \in [K] \\ & \sum_{l=1}^{K}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\mathrm{Tx}} \quad \forall n \in [N] \\ & \boldsymbol{V} \succeq \boldsymbol{0} \end{aligned} \tag{54}$$

where $\mu > 0$ is the regularization parameter. Despite the nonconvexity of the DC problem, problem $\mathscr{P}_{\mathrm{DC}}$ can be efficiently solved by the simplified DC algorithm, i.e., iteratively linearizing the concave part [37]. At the tth iteration, we shall solve

$$\begin{aligned} \underset{\boldsymbol{V},\boldsymbol{\lambda}}{\text{minimize}} \quad & \sum_{n,l}\left(\frac{1}{\eta_n}+w_{nl}^{[j]}P_{nl}^{\mathrm{c}}\right)\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \\ & \quad +\mu\Big(\mathrm{Tr}(\boldsymbol{V}) - \mathrm{Tr}\Big(G^{(t)}\boldsymbol{V}\Big)\Big) \\ \text{subject to} \quad & (44),\ \lambda_k \geq 0 \quad \forall k \in [K] \\ & \sum_{l=1}^{K}\mathrm{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\mathrm{Tx}} \quad \forall n \in [N] \\ & \boldsymbol{V} \succeq \boldsymbol{0} \end{aligned} \tag{55}$$

where $G^{(t)}$ is one subgradient of spectral norm at $\boldsymbol{V}^{(t)}$. It can be computed as $\partial\|\boldsymbol{V}\|_2 = \boldsymbol{u}_1\boldsymbol{u}_1^{\mathsf{H}}$ where $\boldsymbol{u}_1$ is the eigenvector corresponding to the largest eigenvalue of matrix $\boldsymbol{V}$. This DC algorithm guarantees converging to a stationary point of problem $\mathscr{P}_{\mathrm{DC}}$ from arbitrary initial points [37].

When the reweighted ℓ_1 minimization algorithm converges at a rank-one solution $\boldsymbol{V}^{[j]}$, we can extract the aggregative beamforming vector $\boldsymbol{v}^\star$ from the Choleskey decomposition $\boldsymbol{V}^{[j]} = \boldsymbol{v}^\star\boldsymbol{v}^{\star\mathsf{H}}$. The whole procedure of the proposed reweighted power minimization approach is summarized in Algorithm 2.

Algorithm 2: Proposed Reweighted Power Minimization Approach for Problem $\mathscr{P}$

Initialization: $\boldsymbol{V}^{[0]}$, w_{nl}.
while *not converge* **do**
  $\boldsymbol{V}^{(0)} \leftarrow \boldsymbol{V}^{[j]}$
  while *not converge* **do**
    update $\boldsymbol{V}^{(t)}$ as the solution to problem (55)
  end
  $\boldsymbol{V}^{[j+1]} \leftarrow \boldsymbol{V}^{(t)}$
  update the weights $\{w_{nl}^{[j+1]}\}$ according to (52)
end
obtain $\boldsymbol{v}^\star$ through Choleskey decomposition $\boldsymbol{V}^{[j]} = \boldsymbol{v}^\star\boldsymbol{v}^{\star\mathsf{H}}$.
Output: $\boldsymbol{v}^\star$.

V. Numerical Results

In this section, we provide numerical experiments for comparing the proposed framework with other state-of-the-art approaches. We generate the edge inference system with $N = 4$ APs located at $(\pm 400, \pm 400)$ m and $K = 4$ MUs randomly located in the $[-800\ 800]\times[-800\ 800]$ m^2 region. Each AP is equipped with $L = 2$ antennas. The imperfection model of the channel coefficient vector between the nth AP and the kth MU is chosen as $\boldsymbol{h}_{kn} = 10^{-L(d_{kn})/20}(\boldsymbol{c}_{kn}+\boldsymbol{e}_{kn})$. The path-loss model is given by $L(d_{kn}) = 128.1 + 37.6\log_{10} d_{kn}$, the Rayleigh small-scale fading coefficient is given by $\boldsymbol{c}_{kn} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and the additive error is given by $\boldsymbol{e}_{kn} \sim \mathcal{CN}(\boldsymbol{0}, 10^{-4}\boldsymbol{I})$. As presented in Section III-B, D_1 determines the accuracy of the learned shape of the uncertainty set, while D_2 determines the accuracy of the calibrated size of the uncertainty set. To balance these two points, the collected D independent samples of $\boldsymbol{h}_{kn}$s are split evenly for learning the shape and size of the uncertainty ellipsoids, respectively, i.e., $D_1 = D_2 = D/2$. For each AP, the power amplifier efficiency is chosen as $\eta_1 = \cdots = \eta_N = 1/4$, the average maximum transmit power is chosen as $P_1^{\mathrm{Tx}} = \cdots = P_N^{\mathrm{Tx}} = 1$W, and the computation power consumption for each task ϕ_k at the nth AP is chosen as $P_{nk}^{\mathrm{c}} = 0.60$W. We set the target SINR as $\gamma_1 = \cdots = \gamma_K = \gamma$, the tolerance level as $\epsilon = 0.05$, and the confidence level as $\delta = 0.05$. The regularization parameters τ are set as 10^{-6} and μ is set as 10.

A. Benefits of Taking CSI Uncertainty Into Consideration

In this article, we consider the CSI uncertainty in channel sampling and propose to solve it with a learning-based robust optimization approximation approach. To further reduce the channel sampling overhead, we provide a cost-effective sampling strategy in Secion III-D. We now evaluate its advantages over the beamformer design without taking the CSI error into consideration by supposing that each task is performed at all APs. Specifically, we collect $D = 200$ i.i.d. channel samples in the training phase within one coherent interval for CSI. In the test phase, we only collect one channel sample $\boldsymbol{h}^{(1)}$, construct

TABLE II
NUMBER OF TESTS THAT QOS IS MET

User Index	1	2	3	4
Proposed Approach	39946	39946	39946	39946
Without considering uncertainty	15205	15123	15197	15214

$\boldsymbol{H}_k$'s following (40) and solve the problem:

$$\begin{aligned} \underset{\boldsymbol{V},\boldsymbol{\lambda}}{\text{minimize}} \quad & \sum_{n,l}\left(\frac{1}{\eta_n}\text{Tr}(\boldsymbol{V}_{ll}[n,n]) + P_{nl}^{\text{c}}\right) \\ \text{subject to} \quad & (44), \lambda_k \geq 0 \quad \forall k \in [K] \\ & \sum_{l=1}^{K} \text{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\text{Tx}} \quad \forall n \in [N] \\ & \boldsymbol{V} \succeq \boldsymbol{0}. \end{aligned} \tag{56}$$

As comparison, the beamforming design without taking uncertainty into consideration is given by solving the problem

$$\begin{aligned} \underset{\boldsymbol{V},\boldsymbol{\lambda}}{\text{minimize}} \quad & \sum_{n,l}\left(\frac{1}{\eta_n}\text{Tr}(\boldsymbol{V}_{ll}[n,n]) + P_{nl}^{\text{c}}\right) \\ \text{subject to} \quad & \boldsymbol{h}_k^{(1)\text{H}}\left(\frac{1}{\gamma_k}\boldsymbol{V}_{kk} - \sum_{l\neq k}\boldsymbol{V}_{ll}\right)\boldsymbol{h}_k^{(1)} \geq \sigma_k^2 \quad \forall k \\ & \sum_{l=1}^{K} \text{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\text{Tx}} \quad \forall n \\ & \boldsymbol{V} \succeq \boldsymbol{0}. \end{aligned} \tag{57}$$

Note that we use SDR for both approaches for fairness. We compare two approaches by generating 40 000 realizations of i.i.d. channel samples for testing, and regenerate the training data set for the proposed approach every 200 realizations. We compute the achieved SINR for each mobile device with the solution to each approach, i.e., $\text{SINR}_k(\boldsymbol{v}; \tilde{\boldsymbol{h}})$, where $\tilde{\boldsymbol{h}}$ is the true channel coefficient vector, and calculate the number of realizations that the target QoS for each device is met, i.e., $\text{SINR}_k \geq \gamma_k$. The results shown in Table II demonstrate that the proposed robust approximation approach has considerably improved the robustness of QoS against CSI errors by a cost-effective sampling approach.

B. Overcoming the Overconservativeness of Scenario Generation

As we point out in Section II-E, the SG approach is overconservative since it imposes that the target QoS constraints are satisfied for all samples, which would lead to a smaller feasible region. Here, we use numerical experiments to demonstrate the advantage of the presented robust optimization approximation approach in overcoming the overconservativeness. Consider the feasibility problem of the robust optimization approximation approach given by

$$\begin{aligned} \text{find} \quad & \boldsymbol{V}, \boldsymbol{\lambda} \\ \text{subject to} \quad & (44), \lambda_k \geq 0 \quad \forall k \in [K] \\ & \sum_{l=1}^{K} \text{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\text{Tx}} \quad \forall n \in [N] \\ & \boldsymbol{V} \succeq \boldsymbol{0} \end{aligned} \tag{58}$$

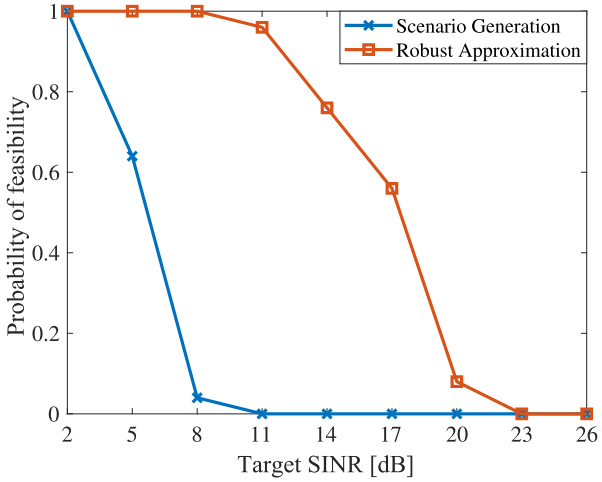


Fig. 5. Probability of feasibility using SG and the robust optimization approximation approach over the target SINR γ.

and the feasibility problem of the scenario approach given by

$$\begin{aligned} \text{find} \quad & \boldsymbol{V} \\ \text{subject to} \quad & \boldsymbol{h}_k^{(i)\text{H}}\left(\frac{1}{\gamma_k}\boldsymbol{V}_{kk} - \sum_{l\neq k}\boldsymbol{V}_{ll}\right)\boldsymbol{h}_k^{(i)} \geq \sigma_k^2 \quad \forall k, i \\ & \sum_{l=1}^{K} \text{Tr}(\boldsymbol{V}_{ll}[n,n]) \leq P_n^{\text{Tx}} \quad \forall n \\ & \boldsymbol{V} \succeq \boldsymbol{0}. \end{aligned} \tag{59}$$

Note that we adopt the SDR technique in both approach for purpose of fairness. We collect $D = 200$ i.i.d. channel samples for each realization, run both algorithms for 25 random realizations of the data set, and compare the probability of yielding feasible solutions using the SG approach and the presented robust optimization approximation approach. The results in Fig. 5 reveal that the statistical learning-based robust approximation considerably improves the probability of feasibility compared with the SG approach though we only obtain sufficient conditions for the robust optimization counterpart using S-procedure in Section III-C.

C. Convergence Behavior

By choosing the reweighting parameter as $c = 1/\ln(1+\tau^{-1})$, the proposed reweighted power minimization approach, i.e., Algorithm 2, essentially approximates the ℓ_0-norm according to $I_{x\neq 0} = \|x\|_0 = \lim_{\tau\to 0}\ln(1+x\tau^{-1})/\ln(1+\tau^{-1})$, and minimizes the approximated objective function

$$f(\boldsymbol{V}) = \sum_{n,l}\left(\frac{1}{\eta_n}\text{Tr}(\boldsymbol{V}_{ll}[n,n]) + P_{nl}^{\text{c}}\frac{\ln\left(1+\tau^{-1}\text{Tr}(\boldsymbol{V}_{ll}[n,n])\right)}{\ln\left(1+\tau^{-1}\right)}\right) + \mu\mathcal{R}(\boldsymbol{V}) \tag{60}$$

under constraints (46) and (47) using a majorization–minimization (MM) technique as stated in [27]. Fig. 6 illustrates the convergence behavior of the proposed

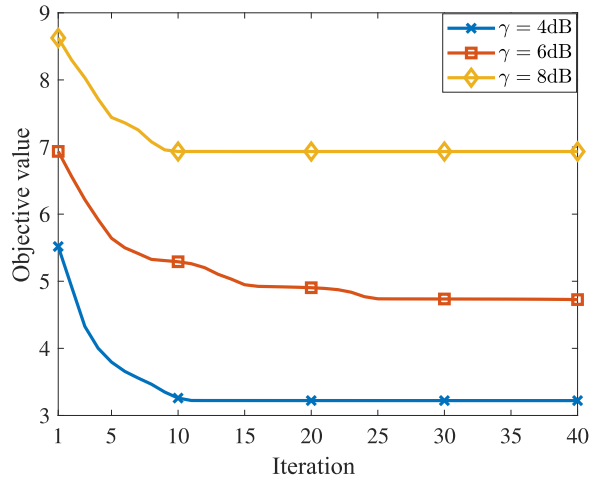


Fig. 6. Convergence behavior of the proposed reweighted power minimization approach with different target SINR γ.

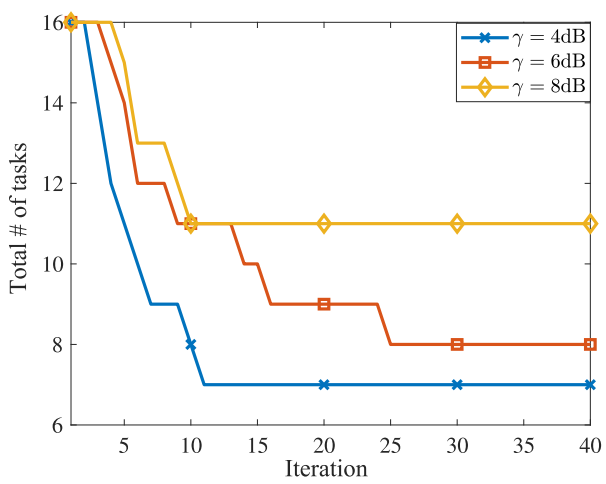


Fig. 7. Trajectories of the total number of inference tasks performed at all edge computing nodes with different target SINR γ.

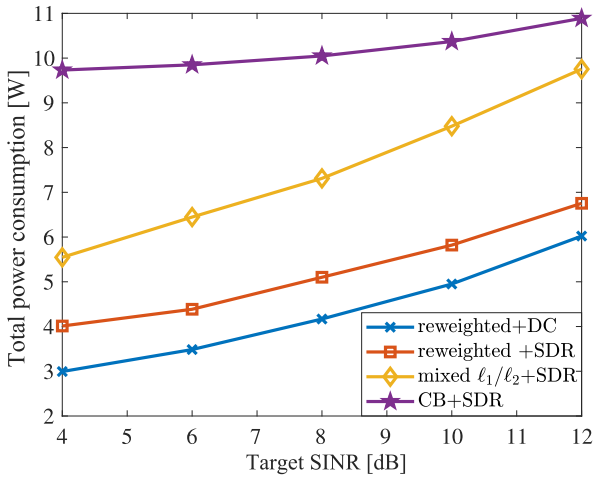


Fig. 8. Total power consumption over target SINR.

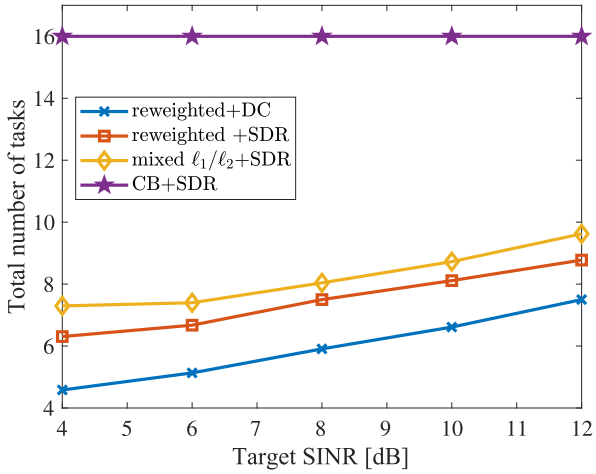


Fig. 9. Total # of tasks performed at APs over target SINR.

reweighted power minimization approach in terms of the objective function f by collecting $D = 200$ channel samples. We also plot the corresponding trajectories of the group sparsity of the aggregative beamforming vector $\boldsymbol{v}$ in Fig. 7, i.e., total number of inference tasks performed at all edge computing nodes. We observe that the number of tasks to be performed at edge computing nodes increases with a greater value of target QoS γ, which leads to higher total power consumption of the edge inference system.

D. Total Power Consumptions Over Target SINR

We then conduct numerical results to compare the performance of different algorithms for problem $\mathscr{P}$ with $D = 200$ i.i.d. channel samples, including the proposed reweighted power minimization approach called "reweighted+DC" and other state-of-the-art algorithms listed as follows.

1) *Mixed ℓ_1/ℓ_2+SDR:* This algorithm is proposed in [26], which adopts the quadratic variational form of the weighted mixed ℓ_1/ℓ_2-norm for inducing group sparsity and SDR to address the nonconvex quadratic constraints.
2) *Reweighted+SDR:* To improve the energy efficiency of downlink transmission in cloud-RAN, we adopt the iteratively reweighted minimization algorithm [27] for inducing the group sparsity and SDR [23] for the nonconvex quadratic constraints.
3) *CB+SDR:* We assume that all tasks are performed at each AP and conduct coordinated beamforming for minimizing the transmission power consumption under probabilistic-QoS constraints.

We also set $c = 1/\ln(1+\tau^{-1})$ as stated in Section V-C. The performances of all algorithms averaged over 100 channel realizations are illustrated in Figs. 8 and 9. Fig. 8 presents the total power consumption of each algorithm and demonstrates that the proposed DC algorithm yields lower total power consumption than other approaches, which is owed to its better capability to induce group sparsity as shown in Fig. 9. Note that the total number of tasks for the "CB+SDR" algorithm is always $KN = 16$.

Through all numerical results, we have seen considerable advantages of the presented statistical learning-based robust

optimization approximation and the proposed reweighted power minimization algorithm in providing energy-efficient processing and robust transmission service for edge inference.

VI. Conclusion

In this article, we presented an energy-efficient processing and robust cooperative transmission framework for executing deep learning inference tasks for mobile devices. Specifically, we proposed to minimize the sum of computation power and transmission power consumption under the probabilistic-QoS constraints via adaptive task selection and coordinated beamforming design. The joint chance constraints therein were further addressed by a statistical learning-based robust optimization approximation approach. This yielded a group sparse beamforming problem with nonconvex quadratic constraints. Then, we developed a reweighted power minimization approach by iteratively solving a DC-regularized reweighted ℓ_1 minimization problem and updating the weights, thereby tackling both the group sparse objective function and nonconvex quadratic constraints. The numerical results demonstrated that the proposed approach achieved the lowest total power consumption among other state-of-the-art algorithms, and avoided the drawbacks of other methods for joint chance-constrained programs.

There are still some open problems to be studied.

1) This article considers the architecture that each inference task is performed at multiple base stations separately. An interesting problem is to consider the hierarchical distributed structure of DNNs over the cloud and the edge [38].
2) In this article, we consider a basic ellipsoid model for each uncertain channel coefficient vector. It is interesting to use a data-driven approach with a more complicated model of the high probability region to reduce its volume, such as clustering the data samples and using a union of ellipsoids as the high probability region.
3) It is still an open problem to provide the theoretical guarantee of the proposed reweighted power minimization algorithm since the conditions for convergence guarantee of reweighted approach in [27] and [39] are not met.

Appendix
Derivation of (35) Using S-Procedure

We first rewrite (33) as

$$\boldsymbol{h}_k\tau_k = \hat{\boldsymbol{h}}_k\tau_k + \boldsymbol{B}_k\tilde{\boldsymbol{u}}_k, \tilde{\boldsymbol{u}}_k^{\mathsf{H}}\tilde{\boldsymbol{u}}_k \le \tau_k^2 \tag{61}$$

where $\boldsymbol{u}_k = \tilde{\boldsymbol{u}}_k/\tau_k \in \mathbb{C}^L$, $\tau_k > 0$. Let

$$\boldsymbol{x}_k = \begin{bmatrix}\tau_k^{\mathsf{H}} & \tilde{\boldsymbol{u}}_k^{\mathsf{H}}\end{bmatrix}^{\mathsf{H}} \in \mathbb{C}^{NL+1} \tag{62}$$

we can obtain that

$$\boldsymbol{h}_k\tau_k = \boldsymbol{H}_k\boldsymbol{x}_k. \tag{63}$$

Thus, we know

$$\boldsymbol{h}_k^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{h}_k - \sigma_k^2 \ge 0 \tag{64}$$

$$\Leftrightarrow (\boldsymbol{h}_k\tau_k)^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{h}_k\tau_k - \sigma_k^2\tau_k^2 \ge 0 \tag{65}$$

$$\Leftrightarrow (\boldsymbol{H}_k\boldsymbol{x}_k)^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{H}_k\boldsymbol{x}_k - \sigma_k^2\tau_k^2 \ge 0 \tag{66}$$

$$\Leftrightarrow \boldsymbol{x}_k^{\mathsf{H}}\boldsymbol{P}_k^0\boldsymbol{x}_k \ge 0 \tag{67}$$

where $\boldsymbol{P}_k^0 \in \mathbb{S}^{NL+1}$ is given by

$$\boldsymbol{H}_k^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{H}_k - \begin{bmatrix}\sigma_k^2 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}. \tag{68}$$

Likewise, $\tilde{\boldsymbol{u}}_k^{\mathsf{H}}\tilde{\boldsymbol{u}}_k \le \tau_k^2$, can be rewritten as

$$\boldsymbol{x}_k^{\mathsf{H}}\boldsymbol{P}_k^1\boldsymbol{x}_k \ge 0 \tag{69}$$

where $\boldsymbol{P}_k^1 \in \mathbb{S}^{NL+1}$ is given by

$$\boldsymbol{P}_k^1 = \begin{bmatrix}1 & \\ & -\boldsymbol{I}_N\end{bmatrix}. \tag{70}$$

Thus, we shall use the S-procedure

$$\boldsymbol{x}_k^{\mathsf{H}}\boldsymbol{P}_k^1\boldsymbol{x}_k \ge 0 \Longrightarrow \boldsymbol{x}_k^{\mathsf{H}}\boldsymbol{P}_k^0\boldsymbol{x}_k \ge 0 \tag{71}$$

which is given by

$$\boldsymbol{P}_k^0 \ge \lambda_k\boldsymbol{P}_k^1, \lambda_k \ge 0. \tag{72}$$

Therefore, we obtain the tractable reformulation for the joint chance constraints (13) as

$$\boldsymbol{H}_k^{\mathsf{H}}\left(\frac{1}{\gamma_k}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{H}} - \sum_{l\neq k}\boldsymbol{v}_l\boldsymbol{v}_l^{\mathsf{H}}\right)\boldsymbol{H}_k \succeq \boldsymbol{Q}_k \tag{73}$$

where $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K] = [\lambda_{nk}] \in \mathbb{R}_+^{N\times K}$ and $\boldsymbol{Q}_k$ is given by

$$\boldsymbol{Q}_k = \begin{bmatrix}\lambda_k + \sigma_k^2 & \\ & -\lambda_k\boldsymbol{I}_{NL}\end{bmatrix} \in \mathbb{C}^{(NL+1)\times(NL+1)}. \tag{74}$$

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[3] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," 2020. [Online]. Available: arXiv:2002.09668.

[5] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[6] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[7] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Commun.*, early access, doi: 10.1109/MWC.001.1900119.

[8] X. Xu *et al.*, "Scaling for edge inference of deep neural networks," *Nature Electron.*, vol. 1, no. 4, pp. 216–222, 2018.

[9] C. Xu, J. Ren, L. She, Y. Zhang, Z. Qin, and K. Ren, "Edgesanitizer: Locally differentially private deep inference at the edge for mobile data analytics," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5140–5151, Jun. 2019.

[10] *Stroke of Genius: GauGAN Turns Doodles Into Stunning, Photorealistic Landscapes*. Accessed: Mar. 18, 2019. [Online]. Available: https://blogs.nvidia.com/blog/2019/03/18/gaugan-photorealistic-landscapes-nvidia-research//

[11] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[12] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[13] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, Jan. 2018.

[14] T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5687–5695.

[15] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[16] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[17] F. Yang, P. Cai, H. Qian, and X. Luo, "Pilot contamination in massive MIMO induced by timing and frequency errors," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4477–4492, Jul. 2018.

[18] J. Mo and R. W. Heath, "Limited feedback in single and multi-user MIMO systems with finite-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3284–3297, May 2018.

[19] 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.

[20] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418–4431, Jul. 2012.

[21] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM J. Optim.*, vol. 17, no. 4, pp. 969–996, 2006.

[22] L. J. Hong, Z. Huang, and H. Lam, "Learning-based robust optimization: Procedures and statistical guarantees," 2017. [Online]. Available: arXiv:1704.04342.

[23] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation bounds for quadratic optimization with homogeneous quadratic constraints," *SIAM J. Optim.*, vol. 18, no. 1, pp. 1–28, 2007.

[24] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.

[25] M. Hong, R. Sun, H. Baligh, and Z. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.

[26] Y. Shi, J. Zhang, and K. B. Letaief, "Robust group sparse beamforming for multicast green cloud-RAN with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4647–4659, Sep. 2015.

[27] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[28] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed L_p-minimization for green cloud-RAN with user admission control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1022–1036, Apr. 2016.

[29] *CNN Energy Estimation Website*. Accessed: Mar. 27, 2017. [Online]. Available: http://energyestimation.mit.edu

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[31] B. Liu, F. Zhou, G. Lu, and R. Q. Hu, "Energy efficient and robust beamforming for MISO cognitive small cell networks," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5002–5014, Dec. 2018.

[32] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874–2885, Dec. 2017.

[33] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.

[34] A. Liu, X. Chen, W. Yu, V. K. N. Lau, and M. Zhao, "Two-timescale hybrid compression and forward for massive MIMO aided C-RAN," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2484–2498, May 2019.

[35] M. F. Hanif, L.-N. Tran, A. Tölli, M. Juntti, and S. Glisic, "Efficient solutions for weighted sum rate maximization in multicellular networks with channel uncertainties," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5659–5674, Nov. 2013.

[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

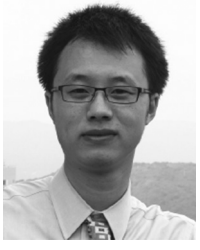
[37] P. D. Tao and L. T. H. An, "Convex analysis approach to DC programming: Theory, algorithms and applications," *Acta Math. Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[38] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Atlanta, GA, USA, 2017, pp. 328–339.

[39] H. Wang, F. Zhang, Q. Wu, Y. Hu, and Y. Shi, "Nonconvex and nonsmooth sparse optimization via adaptively iterative reweighted methods," 2018. [Online]. Available: arXiv:1810.10167.

Kai Yang (Student Member, IEEE) received the B.S. degree in electronic engineering from Dalian University of Technology, Dalian, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, also with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, and also with the University of Chinese Academy of Sciences, Beijing, China.

His research interests include big data processing, mobile edge artificial intelligence, and federated machine learning.



Yuanming Shi (Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in electronic and computer engineering from Hong Kong University of Science and Technology, Hong Kong, in 2015.

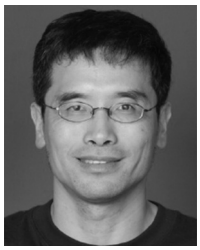
Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, where he is currently a Tenured Associate Professor. He visited the University of California at Berkeley, Berkeley, CA, USA, from October 2016 to February 2017. His research areas include optimization, statistics, machine learning, signal processing, and their applications to 6G, IoT, AI, and FinTech.

Dr. Shi was a recipient of the 2016 IEEE Marconi Prize Paper Award in Wireless Communications and the 2016 Young Author Best Paper Award by the IEEE Signal Processing Society. He is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Wei Yu (Fellow, IEEE) received the B.A.Sc. degree in computer engineering and mathematics from the University of Waterloo, Waterloo, ON, Canada, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1998 and 2002, respectively.

Since 2002, he has been with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, Canada, where he is currently a Professor and holds a Canada Research Chair (Tier 1) of information theory and wireless communications. His main research interests include information theory, optimization, wireless communications, and broadband access networks.

Prof. Yu was a recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2019, the IEEE Communications Society Award for Advances in Communication in 2019; the IEEE Signal Processing Society Best Paper Award in 2017 and 2008; the *Journal of Communications and Networks* Best Paper Award in 2017; the IEEE Communications Society Best Tutorial Paper Award in 2015; the IEEE International Conference on Communications Best Paper Award in 2013; the McCharles Prize for Early Career Research Distinction in 2008; the Early Career Teaching Award from the Faculty of Applied Science and Engineering, University of Toronto in 2007; and the Early Researcher Award from Ontario in 2006. He currently serves as the First Vice President of the IEEE Information Theory Society in 2020. He received the Steacie Memorial Fellowship in 2015. He is currently an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2017 to 2020, and in the past served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2010 to 2013, as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2009 to 2011, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2004 to 2007. He served as the Chair for the Signal Processing for Communications and Networking Technical Committee of the IEEE Signal Processing Society from 2017 to 2018. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2016. He is a Fellow of the Canadian Academy of Engineering, and a member of the College of New Scholars, Artists, and Scientists of the Royal Society of Canada.



Zhi Ding (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 1990.

He is a Professor of electrical and computer engineering with the University of California at Davis, Davis, CA, USA. From 1990 to 2000, he was a Faculty Member with Auburn University, Auburn, AL, USA, and later, University of Iowa, Iowa City, IA, USA. He has held visiting positions with Australian National University, Canberra, ACT, Australia; Hong Kong University of Science and Technology, Hong Kong; NASA Lewis Research Center, Cleveland, OH, USA; and USAF Wright Laboratory,Wright-Patterson AFB, OH, USA. He has active collaboration with researchers from Australia, Canada, China, Finland, Hong Kong, Japan, South Korea, Singapore, and Taiwan. He has coauthored the book *Modern Digital and Analog Communication Systems* (5th ed., Oxford University Press, 2019).

Prof. Ding was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1994 to 1997 and from 2001 to 2004, and IEEE SIGNAL PROCESSING LETTERS from 2002 to 2005. He was a member of technical committee on Statistical Signal and Array Processing and a member of technical committee on Signal Processing for Communications from 1994 to 2003. He was the General Chair of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing and the Technical Program Chair of 2006 IEEE GLOBECOM. He was also an IEEE Distinguished Lecturer (Circuits and Systems Society from 2004 to 2006 and Communications Society from 2008 to 2009). He served on the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS Steering Committee from 2007 to 2009 and as its Chair from 2009 to 2010. He has been an Active Member of IEEE, serving on technical programs of several workshops and conferences.