Reconfigurable Intelligent Surface for Green Edge Inference

Sheng Hua, *Graduate Student Member, IEEE*, Yong Zhou, *Member, IEEE*, Kai Yang, *Member, IEEE*, Yuanming Shi, *Senior Member, IEEE*, and Kunlun Wang, *Member, IEEE*

***Abstract*—Reconfigurable intelligent surface (RIS) as an emerging cost-effective technology can enhance the spectral- and energy-efficiency of wireless networks. In this article, we consider an RIS-aided green edge inference system, where the inference tasks generated from resource-constrained mobile devices (MDs) are uploaded to and cooperatively performed at multiple resource-enhanced base stations (BSs). Taking into account both the computation and uplink/downlink transmit power consumption, we formulate an overall network power consumption minimization problem, which calls for the joint design of the set of tasks performed by each BS, uplink/downlink beamforming vectors of BSs, transmit power of MDs, and uplink/downlink phase-shift matrices at the RIS. However, the resulting combinatorial optimization problem is nonconvex and highly intractable. We tackle the challenge of combinatorial variables by exploiting the group sparsity structure of the beamforming vectors. Moreover, a block-structured optimization with mixed $\ell_{1,2}$-norm and difference-of-convex-functions (DC) based three-stage framework is proposed to solve the problem, where the mixed $\ell_{1,2}$-norm and DC techniques are adopted to induce the group sparsity structure and handle the nonconvex rank-one constraint, respectively. Simulations demonstrate the supreme performance gain of deploying an RIS and confirm the effectiveness of the proposed algorithm over the baseline algorithms in reducing the overall network power consumption.**

***Index Terms*—Reconfigurable intelligent surface, joint uplink and downlink, green edge inference, block-structured optimization, difference-of-convex programming.**

Manuscript received October 13, 2020; revised January 16, 2021; accepted February 6, 2021. Date of publication February 11, 2021; date of current version May 20, 2021. The work of Yong Zhou was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62001294 and Grant 61971286. The work of Kunlun Wang was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61932014. This article was presented in part at the IEEE Global Communications Conference Workshops, Waikoloa, HI, USA, December 2019. The editor coordinating the review of this article was J. Xu. *(Corresponding author: Yong Zhou.)*

Sheng Hua 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: huasheng@shanghaitech.edu.cn).

Yong Zhou and Yuanming Shi are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: zhouyong@shanghaitech.edu.cn; shiym@shanghaitech.edu.cn).

Kai Yang is with JD Technology Group, Beijing 100176, China (email: yangkai@shanghaitech.edu.cn).

Kunlun Wang is with the School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China (e-mail: klwang@cee.ecnu.edu.cn).

Digital Object Identifier 10.1109/TGCN.2021.3058657

I. Introduction

BENEFITING from the availability of big data, recent years have witnessed the prosperity of deep neural network (DNN), which is a branch of artificial intelligence (AI) techniques and has demonstrated its superiority in a variety of intelligent applications (e.g., computer vision and natural language processing). Thanks to its strong representation ability, DNN has been adopted to promote the development of wireless communications and bring convenience to the system implementation in various aspects such as interference management [2]–[3]. On the other hand, wireless communications can promote AI as well, by analyzing the rich data generated by the ever-increasing number of mobile devices (MDs) and thus providing intelligent services for the MDs. With the MDs at the network edge generating 77 exabytes data per month by 2022 [4], the demand of performing inference tasks (e.g., object recognition and machine translation) is anticipated to be ubiquitous, especially in the next generation AI-powered wireless networks [5]. Driven by this trend, it is urgent to push traditional cloud-based DNN models to the network edge so as to unleash the potentials of edge data and in turn provide intelligent services [6]. One possible architecture to perform inference tasks is on-device inference, i.e., running DNN models directly on MDs. While the model compression [7], model selection [8], and hardware acceleration [9] are proposed as promising techniques to help devices run small-sized DNN models, deploying powerful models with millions of parameters is still challenging because of both the memory and battery limitations [10].

By leveraging edge computing [11] and deploying DNN models at the edge base stations (BSs) that have strong computational capacity and large storage resources, edge inference stands out as a promising paradigm to provide intelligent services for MDs [12]. To accomplish the inference tasks, the MDs upload the task-specific data to the BSs and subsequently the BSs deliver the inference results after finishing the inference process. Tailored for latency-critical applications, the authors in [13] and [14] respectively proposed device-edge and edge-cloud synergy frameworks to partition DNN model parameters based on network dynamics to minimize the execution latency. As energy efficiency is a key performance indicator for edge inference systems, the authors in [15] and [16] proposed energy-aware approaches to prune DNN models to minimize the computation power consumption (i.e., power required for the BSs to perform the inference tasks) while maintaining reasonable inference precision. However, the communication power consumption was not considered in [15] and [16]. The authors in [17]–[18]

proposed to minimize the sum of computation and downlink transmit power consumption (i.e., power required for the BSs to deliver inference results to the MDs), while the uplink transmit power (i.e., power required for the MDs to upload data to the BSs) was neglected. However, in edge inference systems, the traffic load in the uplink (e.g., raw images for an object recognition task) is usually comparable to that in the downlink (e.g., labeled images), resulting in high uplink transmit power consumption. Therefore, it is imperative to develop new techniques to reduce both the uplink and downlink transmit power consumption and in turn facilitate an energy-efficient design for edge inference systems.

Recently, a growing line of works focused on an emerging technology named reconfigurable intelligent surface (RIS) [19], which has the potential to significantly reduce the power consumption [1], [20] and improve the energy efficiency [21]. RIS is also envisioned as a key enabler to provide broadband connectivity for the future 6G systems [22]. In particular, an RIS is a low-cost planar array consisting of a large number of passive reflecting elements with reconfigurable phase shifts, each of which can be dynamically tuned via a software controller to reflect the incident signals [23]–[24]. These elements consume negligible energy due to their passive nature. By adaptively adjusting the phase shifts of reflecting elements, an RIS can combine the constructive signals and suppress the interference, thereby greatly enhancing the performance of wireless systems [25]–[26]. By jointly optimizing the beamforming vectors at the BS and the phase-shift matrix at the RIS, deploying an RIS has the potential to reduce the power consumption in various applications, e.g., downlink unicast [20] and broadcast [27] settings, non-orthogonal multiple access [28], and simultaneous wireless information and power transfer [29]. The authors in [30] investigated RIS-aided over-the-air computation to minimize signal distortion. To reduce the complexity of dynamic phase configuration for the RIS, the authors in [31]–[32] proposed deep learning methods to learn the mapping between the locations of transceivers and the optimal phase shifts. In terms of power consumption, all the aforementioned works only considered the downlink transmit power consumed by the BSs. However, in edge inference systems, the computation power consumption is an indispensible component and should be taken into account to accurately characterize the overall network power consumption. In addition, it is essential to optimize both the uplink and downlink phase-shift matrices of the RIS to assist both the uplink and downlink data transmissions. These two key issues make the approaches proposed in the existing works not applicable to RIS-aided edge inference systems.

To guarantee the quality of intelligent services provided for MDs, computation replication [33] allows each inference task to be performed by multiple BSs and creates multiple copies of the inference results at different BSs. These copies enable cooperative downlink transmission among the BSs on delivering the inference results. In terms of the power consumption, however, cooperative transmission and computation replication conflict with each other. Specifically, cooperative transmission reduces the downlink transmit power consumption by exploiting a higher beamforming gain, while computation replication rapidly increases the computation power consumption by repeatedly running the same DNN model for multiple times. Therefore, it is necessary to strike a balance between the computation and communication power consumption via appropriately selecting inference tasks to be performed by each BS and in turn achieve green edge inference, which motivates this work.

In this article, we consider an RIS-aided green edge inference system with multiple BSs cooperatively performing inference tasks for multiple MDs, taking into account both the uplink and downlink transmit power consumption as well as the computation power consumption. Our objective is to minimize the overall network power consumption subject to prescribed quality-of-service (QoS) requirements, by jointly designing the task selection strategy, transmit/receive beamforming vectors of the BSs, the transmit power of the MDs, and the uplink/downlink phase-shift matrices at the RIS. However, the formulated problem is a combinatorial optimization problem with nonconvex constraints and is highly intractable. The main contributions of this article are summarized as follows

- We propose a joint design of the task selection strategy, transmit/receive beamforming vectors, transmit power, and uplink/downlink phase-shift matrices for an RIS-aided green edge inference system. To the best of our knowledge, this is the first attempt to unify beamforming vectors, transmit power, and phase shifts design in both the uplink and downlink transmission into a general edge inference framework.
- The combinatorial nature of the task selection strategy and the coupled optimization variables stand out as two major challenges. We address the challenge of the combinatorial variables by exploiting the group sparsity structure of the beamforming vectors, and tackle the challenge of the coupled variables by proposing a block-structured optimization (BSO) approach.
- With fixed phase shifts, we adopt the weighted mixed $\ell_{1,2}$-norm to induce the group sparsity of beamforming vectors. With fixed beamforming vectors and transmit power, the original problem is transformed to a homogeneous quadratically constrained quadratic programming (QCQP) with a nonconvex rank-one constraint. As the widely adopted semidefinite relaxation (SDR) technique incurs performance degradation when the number of reflecting elements is large, we propose a novel difference-of-convex-functions (DC) representation for this nonconvex constraint, followed by developing an effective DC algorithm. We then propose a BSO with mixed $\ell_{1,2}$-norm and DC based three-stage framework to solve the problem.
- Through extensive simulations, we show that the deployment of an RIS can significantly reduce the overall network power consumption of edge inference systems. Furthermore, the proposed BSO with mixed $\ell_{1,2}$-norm and DC algorithm achieves a significant performance improvement compared to the BSO with mixed $\ell_{1,2}$-norm and SDR algorithm, which demonstrates the effectiveness of DC in yielding the rank-one solutions.

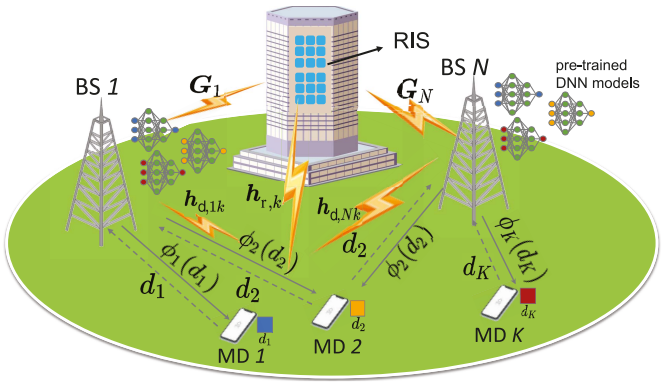


Fig. 1. RIS-aided edge inference system with N BSs collaboratively serving K MDs with the assistance of an RIS deployed on the facade of a building.

The remainder of this article is organized as follows. We present the system model and problem formulation in Section II. A BSO approach is developed in Section III to decouple optimization variables. We propose a three-stage framework in Section IV. Simulation results are illustrated in Section V. Finally, Section VI concludes this article.

Notations: We use boldface lower-case (e.g., $\boldsymbol{h}$) and upper-case letters (e.g., $\boldsymbol{G}$) to represent vectors and matrices, respectively. The transpose, conjugate transpose, trace operator and diagonal matrix are denoted as $(\cdot)^{\mathsf{T}}, (\cdot)^{\mathsf{H}}, \mathrm{Tr}(\cdot)$ and $\mathrm{diag}(\cdot)$, respectively. The symbols $|\cdot|$ and $\Re(\cdot)$ denote the modulus and the real component of a complex number. The $n \times n$ identity matrix is denoted as $\boldsymbol{I}_n$. The complex normal distribution is denoted as $\mathcal{CN}$. The inner product of two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is denoted as $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle$, which is defined as $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{Tr}(\boldsymbol{X}^{\mathsf{H}} \boldsymbol{Y})$. The ℓ_2-norm of a vector is denoted as $\|\cdot\|_2$. The spectral norm and Frobenius norm of a matrix are denoted as $\|\cdot\|$ and $\|\cdot\|_F$, respectively. The i-th largest singular value of matrix $\boldsymbol{X}$ is denoted as $\sigma_i(\boldsymbol{X})$. We use $\mathbf{1}_{\{\cdot\}}$ to denote the indicator function which outputs 1 if the condition $\cdot$ is satisfied, and outputs 0 otherwise. In the rest of this article, the superscripts UL and DL refer to uplink and downlink, respectively, and the letters d and r in the subscripts stand for the *direct link* and the *reflected link*, respectively.

II. System Model and Problem Formulation

In this section, we describe the system model and the power consumption model for performing inference tasks at the network edge, followed by formulating an overall network power consumption minimization problem for green edge inference systems.

A. System Model

We consider an RIS-aided edge inference system, where N L_n-antenna BSs distributed in a small-cell network collaboratively serve K single-antenna MDs with the assistance of an M-element RIS deployed on the facade of a building, as shown in Fig. 1. Let $\mathcal{N} = \{1, \ldots, N\}$, $\mathcal{K} = \{1, \ldots, K\}$, and $\mathcal{M} = \{1, \ldots, M\}$ denote the index sets of BSs, MDs, and reflecting elements, respectively. The BSs are resource-enhanced with strong computation and storage capabilities [34]. Each MD has an inference task (e.g., image recognition) to be processed by a task-specific DNN model (e.g., AlexNet [35]). Specifically, the DNN model denoted as ϕ_k takes MD k's local data d_k (e.g., raw images) as input and generates the inference result $\phi_k(d_k)$ (e.g., labeled images) as output. As it is impractical to run DNN models on resource-constrained MDs, we in this article propose to upload the inference tasks of the MDs to be performed at the BSs. We assume that all the BSs have downloaded the pre-trained DNN models from cloud servers in advance, therefore they can perform tasks for all the MDs [17].

The overall process of accomplishing the inference tasks in the edge inference system is composed of the following three phases.

- *Uplink Transmission:* The MDs upload the collected input data $\{d_k,\ k \in \mathcal{K}\}$ to the BSs.
- *Inference Computation:* The BSs feed data (e.g., d_k) into a specific pre-trained DNN model (e.g., ϕ_k) according to the task type and then obtain the inference results (e.g., $\phi_k(d_k)$).
- *Downlink Transmission:* The BSs deliver the inference results $\{\phi_k(d_k),\ k \in \mathcal{K}\}$ to the corresponding MDs.

By exploiting the broadcast nature of wireless channels, each MD's data can be successfully received by multiple BSs in the uplink, enabling computation replication and creating multiple copies of the inference results at different BSs [33]. In the downlink, the BSs performing the same inference task cooperatively transmit the inference results to the corresponding MD. By exploiting the existing channel estimation approaches [36], we assume that the global channel state information (CSI) is available at the BSs as in [27]–[29]. Let $\mathcal{A}_n \subseteq \mathcal{K}$ denote the set of MD indices whose inference tasks are selectively performed by BS n, and $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_N)$ denote the task selection strategy. We adopt the time-division duplex (TDD) mode to separate the uplink and downlink transmissions.

1) Uplink Transmission: Let $s_k^{\mathrm{UL}} \in \mathbb{C}$ denote the representative information symbol of input data d_k, and $p_k^{\mathrm{UL}} \in \mathbb{R}$ denote the transmit power of MD k. Without loss of generality, $\{s_k^{\mathrm{UL}}, k \in \mathcal{K}\}$ are assumed to have zero mean and unit power. The signal received at BS n can be expressed as

$$\boldsymbol{y}_n^{\mathrm{UL}} = \sum_{k \in \mathcal{K}} \boldsymbol{g}_{nk}^{\mathrm{UL}} \sqrt{p_k^{\mathrm{UL}}} s_k^{\mathrm{UL}} + \boldsymbol{z}_n^{\mathrm{UL}},\ \forall\, n \in \mathcal{N}, \tag{1}$$

where $\boldsymbol{g}_{nk}^{\mathrm{UL}} \in \mathbb{C}^{L_n \times 1}$ is the equivalent baseband channel response from MD k to BS n and $\boldsymbol{z}_n^{\mathrm{UL}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_n^2 \boldsymbol{I}_{L_n})$ is the additive white Gaussian noise (AWGN) at BS n with σ_n^2 being the noise power. With the deployment of an RIS, the equivalent baseband channel from MD k to BS n consists of both the direct link and the reflected link, where the reflected link is a concatenation of the MD-RIS link, the phase shifts at the RIS, and the RIS-BS link [20], [27]–[29]. Therefore, $\boldsymbol{g}_{nk}^{\mathrm{UL}}$ can be modeled as

$$\boldsymbol{g}_{nk}^{\mathrm{UL}} = \underbrace{\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{UL}}}_{\text{direct link}} + \underbrace{\left(\boldsymbol{G}_n^{\mathrm{UL}}\right)^{\mathsf{H}} \left(\boldsymbol{\Theta}^{\mathrm{UL}}\right)^{\mathsf{H}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{UL}}}_{\text{reflected link}}, \tag{2}$$

where $\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{UL}} \in \mathbb{C}^{L_n \times 1}$, $\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{UL}} \in \mathbb{C}^{M \times 1}$, and $\boldsymbol{G}_n^{\mathrm{UL}} \in \mathbb{C}^{M \times L_n}$ denote the channel responses from MD k to BS n, from MD k

to the RIS, and from the RIS to BS n, respectively. In addition, $\mathbf{\Theta}^{\mathrm{UL}} = \beta\,\mathrm{diag}(\theta_1^{\mathrm{UL}}, \ldots, \theta_M^{\mathrm{UL}}) \in \mathbb{C}^{M\times M}$ denotes the diagonal phase-shift matrix for uplink transmission, where $\beta \in [0,1]$ is the amplitude reflection coefficient and $\theta_m^{\mathrm{UL}} = e^{j\varphi_m^{\mathrm{UL}}}$ with $\varphi_m^{\mathrm{UL}} \in [0, 2\pi)$ being the uplink phase shift of the m-th reflecting element of the RIS. The reflected link only accounts for one-time reflection, because the power of signals reflected by two or more times is negligible due to the high path loss [20], [27]–[29].

We consider the linear beamforming strategy, and denote the receive beamforming vector of BS n to decode s_k^{UL} as $\boldsymbol{v}_{nk}^{\mathrm{UL}} \in \mathbb{C}^{L_n\times 1}$. BS n only decodes MD k's transmitted symbol s_k^{UL} if $k \in \mathcal{A}_n$. The estimated symbol at BS n for MD $k \in \mathcal{A}_n$, denoted by $\hat{s}_{nk}^{\mathrm{UL}} \in \mathbb{C}$, is given by

$$\hat{s}_{nk}^{\mathrm{UL}} = \left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}} \boldsymbol{y}_n^{\mathrm{UL}} = \left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}} \boldsymbol{g}_{nk}^{\mathrm{UL}} \sqrt{p_k^{\mathrm{UL}}} s_k^{\mathrm{UL}} + \left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}} \sum_{l\neq k} \boldsymbol{g}_{nl}^{\mathrm{UL}} \sqrt{p_l^{\mathrm{UL}}} s_l^{\mathrm{UL}} + \left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}} \boldsymbol{z}_n^{\mathrm{UL}}. \quad (3)$$

The uplink signal-to-interference-plus-noise ratio (SINR) observed at BS n for MD k is

$$\mathrm{SINR}_{nk}^{\mathrm{UL}} = \frac{p_k^{\mathrm{UL}} \left|\left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}} \boldsymbol{g}_{nk}^{\mathrm{UL}}\right|^2}{\sum_{l\neq k} p_l^{\mathrm{UL}} \left|\left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}} \boldsymbol{g}_{nl}^{\mathrm{UL}}\right|^2 + \sigma_n^2 \left\|\boldsymbol{v}_{nk}^{\mathrm{UL}}\right\|_2^2}, \quad \forall\, k \in \mathcal{A}_n,\ n \in \mathcal{N}. \quad (4)$$

2) Downlink Transmission: After performing the inference tasks, the BSs cooperatively transmit the inference results $\{\phi_k(d_k), k \in \mathcal{K}\}$ to the corresponding MDs through downlink wireless channels. Let $s_k^{\mathrm{DL}} \in \mathbb{C}$ denote the representative symbol of $\phi_k(d_k)$ intended for MD k and $\boldsymbol{v}_{nk}^{\mathrm{DL}}$ denote the downlink beamforming vector from BS n to MD k. Without loss of generality, $\{s_k^{\mathrm{DL}},\ k \in \mathcal{K}\}$ are assumed to have zero mean and unit power. The signal transmitted by BS n, denoted as $\boldsymbol{x}_n^{\mathrm{DL}} \in \mathbb{C}^{L_n\times 1}$, is a summation of beamformed symbols for MD $k \in \mathcal{A}_n$, i.e., $\boldsymbol{x}_n^{\mathrm{DL}} = \sum_{k\in\mathcal{A}_n} \boldsymbol{v}_{nk}^{\mathrm{DL}} s_k^{\mathrm{DL}},\ \forall\, n \in \mathcal{N}$. The signal received by MD k can be expressed as

$$\begin{aligned} y_k^{\mathrm{DL}} &= \sum_{n\in\mathcal{N}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{x}_n^{\mathrm{DL}} + z_k^{\mathrm{DL}} \\ &= \sum_{n\in\mathcal{N}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \times \left(\mathbf{1}_{\{k\in\mathcal{A}_n\}} \boldsymbol{v}_{nk}^{\mathrm{DL}} s_k^{\mathrm{DL}} + \sum_{l\in\mathcal{A}_n,\, l\neq k} \boldsymbol{v}_{nl}^{\mathrm{DL}} s_l^{\mathrm{DL}}\right) + z_k^{\mathrm{DL}} \\ &= \sum_{n\in\mathcal{N}} \mathbf{1}_{\{k\in\mathcal{A}_n\}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{v}_{nk}^{\mathrm{DL}} s_k^{\mathrm{DL}} + \sum_{l\neq k} \sum_{n\in\mathcal{N}} \mathbf{1}_{\{l\in\mathcal{A}_n\}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{v}_{nl}^{\mathrm{DL}} s_l^{\mathrm{DL}} + z_k^{\mathrm{DL}}, \\ &\quad \forall\, k \in \mathcal{K}, \end{aligned} \quad (5)$$

where $z_k^{\mathrm{DL}} \in \mathbb{C}$ is the AWGN at MD k with zero mean and power σ_k^2, and $(\boldsymbol{g}_{nk}^{\mathrm{DL}})^{\mathsf{H}} \in \mathbb{C}^{1\times L_n}$ is the equivalent downlink channel response from BS n to MD k. Similar to the uplink counterpart, $(\boldsymbol{g}_{nk}^{\mathrm{DL}})^{\mathsf{H}}$ can be modeled as

$$\left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} = \underbrace{\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{DL}}\right)^{\mathsf{H}}}_{\text{direct link}} + \underbrace{\left(\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{DL}}\right)^{\mathsf{H}} \mathbf{\Theta}^{\mathrm{DL}} \boldsymbol{G}_n^{\mathrm{DL}}}_{\text{reflected link}}, \quad (6)$$

where $(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{DL}})^{\mathsf{H}} \in \mathbb{C}^{1\times L_n}$, $(\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{DL}})^{\mathsf{H}} \in \mathbb{C}^{1\times M}$, and $\boldsymbol{G}_n^{\mathrm{DL}} \in \mathbb{C}^{M\times L_n}$ denote the channel responses from BS n to MD k, from the RIS to MD k, and from BS n to the RIS, respectively, and $\mathbf{\Theta}^{\mathrm{DL}} = \beta\,\mathrm{diag}(\theta_1^{\mathrm{DL}}, \ldots, \theta_M^{\mathrm{DL}}) \in \mathbb{C}^{M\times M}$ is the downlink phase-shift matrix with diagonal entries $\theta_m^{\mathrm{DL}} = e^{j\varphi_m^{\mathrm{DL}}}$ and $\varphi_m^{\mathrm{DL}} \in [0, 2\pi)$. Although channel reciprocity is often assumed to hold in TDD systems, we consider a general case, where the channel responses in the uplink and downlink can be different. Based on (5), the SINR observed by MD $k \in \mathcal{K}$ in the downlink is given by

$$\begin{aligned} \mathrm{SINR}_k^{\mathrm{DL}} &= \frac{\left|\sum_{n\in\mathcal{N}} \mathbf{1}_{\{k\in\mathcal{A}_n\}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{v}_{nk}^{\mathrm{DL}}\right|^2}{\sum_{l\neq k} \left|\sum_{n\in\mathcal{N}} \mathbf{1}_{\{l\in\mathcal{A}_n\}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{v}_{nl}^{\mathrm{DL}}\right|^2 + \sigma_k^2} \\ &= \frac{\left|\sum_{n\in\mathcal{N}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{v}_{nk}^{\mathrm{DL}}\right|^2}{\sum_{l\neq k} \left|\sum_{n\in\mathcal{N}} \left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}} \boldsymbol{v}_{nl}^{\mathrm{DL}}\right|^2 + \sigma_k^2}, \end{aligned} \quad (7)$$

where the second equality holds because BS n does not transmit data to MD k by setting $\boldsymbol{v}_{nk}^{\mathrm{DL}} = \mathbf{0}$ if $k \notin \mathcal{A}_n$.

B. Power Consumption Model

As running DNN models often incurs high energy consumption due to their high computational complexity [37]–[38] and energy-efficiency is one of the key performance indicators for green communications [39], we in this subsection present the power consumption model of the proposed edge inference system, taking into consideration both the computation power for inference and the communication power for uplink and downlink transmissions.

1) Computation Power Consumption: We denote the power consumption of performing MD k's inference task at BS n as P_{nk}^{c}. Therefore, the total computation power consumption at all BSs is given by $P_{\mathrm{comp}}(\mathcal{A}) = \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{A}_n} P_{nk}^{\mathsf{c}}$. It is worth noting that the majority of the computation power is consumed for running DNN models, which can be estimated by using the energy estimation methodology proposed in [37]. This methodology provides a layer-wise energy breakdown for arbitrary neural networks. In particular, the DNN configurations (e.g., number of filters, number of input feature maps) are taken as inputs and the normalized layer-wise consumptions energy of the neural network (i.e., normalized by the energy consumption per multiply-and-accumulation (MAC) operation) are generated as outputs [40]. The computation time can be calculated via dividing the number of MAC operations by the average throughput of a CPU chip. Therefore, the power consumption for performing an inference task equals to the total energy consumption divided by the corresponding computation time.

For example, the energy consumption of running AlexNet to process one image on a well-designed energy-efficient Eyeriss chip can be approximated by that of performing 4×10^9 MAC

operations, or equivalently 0.45 W when the chip is at core supply voltage 1.2 V [41]. As the typical value of computation power P_{nk}^{c} (e.g., 0.45 W) is comparable to the BSs' transmit power (e.g., 1 W [42]), it is necessary to take into account both the computation and transmit power to facilitate the energy-efficient design.

2) Communication Power Consumption: The communication power consumption consists of the power consumed by the MDs in the uplink transmission and by the BSs in the downlink transmission. According to (2)-(7), the total uplink transmit power consumption is $\sum_{k\in\mathcal{K}} p_k^{\mathrm{UL}}$, while the downlink transmit power consumption of BS n is given by $\mathbf{E}[\sum_{k\in\mathcal{A}_n}\|\boldsymbol{v}_{nk}^{\mathrm{DL}}s_k^{\mathrm{DL}}\|_2^2] = \sum_{k\in\mathcal{A}_n}\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\|_2^2$, where $\mathbf{E}[\,\cdot\,]$ denotes the expectation. Therefore, the total communication power consumption for both uplink and downlink transmissions is given by $P_{\mathrm{comm}}(\mathcal{A},\{p_k^{\mathrm{UL}}\},\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}) = \sum_{k\in\mathcal{K}} p_k^{\mathrm{UL}} + \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{A}_n}\frac{1}{\eta_n}\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\|_2^2$, where η_n is the drain efficiency coefficient of the radio frequency power amplifier of BS n.

In summary, the overall network power consumption, consisting of both the computation and communication power consumption, can be expressed as

$$\begin{aligned} &P_{\mathrm{total}}\Big(\mathcal{A},\Big\{p_k^{\mathrm{UL}}\Big\},\Big\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\}\Big)\\ &\quad= P_{\mathrm{comm}}\Big(\mathcal{A},\Big\{p_k^{\mathrm{UL}}\Big\},\Big\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\}\Big) + P_{\mathrm{comp}}(\mathcal{A})\\ &\quad= \sum_{k\in\mathcal{K}} p_k^{\mathrm{UL}} + \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{A}_n}\frac{1}{\eta_n}\Big\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\|_2^2 + \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{A}_n} P_{nk}^{\mathsf{c}}. \end{aligned} \tag{8}$$

It is worth noting that the static power consumption of the BSs (proportional to N) and that of the RIS (proportional to M) are not included in P_{total} and regarded as constants, because they do not vary with the optimization variables and are neglectable compared to computation and communication power consumption.

C. Problem Formulation and Analysis

In the proposed edge inference system, there exists a fundamental tradeoff between the communication and computation power consumption. Specifically, with computation replication, more BSs performing the same task reduces the downlink transmit power consumption by exploiting a higher cooperative beamforming gain, at the cost of increasing the computation power consumption. Therefore, we propose to achieve green edge inference by minimizing the overall network power consumption via striking a good balance between the communication and computation power consumption.

Let $\{\gamma_k^{\mathrm{UL}}, k\in\mathcal{K}\}$ and $\{\gamma_k^{\mathrm{DL}}, k\in\mathcal{K}\}$ denote the SINR thresholds required to successfully receive the input data and inference results in the uplink and downlink, respectively. Given a task selection strategy $\mathcal{A}$, the network power consumption minimization problem is formulated as

$$\begin{aligned} \mathscr{P}(\mathcal{A}):\quad &\underset{\substack{\boldsymbol{\Theta}^{\mathrm{UL}},\boldsymbol{\Theta}^{\mathrm{DL}}\\ \{\boldsymbol{v}_{nk}^{\mathrm{UL}}\},\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\},\{p_k^{\mathrm{UL}}\}}}{\text{minimize}} \quad P_{\mathrm{total}}\Big(\mathcal{A},\Big\{p_k^{\mathrm{UL}}\Big\},\Big\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\}\Big)\\ \text{subject to}\quad &\mathrm{SINR}_k^{\mathrm{DL}} \ge \gamma_k^{\mathrm{DL}},\ \forall\, k\in\mathcal{K}, \end{aligned} \tag{9a}$$

$$\begin{aligned} &\mathrm{SINR}_{nk}^{\mathrm{UL}} \ge \gamma_k^{\mathrm{UL}},\\ &\forall\, k\in\mathcal{A}_n,\ n\in\mathcal{N}, \end{aligned} \tag{9b}$$

$$\begin{aligned} &\sum_{k\in\mathcal{A}_n}\Big\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\|_2^2 \le P_{n,\max}^{\mathrm{DL}},\\ &\forall\, n\in\mathcal{N}, \end{aligned} \tag{9c}$$

$$\begin{aligned} &p_k^{\mathrm{UL}} \le P_{k,\max}^{\mathrm{UL}},\ \forall\, k\in\mathcal{K}, \end{aligned} \tag{9d}$$

$$\begin{aligned} &\Big|\theta_m^{\mathrm{UL}}\Big| = \Big|\theta_m^{\mathrm{DL}}\Big| = 1,\\ &\forall\, m\in\mathcal{M}, \end{aligned} \tag{9e}$$

$$\begin{aligned} &\boldsymbol{v}_{nk}^{\mathrm{UL}} = \boldsymbol{v}_{nk}^{\mathrm{DL}} = \mathbf{0},\\ &\forall\, k\notin\mathcal{A}_n, n\in\mathcal{N}, \end{aligned} \tag{9f}$$

where $P_{k,\max}^{\mathrm{UL}}$ and $P_{n,\max}^{\mathrm{DL}}$ denote the maximum transmit power of MD k and BS n in the uplink and downlink, respectively, and (9f) are the group sparsity constraints of beamforming vectors. Specifically, if $k\notin\mathcal{A}_n$, BS n does not decode MD k's data in the uplink (i.e., $\boldsymbol{v}_{nk}^{\mathrm{UL}}=\mathbf{0}$) and subsequently cannot transmit inference results to MD k in the downlink (i.e., $\boldsymbol{v}_{nk}^{\mathrm{DL}}=\mathbf{0}$).

As $\mathcal{A}$ is a variable to be designed, we need to search over all possibilities of $\mathcal{A}$ to obtain the optimal task selection strategy $\mathcal{A}^\star$. Therefore the overall optimization problem is given by

$$\mathscr{P}_{\mathrm{original}}:\quad \underset{\mathcal{A}_1\subseteq\mathcal{K},\ldots,\mathcal{A}_N\subseteq\mathcal{K}}{\text{minimize}}\quad p^\star(\mathcal{A}) \tag{10}$$

where $p^\star(\mathcal{A})$ is the objective value of problem $\mathscr{P}(\mathcal{A})$.

In order to solve problem $\mathscr{P}_{\mathrm{original}}$, we are confronted with several main challenges. As set $\mathcal{K}$ has 2^K different subsets, $\mathcal{A}$ has a total of 2^{KN} different possibilities, making it apparently impractical to search over all the possibilities. Despite of the troublesome variable $\mathcal{A}$, the coupled continuous variables phases shifts and beamforming vectors in constraints (9a)-(9b) pose a unique challenge. Moreover, the unit-modulus constraint (9e) imposed by the phase-shift of each RIS element is nonconvex. In the following, we shall exploit the group sparsity structure of beamforming vectors to get rid of the combinatorial variable $\mathcal{A}$ in problem $\mathscr{P}_{\mathrm{original}}$, thereby facilitating efficient algorithm design.

III. Block-Structured Optimization Approach

In general, the combinatorial optimization problem $\mathscr{P}_{\mathrm{original}}$ is hard to tackle. Fortunately, based on the key observation that the task selection strategy $\mathcal{A}$ has an intrinsic connection with the group sparsity structure of beamforming vector $\boldsymbol{v}_{nk} = [(\boldsymbol{v}_{nk}^{\mathrm{UL}})^{\mathsf{T}},(\boldsymbol{v}_{nk}^{\mathrm{DL}})^{\mathsf{T}}]^{\mathsf{T}}$, we can eliminate the troublesome variable $\mathcal{A}$. Specifically, all coefficients in the beamforming group $\boldsymbol{v}_{nk}$ are zero simultaneously if $k\notin\mathcal{A}_n$. In other words, we have $k\notin\mathcal{A}_n \Leftrightarrow \boldsymbol{v}_{nk}=\mathbf{0}$ and $k\in\mathcal{A}_n \Leftrightarrow \boldsymbol{v}_{nk}\neq\mathbf{0}$. Therefore, the overall network power consumption given in (8) can be equivalently rewritten as

$$\begin{aligned} P_{\mathrm{total}}\Big(\Big\{p_k^{\mathrm{UL}}\Big\},\Big\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\}\Big) &= \sum_{k=1}^{K} p_k^{\mathrm{UL}} + \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{\eta_n}\Big\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\Big\|_2^2\\ &\quad+ \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{\boldsymbol{v}_{nk}\neq\mathbf{0}\}}P_{nk}^{\mathsf{c}}. \end{aligned} \tag{11}$$

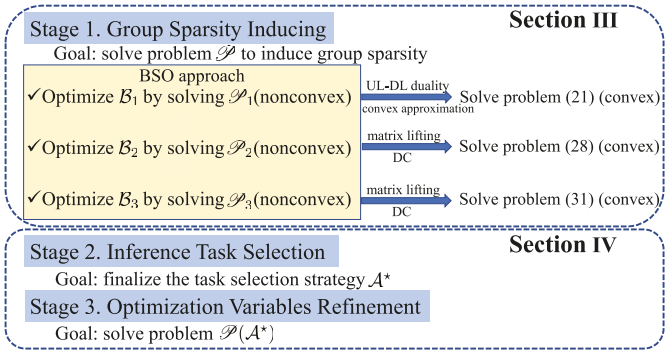


Fig. 2. Overview of the proposed three-stage framework for problem $\mathscr{P}_{\text{original}}$.

As multiple tasks may not be performed by a specific BS, the aggregated beamforming vector $\boldsymbol{v} \in \mathbb{C}^{K\sum_{n=1}^{N} L_n}$ defined as $\boldsymbol{v} = [\boldsymbol{v}_{11}^{\mathsf{T}}, \ldots, \boldsymbol{v}_{1K}^{\mathsf{T}}, \ldots, \boldsymbol{v}_{N1}^{\mathsf{T}}, \ldots, \boldsymbol{v}_{NK}^{\mathsf{T}}]^{\mathsf{T}}$ is expected to have the group sparsity structure with only a few non-zero blocks.

The above discussions indicate that we do not have to explicitly optimize the task selection strategy $\mathcal{A}$. Instead, $\mathcal{A}$ can be determined by the group sparsity pattern of the beamforming vectors, i.e., $\mathcal{A}_n = \{k | \boldsymbol{v}_{nk} \neq \boldsymbol{0},\ k \in \mathcal{K}\},\ \forall\, n \in \mathcal{N}$. Therefore, we propose to solve problem $\mathscr{P}_{\text{original}}$ based on the following three stages. In *Stage 1*, we presume that each BS performs inference tasks for all the MDs (i.e., $\mathcal{A}_n = \mathcal{K}, \forall\, n \in \mathcal{N}$), and solve the following group sparse beamforming problem to obtain the initial group sparse beamforming vectors, which indicate the group sparsity structure of the optimal beamforming vectors

$$\begin{aligned}
\mathscr{P}:\ \underset{\substack{\boldsymbol{\Theta}^{\text{UL}},\boldsymbol{\Theta}^{\text{DL}}\\ \{\boldsymbol{v}_{nk}^{\text{UL}}\},\{\boldsymbol{v}_{nk}^{\text{DL}}\},\{p_k^{\text{UL}}\}}}{\text{minimize}}\ & P_{\text{total}}\Big(\big\{p_k^{\text{UL}}\big\},\big\{\boldsymbol{v}_{nk}^{\text{DL}}\big\}\Big)\\
\text{subject to}\ & \text{SINR}_k^{\text{DL}} \geq \gamma_k^{\text{DL}},\ \forall\, k \in \mathcal{K}, && \text{(12a)}\\
& \text{SINR}_{nk}^{\text{UL}} \geq \gamma_k^{\text{UL}},\ \forall\, n \in \mathcal{N}, k \in \mathcal{K}, && \text{(12b)}\\
& \sum_{k=1}^{K}\Big\|\boldsymbol{v}_{nk}^{\text{DL}}\Big\|_2^2 \leq P_{n,\max}^{\text{DL}},\\
& \forall\, n \in \mathcal{N}, && \text{(12c)}\\
& p_k^{\text{UL}} \leq P_{k,\max}^{\text{UL}},\ \forall\, k \in \mathcal{K}, && \text{(12d)}\\
& \Big|\theta_m^{\text{DL}}\Big| = 1,\ \forall\, m \in \mathcal{M}, && \text{(12e)}\\
& |\theta_m^{\text{UL}}| = 1,\ \forall\, m \in \mathcal{M}. && \text{(12f)}
\end{aligned}$$

Based on the initial group sparse beamforming vectors, we in *Stage 2* determine the task selection strategy $\mathcal{A}^\star$. Finally, we solve problem $\mathscr{P}(\mathcal{A}^\star)$ to refine the optimization variables in *Stage 3*. In the remaining of this section, we propose a BSO approach to solve problem $\mathscr{P}$, while the details of *Stage 2* and *Stage 3* are presented in Section IV. We provide a graphic illustration in Fig. 2 to provide a better overview of the proposed three-stage framework.

The main idea of the BSO approach is to first partition the variables into several blocks, and then alternately optimize one of the blocks in each iteration while keeping the others fixed [43]. Specifically, we partition the five optimization variables into three blocks, denoted as $\mathcal{B}_1 = (\{\boldsymbol{v}_{nk}^{\text{DL}}\}, \{\boldsymbol{v}_{nk}^{\text{UL}}\}, \{p_k^{\text{UL}}\}), \mathcal{B}_2 = \boldsymbol{\Theta}^{\text{DL}}$, and $\mathcal{B}_3 = \boldsymbol{\Theta}^{\text{UL}}$.

A. Optimizing Variables $\{\boldsymbol{v}_{nk}^{\text{UL}}\}, \{p_k^{\text{UL}}\}$, and $\{\boldsymbol{v}_{nk}^{\text{DL}}\}$

When $\mathcal{B}_2$ and $\mathcal{B}_3$ are fixed, problem $\mathscr{P}$ is reduced to the following problem

$$\mathscr{P}_1:\ \underset{\{\boldsymbol{v}_{nk}^{\text{DL}}\},\{\boldsymbol{v}_{nk}^{\text{UL}}\},\{p_k^{\text{UL}}\}}{\text{minimize}}\ P_{\text{total}}\Big(\big\{p_k^{\text{UL}}\big\},\big\{\boldsymbol{v}_{nk}^{\text{DL}}\big\}\Big)\quad \text{subject to}\ \ \text{(12a)-(12d)}.$$

It is observed that optimization variables $\boldsymbol{v}_{nk}^{\text{UL}}$ and $\boldsymbol{v}_{nk}^{\text{DL}}$ are coupled only in the objective function, but not in the constraints. For the sake of analysis convenience, we temporarily dismiss the indicator term in the objective function and split problem $\mathscr{P}_1$ into two parts, i.e., the downlink part $\mathscr{P}_{1\text{-}1}$ and the uplink part $\mathscr{P}_{1\text{-}2}$. Afterwards we will combine the uplink and downlink parts to derive an effective solution to problem $\mathscr{P}_1$. The power minimization problem in the downlink part is

$$\mathscr{P}_{1\text{-}1}:\ \underset{\{\boldsymbol{v}_{nk}^{\text{DL}}\}}{\text{minimize}}\ \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{\eta_n}\Big\|\boldsymbol{v}_{nk}^{\text{DL}}\Big\|_2^2\quad \text{subject to}\ \ \text{(12a)}, \text{(12c)},$$

which is a celebrated problem formulation in unicast multiple-input single-output systems. Due to the fact that an arbitrary phase rotation of vector $\boldsymbol{v}_{nk}^{\text{DL}}$ does not affect the SINR constraints (12a) [44], we can replace (12a) with the following second-order cone (SOC) constraint

$$\sqrt{\sum_{l\neq k}\Big|\big(\boldsymbol{g}_k^{\text{DL}}\big)^{\mathsf{H}}\boldsymbol{v}_l^{\text{DL}}\Big|^2 + \sigma_k^2} \leq \frac{1}{\sqrt{\gamma_k^{\text{DL}}}}\Re\Big(\big(\boldsymbol{g}_k^{\text{DL}}\big)^{\mathsf{H}}\boldsymbol{v}_k^{\text{DL}}\Big), \tag{13}$$

where $\boldsymbol{g}_k^{\text{DL}} = [(\boldsymbol{g}_{1k}^{\text{DL}})^{\mathsf{T}}, \ldots, (\boldsymbol{g}_{Nk}^{\text{DL}})^{\mathsf{T}}]^{\mathsf{T}}$ and $\boldsymbol{v}_k^{\text{DL}} = [(\boldsymbol{v}_{1k}^{\text{DL}})^{\mathsf{T}}, \ldots, (\boldsymbol{v}_{Nk}^{\text{DL}})^{\mathsf{T}}]^{\mathsf{T}}$ denote the aggregated channel response vector and transmit beamforming vector with respect to MD k, respectively. Therefore, problem $\mathscr{P}_{1\text{-}1}$ is recast as a convex second-order cone programming (SOCP), which can be effectively solved by interior-point methods using modern software like CVX [45].

On the other hand, the power minimization problem in the uplink is given by

$$\mathscr{P}_{1\text{-}2}:\ \underset{\{\boldsymbol{v}_{nk}^{\text{UL}}\},\{p_k^{\text{UL}}\}}{\text{minimize}}\ \sum_{k=1}^{K} p_k^{\text{UL}}\quad \text{subject to}\quad \text{(12b)},\ \text{(12d)}.$$

Although the SINR constraints (12b) are similar to those in the downlink counterpart, we cannot convexify them in a similar way because the numerator involves both optimization variables. Moreover, another issue in $\mathscr{P}_{1\text{-}2}$ is that directly optimizing this problem makes $\boldsymbol{v}_{nk}^{\text{UL}}$ to be nearly zero, because an arbitrary scaling of $\boldsymbol{v}_{nk}^{\text{UL}}$ does not affect the uplink SINR constraints (12b) [46]. Specifically, if $(\tilde{\boldsymbol{v}}_{nk}^{\text{UL}}, \tilde{p}_k^{\text{UL}})$ denotes the optimal solution to problem $\mathscr{P}_{1\text{-}2}$, then we have $\tilde{\boldsymbol{v}}_{nk}^{\text{UL}} \approx \boldsymbol{0},\ \forall\, n \in \mathcal{N}, k \in \mathcal{K}$. Although the scaling issue does not

violate the group sparsity structure of $\boldsymbol{v}_{nk}$, it indicates that the receive beamforming vectors do not contribute to the task selection. Based on the uplink-downlink duality, in the following, we shall propose a virtual downlink formulation to overcome the scaling issue.

To facilitate an effective algorithm design, we relax problem $\mathscr{P}_{1\text{-}2}$ to the following problem

$$\begin{aligned}\underset{\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\},\{p_{nk}\}}{\text{minimize}}\ & \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}\\ \text{subject to}\ & p_{nk}\le P_{k,\max}^{\mathrm{UL}},\ \forall\, n\in\mathcal{N},\, k\in\mathcal{K}, \qquad (14a)\\ & \frac{p_{nk}\left|\left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}}\boldsymbol{g}_{nk}^{\mathrm{UL}}\right|^2}{\sum_{l\neq k}p_{nl}\left|\left(\boldsymbol{v}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}}\boldsymbol{g}_{nl}^{\mathrm{UL}}\right|^2+\sigma_n^2\left\|\boldsymbol{v}_{nk}^{\mathrm{UL}}\right\|_2^2}\ge\gamma_k^{\mathrm{UL}},\\ & \forall\, n\in\mathcal{N},\, k\in\mathcal{K}. \qquad (14b)\end{aligned}$$

We can easily verify that problem (14) is indeed a relaxation to problem $\mathscr{P}_{1\text{-}2}$, because given any solution $(\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\},\{p_k^{\mathrm{UL}}\})$ feasible to problem $\mathscr{P}_{1\text{-}2}$, $(\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\},\{p_{nk}\})$ is also feasible to problem (14) when $p_{nk}=p_k^{\mathrm{UL}},\forall\, n\in\mathcal{N}$, and the objective value of problem (14) is no greater than that of $\mathscr{P}_{1\text{-}2}$. The motivation for this relaxation is that problem (14) can be solved in the virtual downlink formulation so as to overcome the scaling issue.

We first consider an ideal scenario that the MDs have unlimited power budgets (i.e., $P_{k,\max}^{\mathrm{UL}}=+\infty,\forall\, k\in\mathcal{K}$). Problem (14) is then equivalent to the following virtual downlink power minimization problem

$$\begin{aligned}\underset{\{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}}{\text{minimize}}\ & \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2 \qquad (15a)\\ \text{subject to}\ & \mathrm{SINR}_{nk}^{\mathrm{VDL}}\ge\gamma_k^{\mathrm{UL}},\ \forall\, n\in\mathcal{N},k\in\mathcal{K}, \qquad (15b)\end{aligned}$$

where $\boldsymbol{v}_{nk}^{\mathrm{VDL}}\in\mathbb{C}^{L_n\times 1}$ denotes the virtual downlink transmit beamforming vector from BS n to MD k, and $\mathrm{SINR}_{nk}^{\mathrm{VDL}}$ is the virtual downlink SINR observed by MD k defined as $\mathrm{SINR}_{nk}^{\mathrm{VDL}}=\frac{|(\boldsymbol{g}_{nk}^{\mathrm{UL}})^{\mathsf{H}}\boldsymbol{v}_{nk}^{\mathrm{VDL}}|^2}{\sum_{l\neq k}|(\boldsymbol{g}_{nk}^{\mathrm{UL}})^{\mathsf{H}}\boldsymbol{v}_{nl}^{\mathrm{VDL}}|^2+\sigma_n^2}$. Note that in $\mathrm{SINR}_{nk}^{\mathrm{VDL}}$, the scaling issue does not exist for $\{\boldsymbol{v}_{nk}^{\mathrm{VDL}},n\in\mathcal{N},k\in\mathcal{K}\}$. The rigorous proof of the equivalence of the uplink power minimization problem (14) and the virtual downlink problem (15) can be derived by Lagrangian duality, as shown in the Appendix. The optimal solutions obtained by solving problem (15) have close connections to solutions to problem (14), i.e., $\boldsymbol{v}_{nk}^{\mathrm{VDL}}=\boldsymbol{v}_{nk}^{\mathrm{UL}}$ and $\sum_{n=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\|_2^2=\sum_{n=1}^{N}\sum_{k=1}^{K}p_{nk}$. However, it is worth mentioning that the equivalence between the virtual downlink beamforming power and the uplink transmit power does not necessarily hold, i.e., $\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\|_2^2\neq p_{nk}$. Therefore if $P_{k,\max}^{\mathrm{UL}}<+\infty$, we cannot directly rewrite the transmit power constraints (14a) as $\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\|_2^2\le P_{k,\max}^{\mathrm{UL}},\forall\, n,\forall\, k$ out of intuition and add them to problem (15). Instead, we consider a sum-power constraint to relax the uplink transmit power constraints (14a), which is given by

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2=\sum_{n=1}^{N}\sum_{k=1}^{K}p_{nk}\le N\sum_{k=1}^{K}P_{k,\max}^{\mathrm{UL}}. \qquad (16)$$

By introducing this mild power control to problem (15), we need to solve the following problem

$$\begin{aligned}\underset{\{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}}{\text{minimize}}\ & \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2\\ \text{subject to}\ & (15\text{b}),\ (16)\,. \qquad (17)\end{aligned}$$

Similar to (13), constraint (15b) can be replaced with the following SOC constraint

$$\begin{aligned}\sqrt{\sum_{l\neq k}\left|\left(\boldsymbol{g}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}}\boldsymbol{v}_{nl}^{\mathrm{VDL}}\right|^2+\sigma_n^2}\le\ & \frac{1}{\sqrt{\gamma_k^{\mathrm{VDL}}}}\Re\\ & \times\left(\left(\boldsymbol{g}_{nk}^{\mathrm{UL}}\right)^{\mathsf{H}}\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right) \qquad (18)\end{aligned}$$

to make problem (17) a convex SOCP problem.

By exploiting the uplink-downlink duality and transforming the uplink model into a virtual downlink model, we address the challenge of the receive beamforming vectors scaling issue. As $\{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}$ have the same group sparsity pattern as $\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}$, combining the downlink and virtual downlink parts, we relax $\mathscr{P}_1$ to the following problem

$$\begin{aligned}\underset{\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\},\{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}}{\text{minimize}}\ & \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2+\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{\eta_n}\left\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\right\|_2^2\\ & +\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{\bar{\boldsymbol{v}}_{nk}\neq\mathbf{0}\}}P_{nk}^{\mathsf{c}}\\ \text{subject to}\ & (12\text{c}),\ (13),\ (16),\ (18), \qquad (19)\end{aligned}$$

where $\bar{\boldsymbol{v}}_{nk}$ in the objective function is defined as $\bar{\boldsymbol{v}}_{nk}=[(\boldsymbol{v}_{nk}^{\mathrm{VDL}})^{\mathsf{T}},(\boldsymbol{v}_{nk}^{\mathrm{DL}})^{\mathsf{T}}]^{\mathsf{T}}$. It is observed that all the constraints of problem (19) are convex.

Although the feasible set is convex, problem (19) is a nonconvex integer programming problem due to the indicator function in the objective function. We identify that the third term in the objective function of problem (19) is a weighted ℓ_0-norm of vector $\bar{\boldsymbol{v}}=[\bar{\boldsymbol{v}}_{11}^{\mathsf{T}},\bar{\boldsymbol{v}}_{12}^{\mathsf{T}},\ldots,\bar{\boldsymbol{v}}_{NK}^{\mathsf{T}}]^{\mathsf{T}}$ with weights P_{nk}^{c}'s, and it is non-convex. As ℓ_1-norm is a well-known convex relaxation to ℓ_0-norm, we relax the weighted ℓ_0-norm as

$$\begin{aligned}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{\bar{\boldsymbol{v}}_{nk}\neq\mathbf{0}\}}P_{nk}^{\mathsf{c}}&\approx\sum_{n=1}^{N}\sum_{k=1}^{K}\|\bar{\boldsymbol{v}}_{nk}\|_2P_{nk}^{\mathsf{c}}\\ &=\sum_{n=1}^{N}\sum_{k=1}^{K}\\ &\quad\times\sqrt{\left\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\right\|_2^2+\left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2}P_{nk}^{\mathsf{c}}.\end{aligned} \qquad (20)$$

Note that (20) is actually the weighted mixed $\ell_{1,2}$-norm of vector $\bar{\boldsymbol{v}}$. The mixed $\ell_{1,2}$-norm behaves like an ℓ_1-norm on vector $[\|\bar{\boldsymbol{v}}_{11}\|_2,\|\bar{\boldsymbol{v}}_{12}\|_2,\ldots,\|\bar{\boldsymbol{v}}_{NK}\|_2]$. The outer ℓ_1-norm induces the sparsity structure, while the inner ℓ_2-norm is responsible for forcing all coefficients in the beamforming group $\bar{\boldsymbol{v}}_{nk}$ to be zero. By adopting mixed $\ell_{1,2}$-norm as the convex relaxation of the indication function term, we can induce the group sparsity structure of beamforming groups $\{\bar{\boldsymbol{v}}_{nk},\ n\in\mathcal{N},\,k\in\mathcal{K}\}$.

Algorithm 1: Mixed $\ell_{1,2}$-Norm Based Group Sparsity Inducing for Problem $\mathscr{P}_1$

Input: $\boldsymbol{\Theta}^{\mathrm{UL}}, \boldsymbol{\Theta}^{\mathrm{DL}}$
1. Solve the convex relaxation problem (21) to obtain $\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}$
2. Set $\boldsymbol{v}_{nk}^{\mathrm{UL}} = \boldsymbol{v}_{nk}^{\mathrm{VDL}}, \forall\, n \in \mathcal{N}, k \in \mathcal{K}$, and then solve problem $\mathscr{P}_{1\text{-}2}$ with fixed $\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}$ to obtain $\{p_k^{\mathrm{UL}}\}$
Output: $\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}, \{p_k^{\mathrm{UL}}\}$

By replacing the indication function with its convex surrogate, we relax problem (19) as the following convex problem

$$\begin{aligned}\underset{\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\},\{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}}{\text{minimize}} \;& \sum_{n=1}^{N}\sum_{k=1}^{K}\left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2 + \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{\eta_n}\left\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\right\|_2^2 \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K}\sqrt{\left\|\boldsymbol{v}_{nk}^{\mathrm{DL}}\right\|_2^2 + \left\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right\|_2^2}\, P_{nk}^{\mathsf{c}} \\ \text{subject to}\;& \text{(12c)},\ \text{(13)},\ \text{(16)},\ \text{(18)}. \end{aligned} \tag{21}$$

Once we solve problem (21) and obtain the solutions $(\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\})$, the receive beamforming vector $\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}$ in $\mathscr{P}_1$ can be obtained by setting $\boldsymbol{v}_{nk}^{\mathrm{UL}} = \boldsymbol{v}_{nk}^{\mathrm{VDL}}, \forall\, n \in \mathcal{N}, k \in \mathcal{K}$, and $\{p_k^{\mathrm{UL}}\}$ can be obtained by solving problem $\mathscr{P}_{1\text{-}2}$ with $\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}$ fixed, which is reduced to a linear program.

The overall algorithm for solving problem $\mathscr{P}_1$ is summarized in Algorithm 1.

B. Optimizing Variable $\boldsymbol{\Theta}^{\mathrm{DL}}$

For given $\mathcal{B}_1$ and $\mathcal{B}_3$, we optimize $\boldsymbol{\Theta}^{\mathrm{DL}}$ to solve the resulting problem, which is termed as $\mathscr{P}_2$. Since $\boldsymbol{\Theta}^{\mathrm{DL}}$ does not appear in the objective function of $\mathscr{P}$, problem $\mathscr{P}_2$ is in fact a downlink feasibility detection problem, which is given by

$$\mathscr{P}_2:\ \text{find}\quad \boldsymbol{\Theta}^{\mathrm{DL}}\quad \text{subject to}\quad \text{(12a)},\ \text{(12e)}.$$

According to the $\mathrm{SINR}^{\mathrm{DL}}$ expression given in (7), we have

$$\begin{aligned}&\left|\sum_{n=1}^{N}\left(\boldsymbol{g}_{nk}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{v}_{nl}^{\mathrm{DL}}\right|^2 \\ &\overset{(a)}{=}\left|\sum_{n=1}^{N}\left(\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{DL}}\right)^{\mathsf{H}} + \left(\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{\Theta}^{\mathrm{DL}}\boldsymbol{G}_n^{\mathrm{DL}}\right)\boldsymbol{v}_{nl}^{\mathrm{DL}}\right|^2 \\ &\overset{(b)}{=}\left|\left(\boldsymbol{h}_{\mathrm{d},k}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{v}_l^{\mathrm{DL}} + \beta\left(\boldsymbol{a}^{\mathrm{DL}}\right)^{\mathsf{H}}\right. \\ &\qquad\left.\times\ \mathrm{diag}\left(\left(\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{DL}}\right)^{\mathsf{H}}\right)\tilde{\boldsymbol{G}}^{\mathrm{DL}}\boldsymbol{v}_l^{\mathrm{DL}}\right|^2, \end{aligned} \tag{22}$$

where (a) follows by substituting (6), and (b) holds by defining $\boldsymbol{h}_{\mathrm{d},k}^{\mathrm{DL}} = [(\boldsymbol{h}_{\mathrm{d},1k}^{\mathrm{DL}})^{\mathsf{T}}, \ldots, (\boldsymbol{h}_{\mathrm{d},Nk}^{\mathrm{DL}})^{\mathsf{T}}]^{\mathsf{T}}$, $\boldsymbol{v}_l^{\mathrm{DL}} = [(\boldsymbol{v}_{1l}^{\mathrm{DL}})^{\mathsf{T}}, \ldots, (\boldsymbol{v}_{Nl}^{\mathrm{DL}})^{\mathsf{T}}]^{\mathsf{T}}$, $\boldsymbol{a}^{\mathrm{DL}} = [\theta_1^{\mathrm{DL}}, \ldots, \theta_M^{\mathrm{DL}}]^{\mathsf{H}}$, $\tilde{\boldsymbol{G}}^{\mathrm{DL}} = [\boldsymbol{G}_1^{\mathrm{DL}}, \ldots, \boldsymbol{G}_N^{\mathrm{DL}}]$. Note that in (22) the only term related to phase shifts is $\boldsymbol{a}^{\mathrm{DL}}$. Therefore for notational ease, we define $\boldsymbol{w}_{kl}^{\mathrm{DL}} = \beta\,\mathrm{diag}((\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{DL}})^{\mathsf{H}})\tilde{\boldsymbol{G}}^{\mathrm{DL}}\boldsymbol{v}_l^{\mathrm{DL}}$, $b_{kl}^{\mathrm{DL}} = (\boldsymbol{h}_{\mathrm{d},k}^{\mathrm{DL}})^{\mathsf{H}}\boldsymbol{v}_l^{\mathrm{DL}}$, and the $\mathrm{SINR}^{\mathrm{DL}}$ expression in (7) can be equivalently rewritten as $\mathrm{SINR}_k^{\mathrm{DL}} = \frac{|b_{kk}^{\mathrm{DL}} + (\boldsymbol{a}^{\mathrm{DL}})^{\mathsf{H}}\boldsymbol{w}_{kk}^{\mathrm{DL}}|^2}{\sum_{l\neq k}|b_{kl}^{\mathrm{DL}} + (\boldsymbol{a}^{\mathrm{DL}})^{\mathsf{H}}\boldsymbol{w}_{kl}^{\mathrm{DL}}|^2 + \sigma_k^2}$, leading to the following inhomogeneous QCQP problem

$$\begin{aligned}\text{find}\;& \boldsymbol{a}^{\mathrm{DL}} \\ \text{subject to}\;& \left|a_m^{\mathrm{DL}}\right|^2 = 1,\ \forall\, m \in \mathcal{M}, \qquad \text{(23a)} \\ &\frac{\left|b_{kk}^{\mathrm{DL}} + \left(\boldsymbol{a}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{w}_{kk}^{\mathrm{DL}}\right|^2}{\sum_{l\neq k}\left|b_{kl}^{\mathrm{DL}} + \left(\boldsymbol{a}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{w}_{kl}^{\mathrm{DL}}\right|^2 + \sigma_k^2} \geq \gamma_k, \\ &\forall\, k \in \mathcal{K}. \qquad \text{(23b)}\end{aligned}$$

By introducing an auxiliary scalar t, and defining $\boldsymbol{R}_{kl}^{\mathrm{DL}} = \begin{bmatrix}\boldsymbol{w}_{kl}^{\mathrm{DL}}(\boldsymbol{w}_{kl}^{\mathrm{DL}})^{\mathsf{H}} & \boldsymbol{w}_{kl}^{\mathrm{DL}}(b_{kl}^{\mathrm{DL}})^{\mathsf{H}} \\ (\boldsymbol{w}_{kl}^{\mathrm{DL}})^{\mathsf{H}}b_{kl}^{\mathrm{DL}} & 0\end{bmatrix}$, and $\bar{\boldsymbol{a}}^{\mathrm{DL}} = \begin{bmatrix}\boldsymbol{a}^{\mathrm{DL}} \\ t^{\mathrm{DL}}\end{bmatrix}$, problem (23) can be transformed into the following homogeneous QCQP problem

$$\begin{aligned}\text{find}\;& \bar{\boldsymbol{a}}^{\mathrm{DL}} \\ \text{subject to}\;& \left|\bar{a}_m^{\mathrm{DL}}\right|^2 = 1,\quad \text{for}\ \ m = 1, \ldots, M+1, \\ &\frac{\left(\bar{\boldsymbol{a}}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{R}_{kk}^{\mathrm{DL}}\bar{\boldsymbol{a}}^{\mathrm{DL}} + \left|b_{kk}^{\mathrm{DL}}\right|^2}{\sum_{l\neq k}\left(\bar{\boldsymbol{a}}^{\mathrm{DL}}\right)^{\mathsf{H}}\boldsymbol{R}_{kl}^{\mathrm{DL}}\bar{\boldsymbol{a}}^{\mathrm{DL}} + \left|b_{kl}^{\mathrm{DL}}\right|^2 + \sigma_k^2} \geq \gamma_k, \\ &\forall\, k \in \mathcal{K}. \qquad \text{(24)}\end{aligned}$$

The non-convexity of problem (24) lies in the unit-modulus constraints. A common technique used to handle the nonconvex QCQP problems is matrix lifting. For ease of notation, we omit the superscript DL if it does not cause any ambiguity. In problem (24), as $\bar{\boldsymbol{a}}^{\mathsf{H}}\boldsymbol{R}_{kk}\bar{\boldsymbol{a}} = \mathrm{Tr}(\boldsymbol{R}_{kk}\bar{\boldsymbol{a}}\bar{\boldsymbol{a}}^{\mathsf{H}})$, and by introducing a new variable $\boldsymbol{A} = \bar{\boldsymbol{a}}\bar{\boldsymbol{a}}^{\mathsf{H}} \in \mathbb{C}^{(M+1)\times(M+1)}$, we rewrite problem (24) as the following feasibility detection problem

$$\begin{aligned}\text{find}\;& \boldsymbol{A} \\ \text{subject to}\;& \mathrm{Tr}(\boldsymbol{R}_{kk}\boldsymbol{A}) + |b_{kk}|^2 \geq \gamma_k\sum_{l\neq k}\mathrm{Tr}(\boldsymbol{R}_{kl}\boldsymbol{A}) \\ &\qquad + \gamma_k\left(\sum_{l\neq k}|b_{kl}|^2 + \sigma_k^2\right), \forall\, k \in \mathcal{K}, \qquad \text{(25a)} \\ &\boldsymbol{A}_{mm} = 1,\ \text{for}\ \ m = 1, \ldots, M+1, \qquad \text{(25b)} \\ &\boldsymbol{A} \succeq \boldsymbol{0},\ \ \mathrm{rank}(\boldsymbol{A}) = 1. \qquad \text{(25c)}\end{aligned}$$

Here $\boldsymbol{A} \succeq \boldsymbol{0}$ indicates that $\boldsymbol{A}$ is a positive semidefinite (PSD) matrix. The challenge in solving problem (25) is the nonconvex rank-one constraint. The SDR technique is widely adopted to tackle the rank-one constraint in QCQP problems [47]. By simply dropping the nonconvex rank-one constraint, SDR relaxes the problem into a convex semidefinite programming (SDP) problem, which can then be solved by CVX. If a feasible $\boldsymbol{A}$ with rank one is found, then $\bar{\boldsymbol{a}}$ can be obtained by singular value decomposition (SVD) of $\boldsymbol{A}$.

However, such a relaxation may not be tight, i.e., the solution obtained by SDR may not satisfy the rank-one constraint. As pointed out in [48], the performance of SDR degrades sharply as the problem size grows. In our case, when M and/or K is large, the probability of returning a rank-one solution is low. If this is the case, additional steps (i.e., Gaussian randomization) are required to construct a rank-one solution from the higher-rank solution obtained by solving problem (25) [20], [47]. However, it is still possible that we

fail to find a feasible solution to problem (24) after a large number of Gaussian randomizations.

In other words, dropping the rank-one constraint cannot accurately detect the feasibility of problem (25). Hence, we propose a novel DC representation for the rank-one constraint, which is guaranteed to obtain a solution satisfying the non-convex rank-one constraint if problem (25) is feasible. Note that for the PSD matrix $\boldsymbol{A}$, the rank-one constraint indicates that $\sigma_1(\boldsymbol{A}) > 0$ and $\sigma_i(\boldsymbol{A}) = 0,\ \forall\, i = 2,\ldots,M+1$, where $\sigma_i(\boldsymbol{A})$ is the i-th largest singular value of $\boldsymbol{A}$. And recall that the trace norm and spectral norm of $\boldsymbol{A}$ are defined as $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{M+1}\sigma_i(\boldsymbol{A})$ and $\|\boldsymbol{A}\| = \sigma_1(\boldsymbol{M})$, respectively. Hence, the rank-one constraint can be equivalently rewritten as the difference of these two convex norms, i.e.,

$$\mathrm{rank}(\boldsymbol{A}) = 1 \Leftrightarrow \mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\| = 0,\ \ \mathrm{Tr}(\boldsymbol{A}) > 0. \quad (26)$$

Based on the DC representation for the nonconvex rank-one constraint, we solve the following DC program

$$\begin{aligned} \underset{\boldsymbol{A}\succeq\boldsymbol{0}}{\text{minimize}}\ \ & g(\boldsymbol{A}) := \mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\| \\ \text{subject to}\ \ & \text{(25a)-(25b)}. \end{aligned} \quad (27)$$

Note that if the objective value $g(\boldsymbol{A})$ is zero, the rank-one constraint is satisfied and we obtain a feasible solution to problem (25). Although problem (27) is still nonconvex due to the concave term $-\|\boldsymbol{A}\|$ in the objective, we can adopt successive convex approximation to solve it in an iterative manner. Specifically, by linearizing the concave term, at iteration i we need to solve the following convex problem

$$\begin{aligned} \underset{\boldsymbol{A}\succeq\boldsymbol{0}}{\text{minimize}}\ \ & \mathrm{Tr}(\boldsymbol{A}) - \left\langle \partial\left\|\boldsymbol{A}^{[i-1]}\right\|, \boldsymbol{A}\right\rangle \\ \text{subject to}\ \ & \text{(25a)-(25b)}. \end{aligned} \quad (28)$$

where $\boldsymbol{A}^{[i-1]}$ is the solution obtained at iteration $i-1$ and $\partial\|\boldsymbol{A}^{[i-1]}\|$ denotes the subgradient of spectral norm at point $\boldsymbol{A}^{[i-1]}$. Note that one subgradient of $\|\boldsymbol{A}\|$ can be efficiently computed as $\boldsymbol{q}_1\boldsymbol{q}_1^{\mathsf{H}}$, where $\boldsymbol{q}_1$ is the vector corresponding to the largest singular value $\sigma_1(\boldsymbol{A})$ [49]. Given an initial $\boldsymbol{A}^{[0]}$ and by iteratively solving (28) until the objective function $g(\boldsymbol{A})$ in (27) becomes zero, we obtain an exact rank-one solution according to (26). We design a practical stopping criterion as $\mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\| < \epsilon_{DC}$, where ϵ_{DC} is a sufficiently small positive constant.

The convergence characteristic of the iterative DC algorithm for problem (27) is presented in the following proposition.

Proposition 1: The generated sequence $\{g(\boldsymbol{A}^{[i]})\}$ is strictly decreasing and the sequence $\{\boldsymbol{A}^{[i]}\}$ converges to a critical point of g from an arbitrary initial point $\boldsymbol{A}^{[0]}$.

Proof: Please refer to [48, Appendix B] for more details. ■

In fact, the proposed DC algorithm can always find a feasible $\boldsymbol{A}$ to problem (25), which guarantees the objective value of problem (27) converges to zero. This is because the feasible region of problem (25) is always non-empty, at least the obtained solution to problem (25) at iteration t (i.e., $\boldsymbol{\Theta}^{(t)}$) is still feasible at iteration $t+1$. Therefore, the strictly decreasing and non-negative sequence $\{g(\boldsymbol{A}^{[i]})\}$ can always converge to zero within finite steps.

After obtaining a feasible $\boldsymbol{A}$, the phase-shift matrix $\boldsymbol{\Theta}^{\mathrm{DL}}$ in $\mathscr{P}_2$ can be recovered as follows. By SVD of $\boldsymbol{A}$, we can obtain the solution to problem (24) as $\bar{\boldsymbol{a}}^{\mathrm{DL}} = [\boldsymbol{a}_0^{\mathrm{DL}}\ t_0^{\mathrm{DL}}]^{\mathsf{T}}$, then the solution to problem (23) can be computed as $\boldsymbol{a}^{\mathrm{DL}} = \boldsymbol{a}_0^{\mathrm{DL}}/t_0^{\mathrm{DL}}$, and the solution to $\mathscr{P}_2$ is given as $\boldsymbol{\Theta}^{\mathrm{DL}} = \beta\,\mathrm{diag}((\boldsymbol{a}^{\mathrm{DL}})^{\mathsf{H}})$.

C. Optimizing Variable $\boldsymbol{\Theta}^{\mathrm{UL}}$

As phase-shift matrix $\boldsymbol{\Theta}^{\mathrm{UL}}$ does not appear in the objective function of $\mathscr{P}$, given $\mathcal{B}_1$ and $\mathcal{B}_2$, the resulting problem denoted as $\mathscr{P}_3$ is also a feasibility detection problem

$$\mathscr{P}_3:\ \text{find}\ \ \boldsymbol{\Theta}^{\mathrm{UL}}\ \ \text{subject to}\ \ \text{(12b)},\ \text{(12f)}.$$

The same derivation process presented in the last subsection is also applicable to transform the uplink problem $\mathscr{P}_3$ into a homogeneous QCQP, therefore details are omitted here for brevity. Specifically, by defining

$$\begin{aligned} \boldsymbol{w}_{nkl}^{\mathrm{UL}} &= \beta\,\mathrm{diag}\left(\left(\boldsymbol{h}_{\mathrm{r},l}^{\mathrm{UL}}\right)^{\mathsf{H}}\right)\boldsymbol{G}_n^{\mathrm{UL}}\boldsymbol{v}_{nk}^{\mathrm{UL}}, \\ b_{nkl} &= \left(\boldsymbol{h}_{\mathrm{d},nl}^{\mathrm{UL}}\right)^{\mathsf{H}}\boldsymbol{v}_{nk}^{\mathrm{UL}}, \\ \boldsymbol{R}_{nkl}^{\mathrm{UL}} &= \begin{bmatrix} \boldsymbol{w}_{nkl}^{\mathrm{UL}}\left(\boldsymbol{w}_{nkl}^{\mathrm{UL}}\right)^{\mathsf{H}} & \boldsymbol{w}_{nkl}^{\mathrm{UL}}\left(b_{nkl}^{\mathrm{UL}}\right)^{\mathsf{H}} \\ \left(\boldsymbol{w}_{nkl}^{\mathrm{UL}}\right)^{\mathsf{H}} b_{nkl}^{\mathrm{UL}} & 0 \end{bmatrix},\ \bar{\boldsymbol{a}}^{\mathrm{UL}} = \begin{bmatrix} \boldsymbol{a}^{\mathrm{UL}} \\ t^{\mathrm{UL}} \end{bmatrix}, \end{aligned}$$

we have the following uplink homogeneous QCQP problem

$$\begin{aligned} \text{find}\ \ & \bar{\boldsymbol{a}}^{\mathrm{UL}} \\ \text{subject to}\ \ & \left|\bar{a}_m^{\mathrm{UL}}\right|^2 = 1,\ \ \text{for}\ m = 1,\ldots,M+1, \\ & \frac{p_k^{\mathrm{UL}}\left((\bar{\boldsymbol{a}}^{\mathrm{UL}})^{\mathsf{H}}\boldsymbol{R}_{nkk}^{\mathrm{UL}}\bar{\boldsymbol{a}}^{\mathrm{UL}} + \left|b_{nkk}^{\mathrm{UL}}\right|^2\right)}{\sum_{l\neq k}p_l^{\mathrm{UL}}\left((\bar{\boldsymbol{a}}^{\mathrm{UL}})^{\mathsf{H}}\boldsymbol{R}_{nkl}^{\mathrm{UL}}\bar{\boldsymbol{a}}^{\mathrm{UL}} + \left|b_{nkl}^{\mathrm{UL}}\right|^2\right) + \sigma_n^2\left\|\boldsymbol{v}_{nk}^{\mathrm{UL}}\right\|_2^2} \\ & \geq \gamma_k, \\ & \forall\, n\in\mathcal{N},\ k\in\mathcal{K}. \end{aligned} \quad (29)$$

Similar to the downlink counterpart, we can lift problem (29) to the following feasibility detection problem

$$\begin{aligned} \text{find}\ \ & \boldsymbol{A} \\ \text{subject to}\ \ & \mathrm{Tr}(\boldsymbol{R}_{nkk}\boldsymbol{A}) + |b_{nkk}|^2 \geq \gamma_k\sum_{l\neq k}\alpha_{kl}\,\mathrm{Tr}(\boldsymbol{R}_{nkl}\boldsymbol{A}) \\ & \qquad + \gamma_k\left(\sum_{l\neq k}\alpha_{kl}|b_{nkl}|^2 + c_{nk}\right),\ \forall\, n, \forall\, k, \quad \text{(30a)} \\ & \boldsymbol{A}_{mm} = 1,\ \ \text{for}\ \ m = 1,\ldots,M+1, \quad \text{(30b)} \\ & \boldsymbol{A}\succeq\boldsymbol{0},\ \ \mathrm{rank}(\boldsymbol{A}) = 1, \quad \text{(30c)} \end{aligned}$$

where $\boldsymbol{A} = \bar{\boldsymbol{a}}^{\mathrm{UL}}(\bar{\boldsymbol{a}}^{\mathrm{UL}})^{\mathsf{H}}$, $\alpha_{kl} = p_l^{\mathrm{UL}}/p_k^{\mathrm{UL}}$ and $c_{nk} = \sigma_n^2\|\boldsymbol{v}_{nk}^{\mathrm{UL}}\|_2^2/p_k^{\mathrm{UL}}$. Similarly, a feasible $\boldsymbol{A}$ to problem (30) can be obtained by iteratively solving the following convex problem

$$\begin{aligned} \underset{\boldsymbol{A}\succeq\boldsymbol{0}}{\text{minimize}}\ \ & \mathrm{Tr}(\boldsymbol{A}) - \left\langle \partial\left\|\boldsymbol{A}^{[i-1]}\right\|, \boldsymbol{A}\right\rangle \\ \text{subject to}\ \ & \text{(30a)-(30b)}, \end{aligned} \quad (31)$$

until the stopping criterion $\mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\| < \epsilon_{DC}$ is satisfied.

The overall algorithm for solving problem $\mathscr{P}_2$ (or $\mathscr{P}_3$) is summarized in Algorithm 2.

Algorithm 2: DC Algorithm for Feasibility Detection Problem $\mathscr{P}_2$ (or $\mathscr{P}_3$)

Input: $\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{VDL}}\}, \{p_k^{\mathrm{UL}}\}$, initial point $\boldsymbol{A}^{[0]}$ and set $i = 0$

while $\mathrm{Tr}(\boldsymbol{A}^{[i]}) - \|\boldsymbol{A}^{[i]}\| < \epsilon_{DC}$ **do**

1. $i \leftarrow i + 1$, $\partial\|\boldsymbol{A}^{[i-1]}\| = \boldsymbol{q}_1\boldsymbol{q}_1^{\mathsf{H}}$
2. Solve problem (28) (or problem (31))

end

Decompose $\boldsymbol{A}$ as $\boldsymbol{A} = \bar{\boldsymbol{a}}\bar{\boldsymbol{a}}^{\mathsf{H}}$; Denote $\bar{\boldsymbol{a}} = [\boldsymbol{a}_0, t_0]^{\mathsf{T}}$;

Obtain $\boldsymbol{a} = \boldsymbol{a}_0/t_0$ and $\boldsymbol{\Theta} = \beta\,\mathrm{diag}\left(\boldsymbol{a}^{\mathsf{H}}\right)$

Output: $\boldsymbol{\Theta}^{\mathrm{UL}}$ (or $\boldsymbol{\Theta}^{\mathrm{DL}}$)

D. Unified BSO Approach

Based on the above discussions, problem $\mathscr{P}$ is solved by iteratively solving problems $\mathscr{P}_1$, $\mathscr{P}_2$, and $\mathscr{P}_3$ in an alternating manner until convergence. We justify the effectiveness and depict the convergence behavior of the proposed BSO approach in the following proposition.

Proposition 2: With the BSO approach, the objective value of $\mathscr{P}_1$ is non-increasing in the consecutive iterations.

Proof: For ease of notation, we denote the objective value of $\mathscr{P}_1$ as $f(\boldsymbol{V}, \boldsymbol{\Omega})$, where the first variable $\boldsymbol{V}$ is an abstraction of three optimization variables $\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}$, and $\{p_k^{\mathrm{UL}}\}$ in $\mathscr{P}_1$, and the second variable $\boldsymbol{\Omega}$ abstracts phase-shift matrices $\boldsymbol{\Theta}^{\mathrm{DL}}$ and $\boldsymbol{\Theta}^{\mathrm{UL}}$ in $\mathscr{P}_2$ and $\mathscr{P}_3$. Assuming that $(\boldsymbol{V}^{(t)}, \boldsymbol{\Omega}^{(t)})$ is obtained at iteration t. If $\mathscr{P}_2$ and $\mathscr{P}_3$ are feasible, i.e., $(\boldsymbol{V}^{(t)}, \boldsymbol{\Omega}^{(t+1)})$ exists, $(\boldsymbol{V}^{(t)}, \boldsymbol{\Omega}^{(t+1)})$ is also feasible to $\mathscr{P}_1$. Therefore, $(\boldsymbol{V}^{(t)}, \boldsymbol{\Omega}^{(t)})$ and $(\boldsymbol{V}^{(t+1)}, \boldsymbol{\Omega}^{(t+1)})$ are feasible solutions to $\mathscr{P}_1$ at the consecutive iterations t and $t+1$, respectively. We have the following inequality $f(\boldsymbol{V}^{(t+1)}, \boldsymbol{\Omega}^{(t+1)}) \overset{(a)}{\leq} f(\boldsymbol{V}^{(t)}, \boldsymbol{\Omega}^{(t+1)}) \overset{(b)}{=} f(\boldsymbol{V}^{(t)}, \boldsymbol{\Omega}^{(t)})$, where (a) holds because $\boldsymbol{V}^{(t+1)}$ is the optimal solution to $\mathscr{P}_1$ for a given $\boldsymbol{\Omega}^{(t+1)}$ at iteration $t+1$, and (b) holds because the objective value P_{total} depends only on the value of $\boldsymbol{V}$ and is independent of $\boldsymbol{\Omega}$. ∎

The obtained beamforming vectors $\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\}$ and $\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\}$ after solving problem $\mathscr{P}$ should have group sparse patterns. In the next section, we shall discuss how to design the task selection strategy $\mathcal{A}$ based on the obtained solutions, and present the holistic three-stage framework for the network power minimization problem $\mathscr{P}_{\mathrm{original}}$.

IV. Three-Stage Framework for Network Power Minimization Problem

In this section, we propose a thorough three-stage framework for problem $\mathscr{P}_{\mathrm{original}}$. In *Stage 1*, we adopt the BSO with mixed $\ell_{1,2}$-norm and DC algorithm to induce the group sparsity structure of the uplink/downlink beamforming vectors and optimize the phase-shift matrices. The obtained solutions are served as a guideline to determine the inference task selection strategy $\mathcal{A}$. In *Stage 2*, an ordering rule is proposed for all tasks according to their priorities, which depend on the structured-sparse beamforming vectors obtained in *Stage 1* as well as several key system parameters (i.e., channel state information, amplifier efficiency and task computation power). Based on the task ordering, we perform a task selection procedure to finalize $\mathcal{A}$. In *Stage 3*, the beamforming vectors and the phase-shift matrices are refined with the finalized task selection strategy $\mathcal{A}$.

A. Stage 1. Group Sparsity Inducing

With randomly initialized phase-shift matrices $\boldsymbol{\Theta}^{\mathrm{UL}}$ and $\boldsymbol{\Theta}^{\mathrm{DL}}$, we repeatedly solve problems $\mathscr{P}_1$, $\mathscr{P}_2$, and $\mathscr{P}_3$ based on Algorithms 1 and 2 respectively in an alternating manner until the following stopping criterion is satisfied: the relative improvement of objective values of problem $\mathscr{P}_1$ defined as $(P_{\mathrm{total}}^{(t)} - P_{\mathrm{total}}^{(t+1)})/P_{\mathrm{total}}^{(t)}$ is below a predefined threshold ε, where $P_{\mathrm{total}}^{(t)}$ and $P_{\mathrm{total}}^{(t+1)}$ represent the objective values obtained in iterations t and $t+1$, respectively. The yielded beamforming vectors should have the group sparsity structures. It is worth mentioning that the relative improvement is expected to be non-negative, because P_{total} is non-increasing as proved in Proposition 2.

B. Stage 2. Inference Task Selection

The next question is how to determine the task selection strategy $\mathcal{A}$. In fact, it is inappropriate to directly use the beamforming vectors obtained in *Stage 1* to finalize $\mathcal{A}$, as the vectors may contain some very small but nonzero coefficients. As illustrated in (11), these nonzero coefficients indicate that the corresponding tasks are performed by the BSs, which may result in unnecessary computation power consumption. To address this issue, we utilize the obtained group-structured beamforming vectors as well as other prior information to provide a guideline to determine $\mathcal{A}$.

For ease of exposition, we define the set of all task indices as $\{(n,k)|n \in \mathcal{N},\ k \in \mathcal{K}\}$. The task indexed by (n, k) is considered to be *active* if $k \in \mathcal{A}_n$, and *inactive* otherwise. Specifically, the priority of task (n, k) is defined as $\tau_{nk} = \sqrt{\frac{\|[\boldsymbol{g}_{nk}^{\mathrm{UL}}, \boldsymbol{g}_{nk}^{\mathrm{DL}}]\|_2^2 \eta_n}{P_{nk}^{\mathrm{c}}}}\|\boldsymbol{v}_{nk}\|_2$. We sort all NK tasks in a descending order according to their priorities, i.e., $\tau_{\pi_1} \geq \tau_{\pi_2} \cdots \geq \tau_{\pi_{NK}}$, where π is a permutation of task indices (n,k)'s. Intuitively, if BS n has a higher power amplifier efficiency, a higher channel gain, and a higher beamforming gain with respect to MD k, but a lower computation power consumption, task (n, k) has a higher priority. A higher τ_{nk} implies that task (n, k) is more power-efficient, therefore it is more likely to be activated.

To finalize the set $\mathcal{A}$, we need to detect the feasibility of a sequence of problems

$$\begin{aligned} \text{find} \quad & \left\{\boldsymbol{v}_{nk}^{\mathrm{DL}}\right\}, \left\{\boldsymbol{v}_{nk}^{\mathrm{UL}}\right\}, \left\{p_k^{\mathrm{UL}}\right\} \\ \text{subject to} \quad & \text{(12a)-(12d)},\ \boldsymbol{v}_{\pi^{[j]}} = \boldsymbol{0}, \end{aligned} \tag{32}$$

where $\pi^{[j]} = \{\pi_{j+1}, \ldots, \pi_{NK}\}$ denotes the inactive task set at iteration j, and $\boldsymbol{v}_{\pi^{[j]}} = \boldsymbol{0}$ represents that all coefficients in those beamforming groups $\boldsymbol{v}_{nk}$'s with index $(n,k) \in \pi^{[j]}$ are set to zero. Note that the number of active tasks is within [K, NK]. Starting with $j = K$, we terminate the feasibility detection

Algorithm 3: BSO With Mixed $\ell_{1,2}$-Norm and DC Based Three-Stage Framework for Nonconvex Combinatorial Problem $\mathscr{P}_{\text{original}}$

Input: initial phase-shift matrices $\mathbf{\Theta}^{\text{UL}}$ and $\mathbf{\Theta}^{\text{DL}}$, and threshold ε
Stage 1: Alternatively optimize $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$
while *the improvement of the objective in problem* $\mathscr{P}_1$ *is greater than* ε **do**
1. solve $\mathscr{P}_1$ for $\boldsymbol{v}_{nk}^{\text{UL}}, p_k^{\text{UL}}, \boldsymbol{v}_{nk}^{\text{DL}}$ using Algorithm 1
2. solve $\mathscr{P}_2$ for $\mathbf{\Theta}^{\text{DL}}$ using Algorithm 2
3. solve $\mathscr{P}_3$ for $\mathbf{\Theta}^{\text{UL}}$ using Algorithm 2
end
Stage 2: Determine the inference task selection
1. Compute task priorities and sort all tasks in a descending order based on their priorities
2. Iteratively solve problem (32) until feasible to finalize the task selection strategy $\mathcal{A}^\star$
Stage 3: Solve problem $\mathscr{P}(\mathcal{A}^\star)$ to refine variables $\{\boldsymbol{v}_{nk}^{\text{DL}}\}, \{\boldsymbol{v}_{nk}^{\text{UL}}\}, \{p_k^{\text{UL}}\}, \mathbf{\Theta}^{\text{DL}}, \mathbf{\Theta}^{\text{UL}}$
Output: $\mathcal{A}^\star$, $\{\boldsymbol{v}_{nk}^{\text{DL}}\}, \{\boldsymbol{v}_{nk}^{\text{UL}}\}, \{p_k^{\text{UL}}\}, \mathbf{\Theta}^{\text{DL}}, \mathbf{\Theta}^{\text{UL}}$

procedure and return $\pi^{[j]}$ until problem (32) is feasible. The task selection strategy $\mathcal{A}^\star$ can be easily obtained from $\pi^{[j]}$.

Comparing problem (32) to $\mathscr{P}_1$, as the added constraint $\boldsymbol{v}_{\pi^{[j]}} = \mathbf{0}$ is convex, we can solve problem (32) using Algorithm 1. Details are thus omitted here for brevity.

C. Stage 3. Optimization Variables Refinement

After determining the task selection strategy $\mathcal{A}^\star$, we should refine the beamforming vectors to make sure $\boldsymbol{v}_{nk} = \mathbf{0}$ if $k \notin \mathcal{A}_n, \forall\, n \in \mathcal{N}$, and the phase-shift matrices at the RIS need to be refined as well. This can be achieved by solving problem $\mathscr{P}(\mathcal{A}^\star)$. Since the group sparsity constraints (9f) are convex, the BSO with mixed $\ell_{1,2}$-norm and DC algorithm to solve problem $\mathscr{P}$ is also applicable here to obtain the solutions. Details are thus omitted here.

D. Complexity Discussions

The overall algorithm for green edge inference is summarized in Algorithm 3. The computational complexity of Algorithm 1 is dominated by solving the SOCP problem (21), which is $\mathcal{O}(L^{3.5}K^{3.5})$ using interior-point methods [50]. The complexity of Algorithm 2 is dominated by iteratively solving the SDP problem (28) (or problem (31)) and computing the SVD of $\boldsymbol{A}$. An SDP problem with an $a \times a$ semidefinite matrix variable and b SDP constraints is solved with complexity $\mathcal{O}(\sqrt{a}(a^3b + a^2b^2 + b^3))$ by interior-point methods [51]. For problems (28) and (31), we have $a = M + 1$, $b = K + 1$ and $a = M + 1$, $b = NK + 1$, respectively. It is also observed in our simulations that the convergence rate of the iterative procedure is fast (less than 10 iterations), therefore, the overall complexity of Algorithm 2 is $\mathcal{O}(\sqrt{M}(M^3NK + M^2N^2K^2 + N^3K^3))$. In short, the computational complexity involved in *Stage 1* is $\mathcal{O}(TR)$, where T is the required iterations for the BSO to converge, and $R = L^{3.5}K^{3.5} + M^{3.5}NK + M^{2.5}N^2K^2 + M^{0.5}N^3K^3$ is the polynomial term.

We need to solve at most *NK* SOCP problems in *Stage 2*, and hence the worst-case complexity is $\mathcal{O}(NK(L^{3.5}K^{3.5}))$. The complexity involved in *Stage 3* is the same as that in *Stage 1*. Therefore, the overall complexity for the proposed three-stage framework is $\mathcal{O}(TR + NL^{3.5}K^{4.5})$. In other words, we propose a polynomial complexity algorithm for the combinatorial problem $\mathscr{P}_{\text{original}}$.

V. Simulation Results

In this section, we present the simulation results to verify the effectiveness of our proposed algorithm. We consider a network with five 8-antenna BSs and six MDs uniformly and randomly distributed in a square region of 300 *m* × 300 m. An RIS with 30 reflecting elements is located at the 3-dimensional coordinate (150, 0, 15). In addition, the BSs are with height 30 m (i.e., the coordinates of the BSs are (*x*, *y*, 30)), while the MDs are with height 0 m (i.e., the coordinates of the MDs are (*x*, *y*, 0)).

We consider the following distance-dependent path loss model $L(d) = T_0(\frac{d}{d_0})^{-\alpha}$, where T_0 is the constant path loss at the reference distance $d_0 = 1$ m, d is the Euclidean distance between the transceivers, and α is the path loss exponent. Each antenna of the BSs is assumed to have an isotropic radiation pattern (i.e., 0 dBi antenna gain), while each element of the RIS has a 3 dBi gain for fair comparison because it reflects signals only in its front half-space [29], [52]. Moreover, Rician fading is considered as the small-scale fading for the BS-RIS and RIS-MD channels. For example, the BS-RIS channel can be expressed as $\boldsymbol{G}_n^x = \sqrt{\frac{\kappa_{\text{BR}}}{1+\kappa_{\text{BR}}}}\boldsymbol{G}_n^{\text{LOS}} + \sqrt{\frac{1}{1+\kappa_{\text{BR}}}}\boldsymbol{G}_n^{\text{NLOS}}$, where κ_{BR} is the Rician factor representing the ratio of power between the deterministic line-of-sight (LOS) path and the scattered paths, $\boldsymbol{G}_n^{\text{LOS}}$ is the LOS component modeled as the product of the steering vectors of the BS-RIS link [53], $\boldsymbol{G}_n^{\text{NLOS}}$ is Rayleigh fading components with entries distributed as $\mathcal{CN}(0,1)$, and $x \in \{\text{UL}, \text{DL}\}$. The entries in $\boldsymbol{G}_n^x$ are then multiplied by the square root of distance-dependent path loss denoted by α_{BR}. According to the 3GPP propagation environment from [54, Table B.1.2.1-1], we set $\alpha_{\text{BR}} = 2.2$. In addition, other channels are similarly generated with $\alpha_{\text{BM}} = 3.67$ and $\kappa_{\text{BM}} = 0$ (i.e., Rayleigh fading to account for rich scattering) for the BS-MD channel, $\alpha_{\text{RM}} = 1.69$ and $\kappa_{\text{RM}} = 3$ for the RIS-MD channel.

We consider a system with bandwidth 1 MHz and set $T_0 = -30$ dB. The effective noise power for the BSs and the MDs are $\sigma_n^2 = -90$ dBm and $\sigma_k^2 = -80$ dBm, respectively. Without specified otherwise, other parameters are set as follows: $P_{n,\max}^{\text{DL}} = 1$ W, $P_{n,\max}^{\text{UL}} = 1$ W, $P_{nk}^{\mathsf{c}} = 0.45$ W, $\eta_n = 0.25$, $\beta = 1$, $\varepsilon = 10^{-2}$, $\gamma_k^{\text{UL}} = \frac{1}{2}\gamma_k^{\text{DL}}$, and $\epsilon_{DC} = 10^{-6}$.

We compare the proposed BSO with mixed $\ell_{1,2}$-norm and DC algorithm (abbreviated as BSO-$\ell_{1,2}$-DC) with the following benchmarks.

- *Without-RIS:* Without the deployment of an RIS, the equivalent channels in (2) and (6) contain only the direct link, i.e., $\boldsymbol{h}_{\text{r},k}^{\text{UL}} = \boldsymbol{h}_{\text{r},k}^{\text{DL}} = \mathbf{0}, \forall\, k$. As we do not need to

TABLE I
COMPARISON BETWEEN SYSTEMS WITH AND WITHOUT AN RIS

	Target SINR [dB]	-20	-15	-10	-5	0	5	10
Feasible Probability	Without-RIS	0.98	0.85	0.54	0.12	0.01	0	0
	BSO-$\ell_{1,2}$-RP	**1.00**	**1.00**	**1.00**	**0.91**	**0.49**	**0.25**	**0.19**
Overall Power Consumption	Without-RIS [W]	3.22	3.32	3.55	5.75	14.35	N/A	N/A
	BSO-$\ell_{1,2}$-RP [W]	**3.22**	**3.30**	**3.47**	**3.99**	**5.21**	**9.85**	**15.32**

optimize phase shifts in this case, the alternating process in *Stage 1* is simplified to solve $\mathscr{P}_1$ only once.

- *BSO with mixed $\ell_{1,2}$-norm and Random Phase* (abbreviated as BSO-$\ell_{1,2}$-RP): In this case, the phase shifts of all reflecting elements in both uplink and downlink transmissions are randomly chosen from $[0, 2\pi)$ and then used to solve problem $\mathscr{P}_1$. We do not solve problems $\mathscr{P}_2$ and $\mathscr{P}_3$ in *Stage 1* subsequently to optimize the phase shifts. This benchmark is designed to reveal the necessity of optimizing the phase-shift matrices.
- *BSO with mixed $\ell_{1,2}$-norm and SDR* (abbreviated as BSO-$\ell_{1,2}$-SDR): In this case, the nonconvex rank-one constraints in (25) and (30) are dropped. Gaussian randomization is then adopted to obtain a feasible solution to problems $\mathscr{P}_2$ and $\mathscr{P}_3$. The number of randomly generated vectors for Gaussian randomization is set as 1000. If Gaussian randomization fails to find a feasible solution, we terminate the alternating process in *Stage 1*.
- *Decoupled BSO with mixed $\ell_{1,2}$-norm and DC* (abbreviated as BSO-$\ell_{1,2}$-DC-decoupled): Instead of solving the coupled uplink and downlink power minimization problem (19) in *Stage 1*, in this case we solve $\mathscr{P}_{1\text{-}1}$ and $\mathscr{P}_{1\text{-}2}$ individually to derive the decoupled uplink and downlink designs. *Stage 2* and *Stage 3* are the same as BSO-$\ell_{1,2}$-DC.

We first compare our RIS-aided communication system with the conventional one without the assistance of an RIS. For fair comparison, we do not explicitly optimize the phase-shift matrices, i.e., we compare the performance of Without-RIS and BSO-$\ell_{1,2}$-RP.

We study the relationship between the feasible probability of problem $\mathscr{P}_{\text{original}}$ and the target SINR γ_k^{DL}. The feasible probability of problem $\mathscr{P}_{\text{original}}$ is defined as

$$\mathbb{P}\{\mathscr{P}_{\text{original}} \text{ is feasible}\} = \frac{\text{number of simulations } \mathscr{P}_{\text{original}} \text{ is feasible}}{\text{total number of simulations}}.$$

As the target SINR requirements become more stringent, i.e., larger values of γ_k^{UL} and γ_k^{DL}, the feasibility probability of problem $\mathscr{P}_{\text{original}}$ is expected to decline. Results illustrated in Table I are averaged over 1000 independently generated channel realizations. We observe that the conventional system without RIS almost fails to support those settings with a target SINR being higher than 0 dB, while the RIS-aided system can still support with a high probability. In terms of the maximum SINR that the communication systems can support, we observe that there exists at least a 10 dB gain of the RIS-aided system over the system without RIS.

The lower part of Table I illustrates the overall network power consumption of systems with and without RIS. Under the same SINR requirement, it is observed that BSO-$\ell_{1,2}$-RP yields significantly lower power consumption. The supreme performance gain demonstrates that the deployment of an RIS in wireless communication systems can greatly boost the SINR and in turn reduce the overall network power consumption.

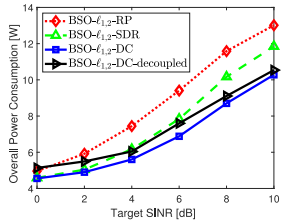


Fig. 3. Overall network power consumption versus target SINR γ_k^{DL}.

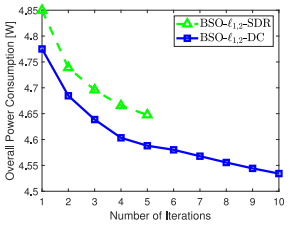


Fig. 4. Convergence behaviors of both BSO-$\ell_{1,2}$-DC and BSO-$\ell_{1,2}$-SDR algorithms.

We also show the superiority of our proposed BSO-$\ell_{1,2}$-DC algorithm in terms of the overall network power consumption, and the results obtained by averaging over 1000 independent channel realizations are shown in Fig. 3. The first observation is that both the BSO-$\ell_{1,2}$-DC and BSO-$\ell_{1,2}$-SDR algorithms significantly outperform BSO-$\ell_{1,2}$-RP, which demonstrates that dynamically optimizing the phase-shift matrices according to the beamforming vectors can reduce the network power consumption to a large extent. In addition, we observe that the proposed BSO-$\ell_{1,2}$-DC algorithm yields much lower overall network power consumption and is more energy efficient than the BSO-$\ell_{1,2}$-SDR algorithm. Given an overall power budget (e.g., 8.5 W), BSO-$\ell_{1,2}$-DC can achieve 1.5 dB higher SINR for the MDs than BSO-$\ell_{1,2}$-SDR. Such a performance gain is mainly because BSO-$\ell_{1,2}$-SDR may early terminate the alternating BSO process in *Stage 1* and cannot further proceed to find feasible solutions to problems $\mathscr{P}_2$ and/or $\mathscr{P}_3$. To make this more explicit, Fig. 4 shows the objective values of problem $\mathscr{P}_1$ in the first 10 alternating iterations in a specific channel realization. It is observed that as the BSO approach proceeds, the overall network power consumption for both BSO-$\ell_{1,2}$-DC and BSO-$\ell_{1,2}$-SDR algorithms are non-increasing, which validates our analysis in Proposition 2. It is also observed that

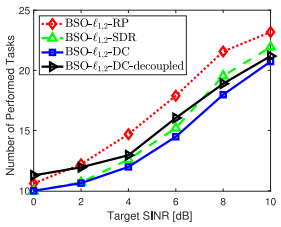


Fig. 5. Number of total performed tasks versus target SINR γ_k^{DL}.

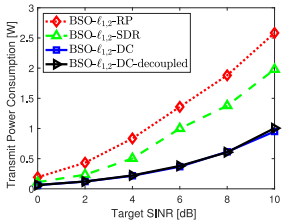


Fig. 6. Transmit power consumption versus target SINR γ_k^{DL}.

the BSO-$\ell_{1,2}$-SDR algorithm terminates at the 5th iteration, as SDR fails to obtain feasible solutions to problems $\mathscr{P}_2$ and/or $\mathscr{P}_3$ even with Gaussian randomization techniques. In contrast, DC can always yield feasible solutions as we have discussed in Section III-B, therefore BSO-$\ell_{1,2}$-DC terminates the alternating process only when the consecutive iterations make little progress. Comparing BSO-$\ell_{1,2}$-DC-decoupled to BSO-$\ell_{1,2}$-DC, we observe that BSO-$\ell_{1,2}$-DC always yields lower power consumption than BSO-$\ell_{1,2}$-DC-decoupled. This is because by coupling uplink and downlink beamforming designs, BSO-$\ell_{1,2}$-DC provides more accurate information to the task ordering rule and thus arranging all tasks in a more reasonable order. Moreover, the performance gap between BSO-$\ell_{1,2}$-DC-decoupled and BSO-$\ell_{1,2}$-DC shrinks as the target SINR grows. This indicates that when the target SINR is high, the impact of task order on reducing the overall power consumption decreases because nearly all tasks should be selected and performed.

Another interesting point worth mentioning in Fig. 3 is that the performance gaps between BSO-$\ell_{1,2}$-DC and BSO-$\ell_{1,2}$-SDR are getting larger as the value of the target SINR increases, which indicates that BSO-$\ell_{1,2}$-DC is especially appealing when high-quality services are required by the MDs. This is because a higher target SINR leads to a narrower feasible region of problems $\mathscr{P}_2$ and $\mathscr{P}_3$, making SDR less likely to find a feasible solution. In short, Fig. 3 shows that the proposed BSO-$\ell_{1,2}$-DC is able to reduce the overall network power consumption by 10% in the low SINR regime, and by up to 30% in the high SINR regime.

The number of tasks performed by all the BSs and the transmit power consumption versus the target SINR are shown in Fig. 5 and Fig. 6, respectively. As the target SINR increases, both the number of performed tasks and transmit power consumption increase. It is observed in Fig. 5 that BSO-$\ell_{1,2}$-DC can always perform fewer tasks to satisfy a certain target SINR. In other words, the long-lasting alternating iterations of BSO-$\ell_{1,2}$-DC shown in Fig. 4 helps in turn promote the group

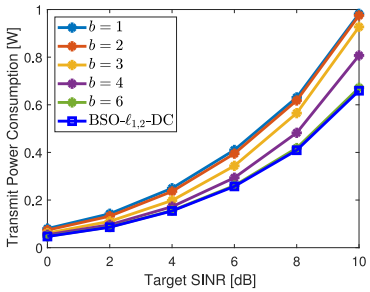


Fig. 7. Discrete and continuous phase shifts.

sparsity structure of beamforming vectors, thereby achieving lower computation power consumption. It is interesting to notice that BSO-$\ell_{1,2}$-DC-decoupled selects even more tasks than BSO-$\ell_{1,2}$-RP when the target SINR is 0 dB. This indicates the urgent demands to couple the uplink and downlink beamforming designs especially in the low SINR regime, so as to accurately characterize the tasks ordering rule and select the performed tasks reasonably. In terms of the transmit power consumption depicted in Fig. 6, we make the similar observation that BSO-$\ell_{1,2}$-DC yields the lowest transmit power consumption. Finally, it is also observed that the performance gaps between BSO-$\ell_{1,2}$-DC and other algorithms tend to be larger in the high SINR regime.

We compare the performances of discrete and continuous phase shifts in Fig. 7. With a b-bit resolution, the set of possible discrete phase shifts at each element is given as $\mathcal{D} = \{0, \Delta\theta, 2\Delta\theta, \ldots, (2^b - 1)\Delta\theta\}$, where $\Delta\theta = 2\pi/2^b$. After solving problems $\mathscr{P}_2$ and $\mathscr{P}_3$ to obtain continuous phase shift at each reflecting element, we quantize it to its nearest neighbor in $\mathcal{D}$. The quantization process takes place in each iteration of *Stage 1*, which influences the subsequent task selection process in *Stage 2*. For fair comparison, in simulations we consider that all NK tasks are performed and only show the impact of discrete phase shifts on the transmit power consumption. It is observed that the performances of discrete and continuous phase shifts are comparable when $b = 6$, which indicates a 6-bit resolution discrete phase shifts is practically sufficient.

VI. Conclusion

In this article, we investigated an RIS-aided edge inference system with multiple BSs cooperatively serving multiple MDs, taking into account both the uplink and downlink transmissions. The design of an energy-efficient edge inference system was formulated as a joint task selection strategy, uplink/downlink beamformers, transmit power, and uplink/downlink phase-shift matrices design problem. A BSO approach was proposed to decouple the optimization variables. For an efficient algorithm design, mixed $\ell_{1,2}$-norm was adopted to induce group sparsity of uplink/downlink beamforming vectors, while the matrix lifting and DC techniques were exploited to handle the nonconvex rank-one constraint and in turn solve the phase-shift matrix optimization problems. We also clarified the convergence behavior of the proposed BSO approach. Through extensive simulations,

$$
\begin{aligned}
L\Big(\boldsymbol{v}_{nk}^{\mathrm{VDL}};\lambda_{nk}\Big) &= \sum_{n=1}^{N}\sum_{k=1}^{K}\Big\|\boldsymbol{v}_{nk}^{\mathrm{VDL}}\Big\|_2^2 - \sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{nk}\left[\left|\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}}\boldsymbol{v}_{nk}^{\mathrm{VDL}}\right|^2/\gamma_k - \sum_{l\neq k}\left|\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}}\boldsymbol{v}_{nl}^{\mathrm{VDL}}\right|^2 - \sigma_n^2\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{nk}\sigma_n^2 + \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\boldsymbol{v}_{nk}^{\mathrm{VDL}}\Big)^{\mathsf{H}}\left[\boldsymbol{I} - \frac{\lambda_{nk}}{\gamma_k}\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}} + \sum_{l\neq k}\lambda_{nl}\boldsymbol{g}_{nl}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nl}^{\mathrm{UL}}\Big)^{\mathsf{H}}\right]\boldsymbol{v}_{nk}^{\mathrm{VDL}} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{nk}\sigma_n^2 + \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\boldsymbol{v}_{nk}^{\mathrm{VDL}}\Big)^{\mathsf{H}}\left[\boldsymbol{I} + \sum_{l=1}^{K}\lambda_{nl}\boldsymbol{g}_{nl}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nl}^{\mathrm{UL}}\Big)^{\mathsf{H}} - \Big(1+\frac{1}{\gamma_k}\Big)\lambda_{nk}\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}}\right]\boldsymbol{v}_{nk}^{\mathrm{VDL}}
\end{aligned} \tag{33}
$$

we demonstrated that deploying an RIS can significantly reduce the overall network power consumption. Furthermore, the effectiveness of the proposed DC algorithm in inducing low-rank solutions was also verified.

We identify the following relevant extensions as our future work. Developing a robust transmission design for RIS-aided edge inference with imperfect CSI is an interesting extension. Taking into account the RIS with coupled reflection amplitude and phase shifts is another interesting extension. In terms of scalability, the statistical CSI and the alternating direction method of multipliers (ADMM)-based parallel convex optimization framework can be utilized to further alleviate the computation burden for large-scale edge inference systems.

APPENDIX

We will show that problem (14) with $P_{k,\max}^{\mathrm{UL}} = +\infty, \forall\, k \in \mathcal{K}$ is equivalent to problem (15). As the fraction $\frac{1}{N}$ in the objective does not affect the equivalence, we dismiss the fraction to simplify the presentation.

The Lagrangian of problem (15) is shown in (33) at the top of the page. The dual function is given by $g(\lambda_{nk}) = \underset{\boldsymbol{v}_{nk}}{\text{minimize}}\ \ L(\boldsymbol{v}_{nk}^{\mathrm{VDL}};\lambda_{nk})$. The Lagrange dual problem is given as

$$
\begin{aligned}
\underset{\{\lambda_{nk}\}}{\text{maximize}}\ & \sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{nk}\sigma_n^2 \\
\text{subject to}\ & \lambda_{nk}\geq 0,\ \boldsymbol{\Lambda}_n \succeq \Big(1+\frac{1}{\gamma_k}\Big)\lambda_{nk}\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}}, \\
& \forall\, n\in\mathcal{N},\ k\in\mathcal{K},
\end{aligned} \tag{34}
$$

where $\boldsymbol{\Lambda}_n = \boldsymbol{I} + \sum_l \lambda_{nl}\boldsymbol{g}_{nl}^{\mathrm{UL}}(\boldsymbol{g}_{nl}^{\mathrm{UL}})^{\mathsf{H}}$, and strong duality holds.

Now we show that problem (14) is equivalent to problem (34). As $\boldsymbol{v}_{nk}^{\mathrm{UL}}$ is not in the objective of problem (14), the optimal $\boldsymbol{v}_{nk}^{\mathrm{UL}}$ that maximizes the SINR (14b) is the MMSE receiver given by $\hat{\boldsymbol{v}}_{nk}^{\mathrm{UL}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{g}_{nk}^{\mathrm{UL}}$, where $\boldsymbol{\Gamma} = \sum_{l=1}^{K} p_{nl}\boldsymbol{g}_{nl}^{\mathrm{UL}}(\boldsymbol{g}_{nl}^{\mathrm{UL}})^{\mathsf{H}} + \sigma_n^2\boldsymbol{I}$ [55]. Substituting $\hat{\boldsymbol{v}}_{nk}^{\mathrm{UL}}$ into (14b), we get

$$
\begin{aligned}
&\Big(1+\frac{1}{\gamma_k}\Big)p_{nk}\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}}\boldsymbol{\Gamma}^{-1}\boldsymbol{g}_{nk}^{\mathrm{UL}} \geq 1 \\
&\overset{(a)}{\Leftrightarrow} \boldsymbol{\Gamma} \succeq \Big(1+\frac{1}{\gamma_k}\Big)p_{nk}\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}},
\end{aligned}
$$

where (a) uses the property $\boldsymbol{y}^{\mathsf{H}}\boldsymbol{X}^{-1}\boldsymbol{y} \geq 1 \Leftrightarrow \boldsymbol{X} \succeq \boldsymbol{y}\boldsymbol{y}^{\mathsf{H}}$ proved in [55, Lemma 1]. By defining $p_{nk} = \lambda_{nk}\sigma_n^2$, $\boldsymbol{\Gamma} = \sigma_n^2\boldsymbol{\Lambda}_n$, the uplink problem (14) is then equivalent to

$$
\begin{aligned}
\underset{\{\lambda_{nk}\}}{\text{minimize}}\ & \sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{nk}\sigma_n^2 \\
\text{subject to}\ & \lambda_{nk}\geq 0,\ \boldsymbol{\Lambda}_n \preceq \Big(1+\frac{1}{\gamma_k}\Big)\lambda_{nk}\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big(\boldsymbol{g}_{nk}^{\mathrm{UL}}\Big)^{\mathsf{H}}, \\
& \forall\, n\in\mathcal{N},\ k\in\mathcal{K}.
\end{aligned} \tag{35}
$$

The optimal $\{\lambda_{nk}\}$ in both problems (34) and (35) will meet the SINR constraints with equality [44]. Therefore, we can safely reverse the SINR constraints as well as minimization to maximization in problem (35), which is now exactly the same as problem (34).

REFERENCES

[1] S. Hua and Y. Shi, "Reconfigurable intelligent surface for green edge inference in machine learning," in *Proc. IEEE Globecom Workshops*, Waikoloa, Hawaii, Dec. 2019, pp. 1–6.

[2] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.

[3] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[4] "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," Cisco, San Jose, CA, USA, White Paper, Feb. 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.pdf

[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. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the Internet of Vehicles," *Proc. IEEE*, vol. 108, no. 2, pp. 246–261, Feb. 2020.

[7] S. Liu, Y. Lin, Z. Zhou, K. Nan, H. Liu, and J. Du, "On-demand deep model compression for mobile devices: A usage-driven model selection framework," in *Proc. 16th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, Munich, Germany, Jun. 2018, pp. 389–400.

[8] B. Taylor, V. S. Marco, W. Wolff, Y. Elkhatib, and Z. Wang, "Adaptive deep learning model selection on embedded systems," *ACM SIGPLAN Notices*, vol. 53, no. 6, pp. 31–43, 2018.

[9] Z. Du *et al.*, "ShiDianNao: Shifting vision processing closer to the sensor," *ACM SIGARCH Comput. Archit. News*, vol. 43, no. 3, pp. 92–104, 2015.

[10] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.

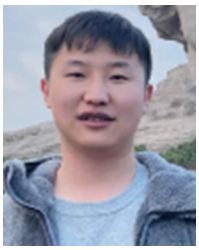
[11] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[12] 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.

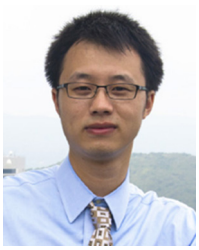
[13] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.
[14] C. Hu, W. Bao, D. Wang, and F. Liu, "Dynamic adaptive DNN surgery for inference acceleration on the edge," in *Proc. IEEE INFOCOM*, Paris, France, Apr. 2019, pp. 1423–1431.
[15] 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. 6071–6079.
[16] C. Louizos, K. Ullrich, and M. Welling, "Bayesian compression for deep learning," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 3288–3293.
[17] K. Yang, Y. Shi, W. Yu, and Z. Ding, "Energy-efficient processing and robust wireless cooperative transmission for edge inference," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9456–9470, Oct. 2020.
[18] X. Yang, S. Hua, Y. Shi, H. Wang, J. Zhang, and K. B. Letaief, "Sparse optimization for green edge AI inference," *J. Commun. Inf. Netw.*, vol. 5, no. 1, pp. 1–15, 2020.
[19] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.
[20] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.
[21] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.
[22] N. Rajatheva *et al.*, "White paper on broadband connectivity in 6G," 2020. [Online]. Available: https://arxiv.org/abs/2004.14247.
[23] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.
[24] C. Huang *et al.*, "Holographic mimo surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.
[25] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commu. Netw.*, vol. 2019, no. 1, pp. 1–20, Jan. 2019.
[26] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 162–169, Sep. 2018.
[27] H. Han, J. Zhao, D. Niyato, M. D. Renzo, and Q.-V. Pham, "Intelligent reflecting surface aided network: Power control for physical-layer broadcasting," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–7.
[28] M. Fu, Y. Zhou, and Y. Shi, "Intelligent reflecting surface for downlink non-orthogonal multiple access networks," in *Proc. IEEE Globecom Workshops*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.
[29] Q. Wu and R. Zhang, "Weighted sum power maximization for intelligent reflecting surface aided SWIPT," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 586–590, May 2020.
[30] Z. Wang, Y. Shi, Y. Zhou, H. Zhou, and N. Zhang, "Wireless-powered over-the-air computation in intelligent reflecting surface-aided IoT networks," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1585–1598, Feb. 2021.
[31] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces," in *Proc. 20th IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.
[32] G. C. Alexandropoulos, S. Samarakoon, M. Bennis, and M. Debbah, "Phase configuration learning in wireless networks with multiple reconfigurable intelligent surfaces," 2020. [Online]. Available: arXiv:2010.04376.
[33] K. Li, M. Tao, and Z. Chen, "Exploiting computation replication for mobile edge computing: A fundamental computation-communication tradeoff study," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4563–4578, Jul. 2020.
[34] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.
[36] Q.-U.-A. Nadeem, H. Alwazani, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface assisted multi-user MISO communication: Channel estimation and beamforming design," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 661–680, 2020.
[37] T.-J. Yang, Y.-H. Chen, J. S. Emer, and V. Sze, "A method to estimate the energy consumption of deep neural networks," in *Proc. 51st Asilomar Conf. Signals Syst. Comput. (ASCSSC)*, Pacific Grove, CA, USA, Oct. 2017, pp. 1916–1920.
[38] X. Xu *et al.*, "Scaling for edge inference of deep neural networks," *Nat. Electron.*, vol. 1, no. 4, pp. 216–222, 2018.
[39] R. Mahapatra, Y. Nijsure, G. Kaddoum, N. U. Hassan, and C. Yuen, "Energy efficiency tradeoff mechanism towards wireless green communication: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 686–705, 1st Quart., 2016.
[40] *CNN Energy Estimation Website*. [Online]. Available: http://energyestimation.mit.edu
[41] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.
[42] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.
[43] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.
[44] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.
[45] M. Grant and S. Boyd. (2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx
[46] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.
[47] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems and applications," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
[48] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.
[49] P. D. Tao and L. T. H. An, "Convex analysis approach to DC programming: Theory, algorithms and applications," *Acta Mathematica Vietnamica*, vol. 22, no. 1, pp. 289–355, Jan. 1997.
[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[51] I. Pólik and T. Terlaky, *Interior Point Methods for Nonlinear Optimization*. Berlin, Germany: Springer, 2010.
[52] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, Mar. 2020.
[53] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided MIMO communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1823–1838, Aug. 2020.
[54] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, 3GPP Standard TS 36.814, Mar. 2010.
[55] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.

Sheng Hua (Graduate Student Member, IEEE) received the B.S. degree from the School of Computer Science, Xidian University, Xi'an, China, in 2018. He is currently pursuing the master's degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His current research interests include reconfigurable intelligent surface and edge artificial intelligence.

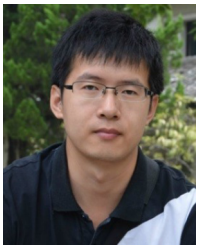
Yong Zhou (Member, IEEE) received the B.Sc. and M.Eng. degrees from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. From November 2015 to January 2018, he worked as a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada. He is currently an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include Internet of Things, edge computing, and reconfigurable intelligent surface.



Kai Yang (Member, IEEE) received the B.S. degree from Dalian University of Technology in 2015, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, in 2020. He is currently an AI Engineer in JD Technology Group focusing on designing efficient systems and algorithms of machine learning (federated learning in particular) and their applications in financial risk management.



Yuanming Shi (Senior 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 the Hong Kong University of Science and Technology in 2015. Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, where he is currently a tenured Associate Professor. He visited the University of California, 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, and AI. She is 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 IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.



Kunlun Wang (Member, IEEE) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016. From 2016 to 2017, he was with Huawei Technologies Company Ltd., where he was involved in energy efficiency algorithm design. From 2017 to 2019, he was with the Key Laboratory of Wireless Sensor Network and Communication, SIMIT, Chinese Academy of Sciences, Shanghai, China. From 2019 to 2020, he was with the School of Information Science and Technology, ShanghaiTech University. Since 2021, he has been a Professor with the School of Information Science and Technology, East China Normal University. His current research interests include energy efficient communications, fog computing networks, resource allocation, and optimization algorithm.