RIS-Aided Cooperative Mobile Edge Computing: Computation Efficiency Maximization via Joint Uplink and Downlink Resource Allocation

Zhenrong Liu, *Graduate Student Member, IEEE,* Zongze Li,
Yi Gong, *Senior Member, IEEE,* and Yik-Chung Wu, *Senior Member, IEEE*

Abstract—In mobile edge computing (MEC) systems, the wireless channel condition is a critical factor affecting both the communication power consumption and computation rate of the offloading tasks. This paper exploits the idea of cooperative transmission and employing reconfigurable intelligent surface (RIS) in MEC to improve the channel condition and maximize computation efficiency (CE). The resulting problem couples various wireless resources in both uplink and downlink, which calls for the joint design of the user association, receive/downlink beamforming vectors, transmit power of users, task partition strategies for local computing and offloading, and uplink/downlink phase shifts at the RIS. To tackle the challenges brought by the combinatorial optimization problem, the group sparsity structure of the beamforming vectors determined by user association is exploited. Furthermore, while the CE does not explicitly depend on the downlink phase shifts, instead of simply finding a feasible solution, we exploit the hidden relationship between them and convert this relationship into an explicit form for optimization. Then the resulting problem is solved via the alternating maximization framework, and the nonconvexity of each subproblem is handled individually. Simulation results show that cooperative transmission and RIS deployment can significantly improve the CE and demonstrate the importance of optimizing the downlink phase shifts with an explicit form.

Index Terms—Mobile edge computing (MEC), reconfigurable intelligent surface (RIS), computation efficiency, user association, cooperative transmission.

I. INTRODUCTION

Mobile users have long been recognized as resource-limited compared to static clients and servers. At any given cost and technology level, weight, size, battery life, ergonomics, and heat dissipation severely limit computational resources in terms of processor speed, memory size, and storage capacity. On the other hand, a wide range of emerging computation-intensive applications, such as mobile virtual reality and augmented reality, call for an unprecedented demand for data computing and processing in mobile users. To liberate the resource-limited users from heavy computation workloads and provide them with high-performance computing services, mobile edge computing (MEC) promotes the use of computing capabilities at the edge servers attached to the wireless access points (APs) [1], [2]. In this way, users' computation-intensive tasks can be offloaded to the nearby APs, and the results are delivered back to the users after the computation is done at the edge servers. Unlike cloud computing, whose vision is to centralize computation, storage, and network management

in a remote cloud center distant from the end users, MEC pushes computing resources, network control, and storage to the network edges [3]. This dramatically reduces latency, mobile energy consumption, and communication overhead of the network backhaul, thus overcoming the key challenges for materializing the 5G vision.

As the focus in MEC shifts from pure communications to communication-assisted computation, it is important to consider optimizing the computation efficiency [4], which is the computing capability (in terms of computation rate) of the MEC system divided by power spent on computing and communications [5]. However, increasing computing capability and decreasing power consumption are two conflicting goals. For example, if users are not offloading their tasks to the edge but computing them locally, power can be saved from communications, but the computation rate would be very low. On the contrary, if all users offload the tasks to the edge servers for computation, it causes high energy consumption for data offloading and results downloading, but the computation capability will be enhanced. Neither of these extremes is a good solution since they result in a waste of resources. Balancing the computation rate and the power consumption is a central problem in wireless MEC systems [6]–[8].

A well-known strategy for reducing downlink transmission energy is to have multiple APs cooperatively serve each user [9]–[11]. Different from the cellular systems, in wireless MEC, the data in downlink cooperative transmission is the computation results from users' offloaded data [12]. To enable cooperative transmission, the APs must either share the computation results of their respective users or each AP compute the results independently. Unfortunately, the computation results exchange between APs alone would cost 15 ms (for a lightly loaded network) to 50 ms (for a heavily loaded network) [13], [14], thus impractical for the latency-critical MEC applications such as mobile virtual reality and augmented reality, whose recommended total latency is around 20 ms [15]. Under the scenario that each cooperating AP independently computes the result for a serving user, more participating APs means more computation energy is spent in the hope of trading for less downlink transmission energy. This raises the fundamental question of which APs should be associated with which user. If the channel from a particular user k to an AP n is in bad condition, AP n should not

cooperatively serve the user k since the offloading energy from user k to AP n will be huge, thus lowering the computation efficiency. But if we are too conservative in choosing APs for cooperation, it defeats the original purpose of improving the quality of downlink transmission. Therefore, user association is critical to the overall MEC system's computation efficiency.

Besides cooperative transmission, another recent technology that could help to increase computation efficiency is reconfigurable intelligent surfaces (RISs). The motivation for incorporating RIS into MEC is driven by its remarkable capability to configure the wireless environment and thus improve the offloading rate and reduce transmission energy. Furthermore, RIS offers the advantage of being cost-effective and easy to deploy, making RIS particularly suitable for enhancing the uplink performance of MEC systems. By capitalizing on these benefits, the integration of RIS into the MEC framework holds great promise.

In this paper, we propose to use both RIS and cooperative transmission to improve the computation efficiency of an MEC system. In particular, we formulate the computation efficiency maximization problem by considering both the uplink and downlink transmission power consumption, and the local and offloaded computation power consumption. Since user association in the proposed MEC system couples with the design of receive and downlink beamforming vectors (i.e., if a user is not selected to be served by a particular AP, the corresponding receive and downlink beamforming vectors should be all-zero vectors), the resulting optimization problem is an intertwined design of user association, power partition parameters for computation offloading, receive and downlink beamforming vectors, and receive/downlink phase-shift matrices at the RIS with the practical phase shift model. Further with the fact that user association is a non-convex and combinatorial problem, the overall optimization problem is difficult to solve.

To address the challenge of the combinatorial optimization problem, we exploit the group sparsity structure of the receive/downlink beamforming vectors and merge the user association design into the beamforming variables. In addition, we observe that although the downlink phase-shift matrix does not appear in the objective function, it indirectly affects power consumption. This inspires us to exploit their hidden relationship to optimize the downlink phase-shift matrix with an explicit form instead of simply finding a feasible one. With the above two ideas, the optimization problem is solved under the alternating maximization (AM) framework. However, each of the subproblems is still non-convex. Tailoring to the different natures of non-convexity, these subproblems are handled by the tightest convex relaxation, Dinkelbach's or quadratic transformation, and the penalty method, respectively. Extensive simulation results show that the proposed RIS-aided cooperative MEC system outperforms systems without RIS or cooperative transmission. At the same time, optimizing the downlink phase shifts with an explicit form can further improve the system's performance.

The rest of the paper is organized as follows. The system model and the computation efficiency maximization problem are formulated in Section II. The discrete user association variable is handled in Section III. Section IV maximizes the
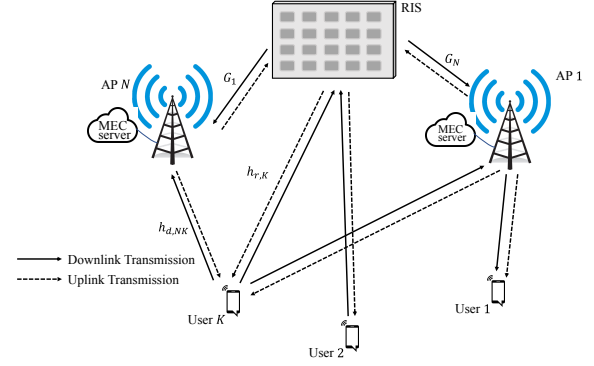

Fig. 1. The RIS-aided MEC system.

computation efficiency under the AM framework. Simulation results are presented in Section V. Finally, a conclusion is drawn in Section VI.

Notations: We use boldface lowercase (e.g., $\boldsymbol{h}$) and uppercase letters (e.g., $\boldsymbol{G}$) to represent vectors and matrices, respectively. The transpose, conjugate, conjugate transpose, and diagonal matrix are denoted as $(\cdot)^{\mathrm{T}}$, $(\cdot)^{*}$, $(\cdot)^{\mathrm{H}}$ and $\mathrm{diag}(\cdot)$, respectively. The symbol $\Re(\cdot)$ denotes the real component of a complex number. The $n \times n$ identity matrix is denoted as $\boldsymbol{I}_n$ and the entries of a matrix $\boldsymbol{X}$ is denoted by $(\boldsymbol{X})_{ij}$. The complex normal distribution is denoted as $\mathcal{CN}$. The ℓ_2-norm of a vector is denoted as $\|\cdot\|_2$ and the Frobenius norm of a matrix is denoted as $\|\cdot\|_{\mathrm{F}}$. $[a]^{+}$denotes $\max(0, a)$. 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 paper, the superscripts U and D 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. Finally, the set of a variable is denoted by $\{variable\}$.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a RIS-aided MEC system as shown in Fig. 1, in which there are N APs, K single-antenna users, and an M-element RIS. Each AP is attached to an MEC edge server and equipped with L antennas. Each user has a limited power budget P_k^c in Watt (W) but has intensive computation tasks to deal with. In practice, if the computation task is too complicated to be completed by a user (possibly due to excessive power or time involved), part of the task should be delegated to the edge server. Hence, a partial offloading mode is adopted. Partial offloading is designed to cope with computation tasks with data partition, in which tasks can be arbitrarily divided to facilitate parallel operations at users for local computing and offloading to the APs for edge computing [16]. We use time division duplex protocol and adjust RIS phase shifts between uplink and downlink to reconfigure the propagation environment in real-time via a control link between RIS and APs [17]. To guarantee the quality of services provided to users, it is assumed that multiple APs in the uplink transmission can successfully receive each user's data, enabling computation repetition at different edge servers and thus creating multiple copies of the computation results at different

APs [18]. These copies in turn enable cooperative downlink transmission among the APs on delivering the computation results. By exploiting existing channel estimation approaches [19], the global channel state information is assumed to be available at the APs.

For uplink transmission, let $\mathcal{N} = \{1, \ldots, N\}$, $\mathcal{K} = \{1, \ldots, K\}$, and $\mathcal{M} = \{1, \ldots, M\}$ denote the index sets of APs, users, and reflecting elements, respectively. In a specific time slot, let $s_k^{\mathrm{U}} \in \mathbb{C}$ with zero mean and unit variance denote the information symbol of user k for the offloading task, and $p_k^{\mathrm{U}} \in \mathbb{R}$ denote the transmit power of user k. The signal received at AP n can be expressed as

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

where $\boldsymbol{\iota}_{nk}^{\mathrm{U}} \in \mathbb{C}^{L \times 1}$ is the equivalent baseband channel from user k to AP n and $\boldsymbol{z}_n^{\mathrm{U}} \sim \mathcal{CN}\left(\boldsymbol{0}, \sigma_n^2 \boldsymbol{I}_L\right)$ is the receiver noise at AP n with σ_n^2 being the noise power. With the deployment of an RIS, the equivalent baseband channel from user k to AP n consists of direct and reflected links. Therefore, $\boldsymbol{\iota}_{nk}^{\mathrm{U}}$ can be modeled as

$$\boldsymbol{\iota}_{nk}^{\mathrm{U}} = \boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{U}} + \left(\boldsymbol{G}_n^{\mathrm{U}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{U}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{U}}, \quad \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{2}$$

where $\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{U}} \in \mathbb{C}^{L \times 1}, \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{U}} \in \mathbb{C}^{M \times 1}$, and $\boldsymbol{G}_n^{\mathrm{U}} \in \mathbb{C}^{M \times L}$ denote the channels from user k to AP n, from user k to the RIS, and from the RIS to AP n, respectively. The uplink RIS matrix $\boldsymbol{\Theta}^{\mathrm{U}} = \mathrm{diag}\left(\rho(\theta_1^{\mathrm{U}})e^{j\theta_1^{\mathrm{U}}}, \ldots, \rho(\theta_M^{\mathrm{U}})e^{j\theta_M^{\mathrm{U}}}\right) \in \mathbb{C}^{M \times M}$ is a diagonal matrix, where $\rho(\theta_m^{\mathrm{U}})$ is the amplitude reflection coefficient that is a function of uplink phase shift at the mth RIS element $\theta_m^{\mathrm{U}} \in [0, 2\pi)$, and can be expressed as [20]

$$\rho\left(\theta_m^{\mathrm{U}}\right) = (1 - \beta_{\min})\left(\frac{\sin\left(\theta_m^{\mathrm{U}} - \phi\right) + 1}{2}\right)^{\alpha} + \beta_{\min}, \forall m \in \mathcal{M}, \tag{3}$$

with $\beta_{\min} \geq 0, \phi \geq 0$, and $\alpha \geq 0$ are parameters related to the specific circuit implementation.

We consider the linear beamforming strategy and denote $\boldsymbol{v}_{nk}^{\mathrm{U}} \in \mathbb{C}^{L \times 1}$ as the receive beamforming vector of AP n for decoding s_k^{U}. AP n only decodes user k's transmitted symbol s_k^{U} if it belongs to the set of AP with indices $\mathcal{A}_k \subseteq \mathcal{N}$, which are responsible for computing the partial computation tasks for user k. If $n \in \mathcal{A}_k$, the received signal at AP n for user k, denoted by $\hat{s}_{nk}^{\mathrm{U}} \in \mathbb{C}$, is given by

$$\hat{s}_{nk}^{\mathrm{U}} = \left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}} \boldsymbol{\iota}_{nk}^{\mathrm{U}} \sqrt{p_k^{\mathrm{U}}} s_k^{\mathrm{U}} + \left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}} \sum_{l \neq k, l \in \mathcal{K}} \boldsymbol{\iota}_{nl}^{\mathrm{U}} \sqrt{p_l^{\mathrm{U}}} s_l^{\mathrm{U}}$$
$$+ \left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}} \boldsymbol{z}_n^{\mathrm{U}}, \quad \forall k \in \mathcal{K}. \tag{4}$$

For the offloading task, we introduce a power partition parameter $a_k \in [0, 1]$ for user k, and $p_k^{\mathrm{U}} = a_k P_k^c$ represents the power used for data offloading. Correspondingly, the uplink signal-to-interference-plus-noise ratio (SINR) observed at AP

n for user k is thus given by

$$\gamma_{nk}^{\mathrm{U}}\left(\{a_k\}, \boldsymbol{v}_{nk}^{\mathrm{U}}, \{\theta_m^{\mathrm{U}}\}\right)$$
$$= \frac{p_k^{\mathrm{U}} \left|\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{U}} + \left(\boldsymbol{G}_n^{\mathrm{U}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{U}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{U}}\right)\right|^2}{\sum\limits_{l \neq k, l \in \mathcal{K}} p_l^{\mathrm{U}} \left|\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\left(\boldsymbol{h}_{\mathrm{d},nl}^{\mathrm{U}} + \left(\boldsymbol{G}_n^{\mathrm{U}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{U}} \boldsymbol{h}_{\mathrm{r},l}^{\mathrm{U}}\right)\right|^2 + \sigma_n^2 \left\|\boldsymbol{v}_{nk}^{\mathrm{U}}\right\|_2^2},$$
$$\forall n \in \mathcal{A}_k, k \in \mathcal{K}. \tag{5}$$

After performing the computation tasks, the APs cooperatively transmit the computation results to the corresponding users through downlink wireless channels. We assume the downlink transmission starts simultaneously from all serving APs in our cooperative transmission protocol[1]. Let $\{\mathcal{A}_k\}$ denote the task selection strategy, $s_k^{\mathrm{D}} \in \mathbb{C}$ with zero mean and unit variance denote the symbol intended for user k, and $\boldsymbol{v}_{nk}^{\mathrm{D}} \in \mathbb{C}^{L \times 1}$ denote the downlink beamforming vector from AP n to user k. The signal transmitted by AP n, denoted as $\boldsymbol{x}_n^{\mathrm{D}} \in \mathbb{C}^{L \times 1}$, is a summation of beamformed symbols for all users with $n \in \mathcal{A}_k$, i.e., $\boldsymbol{x}_n^{\mathrm{D}} = \sum_{\{k | n \in \mathcal{A}_k, k \in \mathcal{K}\}} \boldsymbol{v}_{nk}^{\mathrm{D}} s_k^{\mathrm{D}}, \forall n \in \mathcal{N}$. Then, the signal received by user k is expressed as

$$y_k^{\mathrm{D}} = \sum_{n \in \mathcal{N}} \mathbf{1}_{\{n \in \mathcal{A}_k\}} \left(\boldsymbol{\iota}_{nk}^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{v}_{nk}^{\mathrm{D}} s_k^{\mathrm{D}}$$
$$+ \sum_{l \neq k} \sum_{n \in \mathcal{N}} \mathbf{1}_{\{n \in \mathcal{A}_l\}} \left(\boldsymbol{\iota}_{nk}^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{v}_{nl}^{\mathrm{D}} s_l^{\mathrm{D}} + z_k^{\mathrm{D}}, \quad \forall k \in \mathcal{K}, \tag{6}$$

where $z_k^{\mathrm{D}} \in \mathbb{C}$ is the AWGN at user k with zero mean and variance σ_k^2, and $\boldsymbol{\iota}_{nk}^{\mathrm{D}} \in \mathbb{C}^{L \times 1}$ is the equivalent baseband channel from AP n to user k. Similar to (2), $\boldsymbol{\iota}_{nk}^{\mathrm{D}}$ is expressed as

$$\boldsymbol{\iota}_{nk}^{\mathrm{D}} = \boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}, \quad \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{7}$$

where $\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} \in \mathbb{C}^{L \times 1}, \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}} \in \mathbb{C}^{M \times 1}$, and $\boldsymbol{G}_n^{\mathrm{D}} \in \mathbb{C}^{M \times L}$ denote the channels from AP n to user k, from the RIS to user k, and from AP n to the RIS, respectively. Furthermore, $\boldsymbol{\Theta}^{\mathrm{D}} = \mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}}, \ldots, \rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}}\right) \in \mathbb{C}^{M \times M}$ is the downlink RIS matrix with $\theta_m^{\mathrm{D}} \in [0, 2\pi)$. Based on the above equation, the SINR observed by user $k \in \mathcal{K}$ in the downlink transmission is given by

$$\mathrm{SINR}_k^{\mathrm{D}}\left(\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{\theta_m^{\mathrm{D}}\}\right)$$
$$= \frac{\left|\sum_{n \in \mathcal{N}} \mathbf{1}_{\{n \in \mathcal{A}_k\}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{v}_{nk}^{\mathrm{D}}\right|^2}{\sum_{l \neq k}\left|\sum_{n \in \mathcal{N}} \mathbf{1}_{\{n \in \mathcal{A}_l\}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{v}_{nl}^{\mathrm{D}}\right|^2 + \sigma_k^2},$$

---

[1]During the downlink phase, APs can employ alignment procedures to transmit only the finished bits across all the APs for a specific user at this time slot, and any remaining bits can be transmitted in the subsequent time slot's downlink phase.

$$= \frac{\left| \sum_{n\in\mathcal{N}} \left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}} \right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}} \right)^{\mathrm{H}} \boldsymbol{v}_{nk}^{\mathrm{D}} \right|^2}{\sum_{l\neq k} \left| \sum_{n\in\mathcal{N}} \left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}} \right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}} \right)^{\mathrm{H}} \boldsymbol{v}_{nl}^{\mathrm{D}} \right|^2 + \sigma_k^2}, \tag{8}$$

where the second equality holds because AP n does not transmit data to user k by setting $\boldsymbol{v}_{nk}^{\mathrm{D}} = \boldsymbol{0}$ if $n \notin \mathcal{A}_k$.

Remark 1. *Given the coverage area of RIS in an experimental system [21], as well as the density/coverage area of the base station in 5G NR [22], [23], utilizing a single RIS for multiple MEC servers is possible in practical deployments. This assumption is also widely accepted in RIS-aided MEC research [24], [25].*

Remark 2. *The RIS model described by (3) can be switched to the simultaneously transmitting and reflecting (STAR-) RIS model, thus enabling serving both users at the transmission and reflection side [26]. The resulting problem is simpler than the adopted practical phase shift model since the amplitude coefficients and phase shifts of the STAR-RIS are not coupled.*

B. Offloading Model

In our system model, the communication rate of user k offloaded to AP n is

$$R_{nk}\left(\{a_k\}, \boldsymbol{v}_{nk}^{\mathrm{U}}, \{\theta_m^{\mathrm{U}}\} \right) = B \log_2 \left(1 + \gamma_{nk}^{\mathrm{U}} \left(\{a_k\}, \boldsymbol{v}_{nk}^{\mathrm{U}}, \{\theta_m^{\mathrm{U}}\} \right) \right),$$
$$\forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{9}$$

where B is the bandwidth of the system. However, due to users offloading to multiple APs through multicast transmission, for user k, its communication rate is determined by the slowest rate among all the APs they are connected to. This is mathematically expressed as $\min_{n\in\mathcal{A}_k} R_{nk}$. Each user is allocated a data transmission time t_k, and the average computation rate for user k is expressed as $(t_k \min_{n\in\mathcal{A}_k} R_{nk})/T$, where T represents the time slot length. After data uploading to AP n, a computation frequency f_{nk} is allocated to each user k. If C_k is the amount of required computing resource (i.e., the number of CPU cycles) for computing 1-bit of the user k's input data, then we can calculate the maximum computation time among all the connected APs for user k's task as $(C_k t_k \min_{n\in\mathcal{A}_k} R_{nk})/\min_{n\in\mathcal{A}_k} f_{nk}$.

As for the case of local computing, the dynamic voltage and frequency scaling technique is adopted by all users for increasing the computation energy efficiency through adaptively controlling the CPU frequency used for computing [27]. In particular, the computation energy consumption of user $k \in \mathcal{K}$ can be expressed as $T\kappa_k f_k^2$, where f_k is the local computation frequency, κ_k is the effective capacitance coefficient of user k. Also, as $(1-a_k)P_k^c$ is the allocated power for local computing, we have $T(1-a_k)P_k^c = T\kappa_k f_k^2$, and thus we can calculate f_k as $f_k = \sqrt{((1-a_k)P_k^c)/\kappa_k}$, $\forall k \in \mathcal{K}$. The computation rate of user k for local computing is then given by

$$R_k^{\mathrm{loc}}(a_k) = \frac{f_k}{C_k} = \frac{1}{C_k}\sqrt{\frac{(1-a_k)P_k^c}{\kappa_k}}, \quad \forall k \in \mathcal{K}. \tag{10}$$

C. Power Consumption and Latency Model

The power consumption of each user's computation task includes both that used in local computing and computation at APs for offloaded tasks. To be specific, at user k, $(1-a_k)P_k^c$ W will be used for local computing [28]. In terms of the edge computation at the AP n, the average computation power is given as $\left((C_k t_k \min_{n\in\mathcal{A}_k} R_{nk}/f_{nk})f_{nk}^2\kappa_n \right)/T$. Therefore, the total computation power consumption for all APs and users is given by $\sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{A}_k} (C_k t_k f_{nk}\kappa_n \min_{n\in\mathcal{A}_k} R_{nk})/T + \sum_{k\in\mathcal{K}}(1-a_k)P_k^c$.

In terms of communication power consumption, it consists of the power consumed by the users in the data offloading and APs in the downlink transmission. The total data offloading power consumption is $\sum_{k\in\mathcal{K}} p_k^{\mathrm{U}} = \sum_{k\in\mathcal{K}} a_k P_k^c$, while the downlink transmit power consumption from all the APs to a specific user k is represented by $\sum_{n\in\mathcal{A}_k} \left\| \boldsymbol{v}_{nk}^{\mathrm{D}} \right\|_2^2$. Therefore, the total communication power consumption for both uplink and downlink transmissions is given by $\sum_{k\in\mathcal{K}} a_k P_k^c + \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{A}_k} \left\| \boldsymbol{v}_{nk}^{\mathrm{D}} \right\|_2^2$. Combining both computation and communication power consumption, the overall system power consumption can be expressed as

$$P_{\mathrm{total}}\left(\{\mathcal{A}_k\}, \{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{U}}\}, \{f_{nk}\}, \{a_k\}, \{t_k\}, \{\theta_m^{\mathrm{U}}\} \right) = \sum_{k\in\mathcal{K}} P_k^c$$
$$+ \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{A}_k} \left\| \boldsymbol{v}_{nk}^{\mathrm{D}} \right\|_2^2 + \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{A}_k} \frac{(C_k t_k f_{nk}\kappa_n \min_{n\in\mathcal{A}_k} R_{nk})}{T}. \tag{11}$$

Assuming the minimum required task's data size for user k is denoted as U_k bits [29]. Since local computing is ongoing during the entire time slot, the number of computed bits is TR_k^{loc} [5], and the remaining data bits for offloading is $U_k - TR_k^{\mathrm{loc}}$. Then the offloading latency for user k plus computation latency at AP n is calculated as

$$\tau_{nk} = (U_k - TR_k^{\mathrm{loc}})/\min_{n\in\mathcal{A}_k} R_{nk} + (U_k - TR_k^{\mathrm{loc}})/f_{nk}, \tag{12}$$

where the first term represents data offloading latency, and the second term accounts for computation latency at the AP.

D. Problem Formulation

In the proposed RIS-aided MEC system, there exists a fundamental tradeoff between the quality of service and the system power consumption. Specifically, with computation replication, more APs performing the same task improves the downlink diversity gain by exploiting cooperative transmission, at the cost of increasing the computation power consumption. To strike a good balance between communication power consumption, computation power consumption, and computation rate, we aim to maximize computation efficiency, defined as the ratio of total computation rate to total power consumption. This leads to the optimization problem being formulated as (13).

In the formulated problem, (13c) indicates that for each user k, the latency among all APs serving that user must be less than a threshold $\chi < T$. (13d) implies that for user k to perform uplink offloading, the subsequent downlink SINR must exceed the specified threshold γ_k^{D} for reliable reception

$$\text{P}_0: \max_{\substack{\{\boldsymbol{v}_{nk}^{\text{U}}\},\{\boldsymbol{v}_{nk}^{\text{D}}\},\{\mathcal{A}_k\},\{f_{nk}\}, \\ \{\theta_m^{\text{U}}\},\{\theta_m^{\text{D}}\},\{a_k\},\{t_k\}}} \frac{\sum_{k=1}^{K} t_k \min_{n\in\mathcal{A}_k} R_{nk}\left(\{a_k\},\boldsymbol{v}_{nk}^{\text{U}},\{\theta_m^{\text{U}}\}\right)/T + \sum_{k=1}^{K} R_k^{\text{loc}}(a_k)}{P_{\text{total}}\left(\{\mathcal{A}_k\},\{\boldsymbol{v}_{nk}^{\text{D}}\},\{\boldsymbol{v}_{nk}^{\text{U}}\},\{f_{nk}\},\{a_k\},\{t_k\},\{\theta_m^{\text{U}}\}\right)}, \tag{13a}$$

$$\text{s.t.} \quad a_k \in [0,1], \quad \forall k \in \mathcal{K}, \tag{13b}$$

$$\tau_{nk} \leq \chi, \quad \forall k \in \mathcal{K}, \ \forall n \in \mathcal{A}_k, \tag{13c}$$

$$\mathbf{1}_{\{a_k \neq 0\}}\text{SINR}_k^{\text{D}}\left(\{\boldsymbol{v}_{nk}^{\text{D}}\},\{\theta_m^{\text{D}}\}\right) \geq \mathbf{1}_{\{a_k \neq 0\}}\gamma_k^{\text{D}}, \quad k \in \mathcal{K}, \tag{13d}$$

$$\sum_{k\in\mathcal{K}} \left\|\boldsymbol{v}_{nk}^{\text{D}}\right\|_2^2 \leq P_{n,\max}^{\text{D}}, \quad \forall n \in \mathcal{N}, \tag{13e}$$

$$\boldsymbol{v}_{nk}^{\text{U}} = \boldsymbol{v}_{nk}^{\text{D}} = \mathbf{0}, \quad \forall n \notin \mathcal{A}_k, k \in \mathcal{K}, \tag{13f}$$

$$\boldsymbol{\Theta}^{\text{U/D}} = \text{diag}\left(\rho(\theta_1^{\text{U/D}})e^{j\theta_1^{\text{U/D}}},\ldots,\rho(\theta_M^{\text{U/D}})e^{j\theta_M^{\text{U/D}}}\right), \quad \forall m \in \mathcal{M}, \tag{13g}$$

$$t_k + (C_k t_k \min_{n\in\mathcal{A}_k} R_{nk})/\min_{n\in\mathcal{A}_k} f_{nk} \leq T, \quad \forall k \in \mathcal{K}, \tag{13h}$$

$$\sum_{k\in\mathcal{K}} f_{nk} \leq f_n, \ f_{nk} \geq 0, \quad \forall k \in \mathcal{K}, \forall n \in \mathcal{N}. \tag{13i}$$

of computation results. An indicator function exists in (13d) since if the user is determined not to offload, the downlink SINR does not matter. In this case, both sides of the constraint become zero, and the constraint is satisfied automatically. $P_{n,\max}^{\text{D}}$ in (13e) denote the maximum transmit power of AP n in the downlink. (13f) specifies that if $n \notin \mathcal{A}_k$, AP n does not decode user k's data in the uplink (i.e., $\boldsymbol{v}_{nk}^{\text{U}} = \mathbf{0}$) and subsequently would not transmit computation results to user k in the downlink (i.e., $\boldsymbol{v}_{nk}^{\text{D}} = \mathbf{0}$). (13f) reveals that the full set of the beamforming vectors $\{\boldsymbol{v}_{11}^{\text{U}},\cdots,\boldsymbol{v}_{NK}^{\text{U}},\boldsymbol{v}_{11}^{\text{D}},\cdots,\boldsymbol{v}_{NK}^{\text{D}}\}$ has a special structure where some components in the set are simultaneously zero. Constraint (13h) ensures the user's data transmission and computation time do not exceed the time slot length. Further, (13i) limits the cumulative computational frequency f_{nk} for each user k to not surpass AP n's total computational frequency f_n [29]. Notice that in the formulated problem, the norm $\|\boldsymbol{v}_{nk}^{\text{U}}\|$ does not matter as $\boldsymbol{v}_{nk}^{\text{U}}$ only appears in the objective function through R_{nk}, and from (5), $\boldsymbol{v}_{nk}^{\text{U}}$ appears in both the numerator and denominator of the SINR expression.

Efficiently solving P_0 is a highly challenging endeavor. Firstly, P_0 is intricate due to its general non-convexity in both the objective function and constraints, compounded by the non-differentiability from the maximin optimization structure within $\min_{n\in\mathcal{A}_k} R_{nk}$ and $\min_{n\in\mathcal{A}_k} f_{nk}$. Secondly, since discrete variables are involved in $\{\mathcal{A}_k\}$, this makes P_0 a combinatorial optimization problem. While brute force and randomized search algorithms are widely accepted techniques for solving the combinatorial optimization problem, they are not desirable solutions as brute force search is computationally prohibitive due to the exponential complexity [30] while randomized search lacks solution quality guarantee [31], [32]. Thirdly, since variable $\{\theta_m^{\text{D}}\}$ does not appear in the objective function but indirectly affects the objective value through (13d), it is nontrivial to derive an optimal solution for $\{\theta_m^{\text{D}}\}$. In the subsequent sections, we begin by transforming the maximin optimization problem into a more tractable form. Following that, we leverage on the distinctive structure of the

beamforming vectors to address the discrete variable $\{\mathcal{A}_k\}$ within P_0, thus simplifying the subsequent algorithm design.

Remark 3. *In the proposed system, we assume that users determine the parameter γ_k^{D} based on their specific requirements and is independent of the optimization variable a_k as in [33], [34]. This is because the number of bits uploaded to the AP does not directly relate to the number of bits in the downlink in wireless MEC for applications such as image classification and video tracking [35].*

III. TRANSFORMATION OF MAXIMIN OPTIMIZATION AND HANDLING DISCRETE VARIABLE IN P_0

The objective function of P_0 is non-differentiable due to the maximin optimization structure, which is embedded in the terms $\min_{n\in\mathcal{A}_k} R_{nk}$ in the objective function and the terms $\min_{n\in\mathcal{A}_k} R_{nk}$ and $\min_{n\in\mathcal{A}_k} f_{nk}$ within the constraints (13h). To address this, we reformulate it as follows:

$\text{P}_1:$

$$\max_{\substack{\{\boldsymbol{v}_{nk}^{\text{U}}\},\{\boldsymbol{v}_{nk}^{\text{D}}\},\{f_{nk}\}, \\ \{\mathcal{A}_k\},\{\theta_m^{\text{U}}\},\{\theta_m^{\text{D}}\},\{t_k\} \\ \{R_k\},\{a_k\},\{f_k^{\min}\}}} \frac{\sum_{k=1}^{K} t_k R_k/T + \sum_{k=1}^{K} R_k^{\text{loc}}(a_k)}{P_{\text{total}}\left(\{\mathcal{A}_k\},\{R_k\},\{\boldsymbol{v}_{nk}^{\text{D}}\},\{f_{nk}\},\{t_k\}\right)}, \tag{14a}$$

$$\text{s.t.} \ (13b) - (13g), (13i),$$

$$t_k + (C_k t_k R_k)/f_k^{\min} \leq T, \quad \forall k \in \mathcal{K}, \tag{14b}$$

$$R_k \leq \min_{n\in\mathcal{A}_k} R_{nk}\left(\{a_k\},\boldsymbol{v}_{nk}^{\text{U}},\{\theta_m^{\text{U}}\}\right), \quad \forall k \in \mathcal{K}, \tag{14c}$$

$$f_k^{\min} \leq \min_{n\in\mathcal{A}_k} f_{nk}, \quad \forall k \in \mathcal{K}, \tag{14d}$$

where R_k and $f_k^{\min}$ are introduced as auxiliary variables. Note that P_1 and P_0 are considered equivalent since to maximize the objective function, the value of R_k and $f_k^{\min}$ needs to be set equal to $\min_{n\in\mathcal{A}_k} R_{nk}\left(\{a_k\},\boldsymbol{v}_{nk}^{\text{U}},\{\theta_m^{\text{U}}\}\right)$ and $\min_{n\in\mathcal{A}_k} f_{nk}$, respectively [36], [37].

Regarding the discrete variable $\{\mathcal{A}_k\}$, notice that it has an intrinsic connection with the group sparsity structure of the beamforming vectors. Specifically, $\boldsymbol{v}_{nk}^{\text{U}}$ and $\boldsymbol{v}_{nk}^{\text{D}}$ will be zero

at the same time if $n \notin \mathcal{A}_k$ or non-zero at the same time if $n \in \mathcal{A}_k$. This inspires us to impose a value restriction on the mixed $\ell_{1,2}$ norm [35],[38]: $\sum_{n=1}^{N} \sum_{k=1}^{K} \left| \|[(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}} \ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}]\|_2 \right|$ to enforce the group sparsity structure. In this formulation, the outer ℓ_1-norm induces sparsity, while the inner ℓ_2-norm is responsible for forcing all coefficients in the beamforming group $[(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}} \ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}]$ to be zero. By doing so, we do not have to explicitly optimize the task selection strategy $\{\mathcal{A}_k\}$. Instead, $\{\mathcal{A}_k\}$ can be determined by the group sparsity pattern of the beamforming vectors, i.e., $\mathcal{A}_k = \left\{ n \mid [(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}} \ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}] \neq \boldsymbol{0}, n \in \mathcal{N} \right\}$.

Next, to leverage the sparsity structure for more efficient problem-solving, we need to transform the constraints (13c), (14c) and (14d), which involve the variable $\{\mathcal{A}_k\}$. The equivalent transformation for the latency constraints (13c) is given by

$$\mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \left(\frac{U_k - T R_k^{\mathrm{loc}}}{R_k} + \frac{U_k - T R_k^{\mathrm{loc}}}{f_{nk}} \right) \leq \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \chi, \quad (15)$$
$$k \in \mathcal{K}, n \in \mathcal{N},$$

which is obtained by putting the expression for τ_{nk} from equation (12) into (13c). In (15), the indicator function signifies that for user k, if AP n is not in $\mathcal{A}_k$, its latency is disregarded (as both sides of the inequality in (15) become 0, automatically satisfying the constraint). Similarly, for (14c) and (14d), we can equivalently transform them as

$$\mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} R_k \leq \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} R_{nk}, \quad k \in \mathcal{K}, n \in \mathcal{N}, \quad (16)$$

and

$$\mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} f_k^{\min} \leq \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} f_{nk}, \quad k \in \mathcal{K}, n \in \mathcal{N}, \quad (17)$$

respectively. Therefore, the transformed problem of P_1 is given by

$$\mathrm{P}_2: \max_{\substack{\{\boldsymbol{v}_{nk}^{\mathrm{U}}\},\{\boldsymbol{v}_{nk}^{\mathrm{D}}\},\{R_k\}, \\ \{f_{nk}\},\{\theta_m^{\mathrm{U}}\},\{\theta_m^{\mathrm{D}}\}, \\ \{a_k\},\{f_k^{\min}\},\{t_k\}}} \frac{\sum_{k=1}^{K} t_k R_k / T + \sum_{k=1}^{K} R_k^{\mathrm{loc}}(a_k)}{P_{\mathrm{total}}\left(\{R_k\}, \{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{f_{nk}\}, \{t_k\}\right)},$$
$$(18\mathrm{a})$$

$$\text{s.t.} \quad (13\mathrm{b}), (13\mathrm{d}), (13\mathrm{e}), (13\mathrm{g}),$$
$$(13\mathrm{i}), (14\mathrm{b}), (15), (16), (17),$$
$$\sum_{n=1}^{N} \sum_{k=1}^{K} \left| \|[(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}} \ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}]\|_2 \right| \leq \beta \quad (18\mathrm{b})$$

for some $\beta > 0$ and $P_{\mathrm{total}}\left(\{R_k\}, \{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{f_{nk}\}, \{t_k\}\right)$ can be equivalently rewritten as

$$P_{\mathrm{total}}\left(\{R_k\}, \{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{f_{nk}\}, \{t_k\}\right) = \sum_{k \in \mathcal{K}} P_k^c$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \left(\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2^2 + \frac{(C_k t_k R_k f_{nk} \kappa_n)}{T} \right), \quad (19)$$

where an indicator function is used in the second term due to a change from summation over $\mathcal{A}_k$ to summation over all n.

Finally, to address the non-convex constraints (16), we employ the logarithmic barrier function $\frac{1}{w} \log(-x)$ to replace the inequality and move it as a penalty term

$$\mathrm{P}_{2w}: \max_{\substack{\{\boldsymbol{v}_{nk}^{\mathrm{U}}\},\{\boldsymbol{v}_{nk}^{\mathrm{D}}\},\{R_k\}, \\ \{f_{nk}\},\{\theta_m^{\mathrm{U}}\},\{\theta_m^{\mathrm{D}}\}, \\ \{a_k\},\{f_k^{\min}\},\{t_k\}}} \frac{\sum_{k=1}^{K} t_k R_k / T + \sum_{k=1}^{K} R_k^{\mathrm{loc}}(a_k)}{P_{\mathrm{total}}\left(\{R_k\}, \{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{f_{nk}\}, \{t_k\}\right)}$$
$$+ \frac{1}{w} \sum_k \sum_n \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \left(\log(R_{nk} - R_k) \right),$$
$$(20)$$

$$\text{s.t.} \quad (13\mathrm{b}), (13\mathrm{d}), (13\mathrm{e}), (13\mathrm{g}),$$
$$(13\mathrm{i}), (14\mathrm{b}), (15), (17), (18\mathrm{b}).$$

As w increases towards infinity, P_{2w} becomes equivalent to P_2 [37]. Therefore, we implement a strategy for solving a series of progressively penalized problems. This approach involves starting with w at a small value and incrementally increasing it by a factor $\vartheta > 1$. For each subsequent problem with a higher value of w, we utilize the solution of the previous penalized problem as the initialization. This iterative process continues until w reaches a pre-set threshold value.

Upon merging the task selection strategy $\{\mathcal{A}_k\}$ into the beamforming vectors design, we note that the objective function in P_{2w} is in a fractional structure. In the subsequent section, we will employ quadratic and Dinkelbach's transformation techniques [38] to convert the problem into a series of concave subproblems, thus enabling efficient solutions.

IV. MAXIMIZING COMPUTATION EFFICIENCY BY OPTIMIZING P_{2w}

Since the optimization variables in P_{2w} are strongly coupled in the objective function, it is difficult to solve P_{2w} directly. To this end, we optimize P_{2w} under the AM framework. The main idea of the AM is to alternately optimize one block of variables at a time while keeping the others fixed [39]. Specifically, we partition the optimization problem P_{2w} into seven subproblems and solve them in an AM fashion. This section will cover the optimization with respect to $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}, \{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{f_{nk}, f_k^{\min}\}, \{\theta_m^{\mathrm{U}}\}, \{a_k\}$ and $\{R_k, t_k\}$. The next section will be dedicated to the optimization of $\{\theta_m^{\mathrm{D}}\}$.

A. Optimizing Variables $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$

When other variables are fixed, the subproblem for updating $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$ is

$$\min_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \left(\frac{(C_k t_k R_k f_{nk} \kappa_n)}{T} + \|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2^2 \right), \quad (21\mathrm{a})$$

$$\text{s.t.} \ \mathbf{1}_{\left\{a_k \neq 0\right\}} \mathrm{SINR}_k^{\mathrm{D}} \left(\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}, \{\theta_m^{\mathrm{D}}\} \right) \geq \mathbf{1}_{\left\{a_k \neq 0\right\}} \gamma_k^{\mathrm{D}}, \ k \in \mathcal{K}, \quad (21\mathrm{b})$$

$$\sum_{k \in \mathcal{K}} \|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2^2 \leq P_{n,\max}^{\mathrm{D}}, \quad \forall n \in \mathcal{N}, \quad (21\mathrm{c})$$

$$\mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \left(\frac{U_k - T R_k^{\mathrm{loc}}}{R_k} + \frac{U_k - T R_k^{\mathrm{loc}}}{f_{nk}} \right) \leq \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} \chi,$$
$$k \in \mathcal{K}, n \in \mathcal{N},$$
$$(21\mathrm{d})$$

$$\mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} R_k \leq \mathbf{1}_{\left\{\boldsymbol{v}_{nk}^{\mathrm{D}} \neq \boldsymbol{0}\right\}} R_{nk}, \quad k \in \mathcal{K}, n \in \mathcal{N}, \quad (21\mathrm{e})$$

$$\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}f_k^{\min}\leq\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}f_{nk},\ k\in\mathcal{K},n\in\mathcal{N},\quad(21\mathrm{f})$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\left|\|[(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}}\ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}]\|_2\right|\leq\beta.\quad(21\mathrm{g})$$

Since an arbitrary phase rotation of vector $\boldsymbol{v}_{nk}^{\mathrm{D}}$ does not affect the SINR constraints (21b), we can replace the non-convex constraints (21b) with second-order cone (SOC) constraints [40]:

$$\mathbf{1}_{\{a_k\neq0\}}\sqrt{\sum_{l\neq k}\left|\sum_{n\in\mathcal{N}}(\boldsymbol{\iota}_{nk}^{\mathrm{D}})^{\mathrm{H}}\boldsymbol{v}_{nl}^{\mathrm{D}}\right|^2+\sigma_k^2}$$
$$\leq\mathbf{1}_{\{a_k\neq0\}}\frac{1}{\sqrt{\gamma_k^{\mathrm{D}}}}\Re\left(\sum_{n\in\mathcal{N}}(\boldsymbol{\iota}_{nk}^{\mathrm{D}})^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{D}}\right),\quad\forall k\in\mathcal{K}.\quad(22)$$

For the objective function in (21), we identify that it has an indicator function which is equivalent to a weighted group ℓ_0-"norm" of $\left[(\boldsymbol{v}_{11}^{\mathrm{D}})^{\mathrm{T}},\cdots,(\boldsymbol{v}_{1K}^{\mathrm{D}})^{\mathrm{T}},\cdots,(\boldsymbol{v}_{NK}^{\mathrm{D}})^{\mathrm{T}}\right]^{\mathrm{T}}$, as given by $\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}$, with weights $(C_kt_kR_kf_{nk}\kappa_n)/T+\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2^2$ [41], and it is non-convex. As group ℓ_1-norm is the tightest convex relaxation to group ℓ_0-"norm", the objective function of (21) can be relaxed as $\sum_{n=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\left((C_kt_kR_kf_{nk}\kappa_n)/T+\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2^2\right)$. Furthermore, considering the constraints (21d), (21e) and (21f), because the ℓ_0-"norm" and ℓ_1-norm are always non-negative, replacing the ℓ_0-"norm" with the ℓ_1-norm will not change the relationship of the inequality. Therefore, the constraints (21d), (21e) and (21f) can be transformed into equivalent forms as

$$\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\left(\frac{U_k-TR_k^{\mathrm{loc}}}{R_k}+\frac{U_k-TR_k^{\mathrm{loc}}}{f_{nk}}\right)\leq\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\,\chi,$$
$$k\in\mathcal{K},n\in\mathcal{N},\quad(23)$$

$$\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\,R_k\leq\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\,R_{nk},\quad k\in\mathcal{K},n\in\mathcal{N},\quad(24)$$

and

$$\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\,f_k^{\min}\leq\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\,f_{nk},\ k\in\mathcal{K},n\in\mathcal{N},\quad(25)$$

respectively.

Therefore, the transformed problem of (21) is given as

$$\min_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}}\ \sum_{n=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2\left(\frac{(C_kt_kR_kf_{nk}\kappa_n)}{T}+\|\boldsymbol{v}_{nk}^{\mathrm{D}}\|_2^2\right),\quad(26)$$
$$\mathrm{s.t.}\quad(21\mathrm{c}),(21\mathrm{g}),(22),(23),(24),(25).$$

Since (26) is a convex optimization problem, it can be directly solved via the interior-point method [42]. Theoretical results from [43] show that the optimal solution to (26) provides a feasible solution to (21).

B. Optimizing Variables $\{f_{nk}\}$ and $\{f_k^{\min}\}$

Regarding the optimization of $\{f_{nk}\}$ and $\{f_k^{\min}\}$, the subproblem is given as

$$\min_{\{f_{nk}\},\{f_k^{\min}\}}\ \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}\left(\frac{t_kR_k}{T}f_{nk}\kappa_nC_k\right)\quad(27)$$
$$\mathrm{s.t.}\ (13\mathrm{i}),(14\mathrm{b}),(15),(17).$$

By multiplying both sides of the constraint (15) by f_{nk} and both sides of the constraint (14b) by $f_k^{\min}$, these constraints are transformed into linear constraints in terms of f_{nk} and $f_k^{\min}$, respectively. Consequently, this manipulation renders the optimization problem (27) as a convex optimization problem, which can be efficiently solved using the interior-point method.

C. Optimizing Variables $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}$

Considering P_{2w}, when the remaining variables are held fixed, the subproblem for updating $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}$ is

$$\max_{\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}}\quad\frac{1}{w}\sum_k\sum_n\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}\log(R_{nk}(\boldsymbol{v}_{nk}^{\mathrm{U}})-R_k),\quad(28\mathrm{a})$$

$$\mathrm{s.t.}\quad\sum_{n=1}^{N}\sum_{k=1}^{K}\left|\|[(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}}\ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}]\|_2\right|\leq\beta.\quad(28\mathrm{b})$$

Since $R_{nk}(\boldsymbol{v}_{nk}^{\mathrm{U}})$ has a fractional structure, we employ a quadratic transform to reconfigure it as

$$\max_{\{\boldsymbol{v}_{nk}^{\mathrm{U}}\},\boldsymbol{s}}\frac{1}{w}\sum_k\sum_n\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}\log(\log_2(Q_{nk}(\boldsymbol{v}_{nk}^{\mathrm{U}}))-R_k),\quad(29\mathrm{a})$$

$$\mathrm{s.t.}\quad\sum_{n=1}^{N}\sum_{k=1}^{K}\left|\|[(\boldsymbol{v}_{nk}^{\mathrm{D}})^{\mathrm{T}}\ (\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{T}}]\|_2\right|\leq\beta,\quad(29\mathrm{b})$$

where

$$Q_{nk}(\boldsymbol{v}_{nk}^{\mathrm{U}})=1+2\mathrm{Re}(s_{nk}p_k^{\mathrm{U}}(\boldsymbol{\iota}_{nk}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{U}})$$
$$-s_{nk}\left(\sum_{l\neq k}p_l^{\mathrm{U}}(\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{\iota}_{nl}^{\mathrm{U}}(\boldsymbol{\iota}_{nl}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{U}}+\sigma_n^2(\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{U}}\right)s_{nk}^{\mathrm{H}}\quad(30)$$

and $\boldsymbol{s}=\{s_{nk}\}$ is the set of auxiliary variables introduced through the quadratic transformation. Then iteratively solving (29) can obtain a stationary point of (28). In particular, when $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}$ are held fixed, the optimal s_{nk} is given by

$$s_{nk}^{\star}=\frac{p_k^{\mathrm{U}}(\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{\iota}_{nk}^{\mathrm{U}}}{\left(\sum_{l\neq k}p_l^{\mathrm{U}}(\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{\iota}_{nl}^{\mathrm{U}}(\boldsymbol{\iota}_{nl}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{U}}+\sigma_n^2(\boldsymbol{v}_{nk}^{\mathrm{U}})^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{U}}\right)}.\quad(31)$$

On the other hand, when $\boldsymbol{s}$ is fixed, (29) becomes a convex problem with respect to $\boldsymbol{v}_{nk}^{\mathrm{U}}$. This is because $Q_{nk}(\boldsymbol{v}_{nk}^{\mathrm{U}})$ is a concave function with respect to $\boldsymbol{v}_{nk}^{\mathrm{U}}$, and the composition of log and $Q_{nk}(\boldsymbol{v}_{nk}^{\mathrm{U}})$ preserves concavity. Since the subproblem for solving $\boldsymbol{v}_{nk}^{\mathrm{U}}$ is a convex optimization problem, it can be efficiently solved using the interior-point method.

D. Optimizing Variables $\{\theta_m^{\mathrm{U}}\}$

When it comes to the uplink RIS reflecting coefficients design, the subproblem for updating $\{\theta_m^{\mathrm{U}}\}$ is given by

$$\max_{\{\theta_m^{\mathrm{U}}\}}\quad\frac{1}{w}\sum_k\sum_n\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}\log(R_{nk}(\{\theta_m^{\mathrm{U}}\})-R_k),\quad(32)$$

where we substitute (13g) in the objective function. Since (32) is an unconstrained optimization problem with a differentiable objective function, it can be readily solved by the gradient

descent method. To be specific, the update of θ_m^U at the ith iteration is given by

$$
\theta_m^U(i+1) = \theta_m^U(i)
$$
$$
+ \mathcal{I}(i)\Big(\frac{1}{w}\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\frac{\nabla_{\theta_m^U}R_{nk}\left(\theta_m^U(i)\right)}{R_{nk}\left(\{\theta_m^U\}\right)-R_k}\Big), \quad (33)
$$

where $\mathcal{I}(i)$ is the step size chosen by the Armijo rule to guarantee convergence [44] and $\nabla_{\theta_m^U}R_{nk}\left(\theta_m^U(i)\right)$ is the gradient of $R_{nk}\left(\theta_m^U\right)$ at point $\theta_m^U(i)$ as given by (34). By updating $\{\theta_m^U\}$ based on (33), it is guaranteed to converge to a stationary point of the problem (32).

E. Optimizing Variables $\{a_k\}$

In terms of variable $\{a_k\}$, when other variables in P_{2w} are fixed, the subproblem for updating $\{a_k\}$ is

$$
\max_{\{a_k\}} \frac{\sum_{k=1}^{K}t_kR_k/T + \sum_{k=1}^{K}R_k^{loc}(a_k)}{\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\frac{(C_kt_kR_kf_{nk}\kappa_n)}{T}}
$$
$$
+ \frac{1}{w}\sum_k\sum_n\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\log(R_{nk}(\{a_k\})-R_k), \quad (35a)
$$

$$
\text{s.t. } a_k \in [0,1], \quad \forall k\in\mathcal{K}, \quad (35b)
$$
$$
\mathbf{1}_{\{a_k\neq0\}}\text{SINR}_k^D\left(\{v_{nk}^D\},\{\theta_m^D\}\right)\geq\mathbf{1}_{\{a_k\neq0\}}\gamma_k^D, \; k\in\mathcal{K}, \quad (35c)
$$
$$
\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\left(\frac{U_k-TR_k^{loc}}{R_k}+\frac{U_k-TR_k^{loc}}{f_{nk}}\right)\leq\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\chi,
$$
$$
k\in\mathcal{K}, n\in\mathcal{N}. \quad (35d)
$$

By identifying that (35c) is a weight ℓ_0-"norm" as introduced in IV-A, we equivalently transform it as

$$
\|a_k\|_2\,\text{SINR}_k^D\left(\{v_{nk}^D\},\{\theta_m^D\}\right)\geq\|a_k\|_2\,\gamma_k^D, \quad k\in\mathcal{K}. \quad (36)
$$

Next, as $\{a_k\}$ in $R_{nk}(\{a_k\})$ also possesses a fractional structure, akin to our treatment of $\{v_{nk}^U\}$, we also apply quadratic transformation:

$$
\max_{\{a_k\},o} \frac{\sum_{k=1}^{K}t_kR_k/T + \sum_{k=1}^{K}R_k^{loc}(a_k)}{\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\frac{(C_kt_kR_kf_{nk}\kappa_n)}{T}}
$$
$$
+ \frac{1}{w}\sum_k\sum_n\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\log(D_{nk}(\{a_k\})-R_k), \quad (37)
$$
$$
\text{s.t. } (35b),(35d),(36),
$$

where o refers to the collection of $\{o_{nk}\}$ and $D_{nk}(\{a_k\})$ is given as

$$
D_{nk}(\{a_k\}) = \log_2\Big(1 + 2o_{nk}\sqrt{a_kP_k^c}\left|(v_{nk}^U)^H\iota_{nk}^U\right|
$$
$$
- o_{nk}^2\Big(\sum_{l\neq k}a_lP_l^c\left|(v_{nk}^U)^H\iota_{nl}^U\right|^2+\sigma_n^2\|v_{nk}^U\|^2\Big)\Big). \quad (38)
$$

Similar to (29), when $\{a_k\}$ is held fixed, the optimal o_{nk} can be directly given by

$$
o_{nk}^\star = \frac{a_kP_k^c\left|(v_{nk}^U)^H\iota_{nk}^U\right|^2}{\sum_{l\neq k}a_lP_l^c\left|(v_{nk}^U)^H\iota_{nl}^U\right|^2+\sigma_n^2\left|v_{nk}^U\right|^2}. \quad (39)
$$

When o is fixed, (37) is a convex problem with respect to a_k, and therefore, the optimal solution can be efficiently found using the standard numerical method.

F. Optimizing Variables $\{R_k\}$ and $\{t_k\}$

Regarding the optimization of $\{R_k\}$ and $\{t_k\}$, the subproblem is given as

$$
\max_{\{R_k\},\{t_k\}} \frac{\sum_{k=1}^{K}t_kR_k/T + \sum_{k=1}^{K}R_k^{loc}}{\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\left(\frac{t_kR_k}{T}f_{nk}\kappa_nC_k\right)}
$$
$$
+ \underbrace{\frac{1}{w}\sum_k\sum_n\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\left(\log(R_{nk}-R_k)\right)}_{:=e(\{R_k\})}, \quad (40a)
$$

$$
\text{s.t. } t_k + (C_kt_kR_k)/f_k^{min}\leq T, \quad \forall k\in\mathcal{K}, \quad (40b)
$$
$$
(35d).
$$

Then applying Dinkelbach's transformation to separate the numerator and denominator, we have

$$
\max_{\{R_k\},\{t_k\},z} \frac{1}{w}e(\{R_k\}) + \sum_{k=1}^{K}\left(\frac{t_kR_k}{T}+R_k^{loc}\right)
$$
$$
- z\left(\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\left(\frac{t_kR_k}{T}f_{nk}\kappa_nC_k\right)\right), \quad (41)
$$
$$
\text{s.t. } (35d),(40b),
$$

where z is the auxiliary variable. Similarly, when $\{R_k\}$ and $\{t_k\}$ are fixed, the optimal z is given by the iterative update rule of the Dinklebach transform as

$$
z^\star = \frac{\sum_{k=1}^{K}t_kR_k/T + \sum_{k=1}^{K}R_k^{loc}}{\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\left(\frac{t_kR_k}{T}f_{nk}\kappa_nC_k\right)}. \quad (42)
$$

With other variables fixed, the subproblem for updating $\{R_k\}$ is given as

$$
\max_{\{R_k\}} \frac{1}{w}e(\{R_k\}) + \sum_{k=1}^{K}\left(\frac{t_kR_k}{T}+R_k^{loc}\right)
$$
$$
- z\left(\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}_{\{v_{nk}^D\neq\mathbf{0}\}}\left(\frac{t_kR_k}{T}f_{nk}\kappa_nC_k\right)\right), \quad (43)
$$
$$
\text{s.t. } (35d),(40b).
$$

The constraints (35d) can be linearized by multiplying both sides of the constraint by R_k. This transformation converts the optimization problem (43) into a convex one, which can then be solved using standard numerical methods.

Finally, regarding $\{t_k\}$, with other variables are fixed, the

$$\nabla_{\theta_m^{\mathrm{U}}} R_{nk}\left(\theta_m^{\mathrm{U}}(i)\right) = \mathrm{Re}\left(\frac{\psi_m(i)}{\left(1+\gamma_{nk}^{\mathrm{U}}\right)\ln 2}\left(\frac{2p_k^{\mathrm{U}}\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\boldsymbol{\iota}_{nk}^{\mathrm{U}}\left(\boldsymbol{K}_{nkk}\right)_{m,m}}{\sum_{l\neq k}p_l^{\mathrm{U}}\left|\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\boldsymbol{\iota}_{nl}^{\mathrm{U}}\right|^2+\sigma_n^2\left\|\boldsymbol{v}_{nk}^{\mathrm{U}}\right\|_2^2}\right.\right.$$
$$\left.\left.-\frac{p_k^{\mathrm{U}}\left|\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\boldsymbol{\iota}_{nk}^{\mathrm{U}}\right|^2\sum_{l\neq k}2p_l^{\mathrm{U}}\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\boldsymbol{\iota}_{nl}^{\mathrm{U}}\left(\boldsymbol{K}_{nkl}\right)_{m,m}}{\left(\sum_{l\neq k}p_l^{\mathrm{U}}\left|\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^{\mathrm{H}}\boldsymbol{\iota}_{nl}^{\mathrm{U}}\right|^2+\sigma_n^2\left\|\boldsymbol{v}_{nk}^{\mathrm{U}}\right\|_2^2\right)^2}\right)\right)$$
$$(34)$$

where $\boldsymbol{K}_{nkl} = \left(\boldsymbol{G}_n^{\mathrm{U}}\right)^*\left(\boldsymbol{v}_{nk}^{\mathrm{U}}\right)^*\left(\boldsymbol{h}_{\mathrm{r},l}^{\mathrm{U}}\right)^{\mathrm{T}}$ and $\psi_m(i) = je^{j\theta_m^{\mathrm{U}}(i)}\left(\beta_{\min}+2^{-\alpha}(1-\beta_{\min})\left(1+\sin\left(\theta_m^{\mathrm{U}}(i)-\phi\right)\right)^{\alpha}\right)+$
$\alpha 2^{-\alpha}e^{j\theta_m^{\mathrm{U}}(i)}(1-\beta_{\min})\left(1+\sin(\theta_m^{\mathrm{U}}(i)-\phi)\right)^{\alpha-1}\cos(\theta_m^{\mathrm{U}}(i)-\phi).$

subproblem for updating $\{t_k\}$ is given as

$$\max_{\{t_k\}}\ \sum_{k=1}^{K}\left(\frac{t_k R_k}{T}+R_k^{\mathrm{loc}}\right)$$
$$-z\left(\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbf{1}_{\{\boldsymbol{v}_{nk}^{\mathrm{D}}\neq\mathbf{0}\}}\left(\frac{t_k R_k}{T}f_{nk}\kappa_n C_k\right)\right),\quad(44)$$
$$\text{s.t.}\ (40\mathrm{b}).$$

It is important to note that the optimization problem (44) is a linear programming problem. Therefore, it can be solved using standard numerical methods.

V. DESIGNING $\{\theta_m^{\mathrm{D}}\}$ BEYOND A RANDOM FEASIBLE POINT IN P_{2w}

Unlike other variables in P_{2w}, $\{\theta_m^{\mathrm{D}}\}$ does not appear in the objective function of P_{2w}. This absence of dependence often leads to a misconception that any $\{\theta_m^{\mathrm{D}}\}$ satisfying constraint (13d) would be a valid solution, as has been done in previous works [45]–[47]. However, it is known that an optimized RIS together with properly designed beamforming vector $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$ at the APs can help to construct a high-quality channel link, thus reduce the downlink transmission power at APs while satisfying the SINR constraints (13d). This intuition contradicts to simply finding a feasible solution for (13d).

Recognizing that the variables $\{\theta_m^{\mathrm{D}}\}$ work cooperatively with $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$ to reduce the downlink transmission power, we resort to the optimization problem (21) that updates $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$. More insight can be revealed by noticing that $\{\theta_m^{\mathrm{D}}\}$ are hidden in $\boldsymbol{\Theta}^{\mathrm{D}}$ and is coupled with $\boldsymbol{v}_{nk}^{\mathrm{D}}$ in the SINR constraints (22) (an equivalent SOC form of (13d)). We can potentially obtain a better solution of (21) by adjusting $\{\theta_m^{\mathrm{D}}\}$ such that the active constraints in (22) are turned into inactive constraints[2] [48].

Motivated by the fact that the difference between the right-hand-side and the left-hand-side of equation (22) is bigger than zero if the constraint is inactive and equals zero if it is active, we compute the minimum differences between the right-hand-side and the left-hand-side of the equation (22) among the K users. Then, we tune the $\{\theta_m^{\mathrm{D}}\}$ such that the minimum difference is maximized, giving the following

maximin optimization problem:

$$\max_{\{\theta_m^{\mathrm{D}}\}}\min_{\substack{k\in\mathcal{K},\\a_k\neq 0}}\ \frac{1}{\sqrt{\gamma_k^{\mathrm{D}}}}\Re\left(\sum_{n\in\mathcal{N}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}}+\left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{D}}\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{D}}\right)$$
$$-\sqrt{\sum_{l\neq k}\left|\sum_{n\in\mathcal{N}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}}+\left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{D}}\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{v}_{nl}^{\mathrm{D}}\right|^2+\sigma_k^2},$$
$$(45\mathrm{a})$$
$$\text{s.t.}\quad\boldsymbol{\Theta}^{\mathrm{D}}=\mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}},\ldots,\rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}}\right),$$
$$\forall m\in\mathcal{M},\qquad(45\mathrm{b})$$

where we expressed $\boldsymbol{\iota}_{nk}^{\mathrm{D}}$ as $\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}}+\left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{D}}\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}$ according to (7).

By introducing a slack variable y, the above optimization problem can be recast as [49]

$$\max_{\{\theta_m^{\mathrm{D}}\},y}\ y,\qquad(46\mathrm{a})$$
$$\text{s.t.}\ \frac{1}{\sqrt{\gamma_k^{\mathrm{D}}}}\Re\left(\sum_{n\in\mathcal{N}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}}+\left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{D}}\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{v}_{nk}^{\mathrm{D}}\right)$$
$$-\sqrt{\sum_{l\neq k}\left|\sum_{n\in\mathcal{N}}\left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}}+\left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{D}}\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}}\right)^{\mathrm{H}}\boldsymbol{v}_{nl}^{\mathrm{D}}\right|^2+\sigma_k^2}\geq y,$$
$$\bigcap_{k\in\mathcal{K},a_k\neq 0}\{k\},$$
$$(46\mathrm{b})$$
$$\boldsymbol{\Theta}^{\mathrm{D}}=\mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}},\ldots,\rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}}\right),\quad\forall m\in\mathcal{M}.$$
$$(46\mathrm{c})$$

Furthermore, to deal with the non-convexity introduced by the practical phase shift model in (46c), we treat $\boldsymbol{\Theta}^{\mathrm{D}}$ as an auxiliary variable and resort to the penalty-based method by adding an equality constraint-related penalty term to the objective function of (46), yielding the following optimization problem

$$\max_{\{\theta_m^{\mathrm{D}}\},\boldsymbol{\Theta}^{\mathrm{D}},y}y$$
$$-\mu\left\|\boldsymbol{\Theta}^{\mathrm{D}}-\mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}},\ldots,\rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}}\right)\right\|_{\mathrm{F}}^2,$$
$$(47\mathrm{a})$$
$$\text{s.t.}\quad(46\mathrm{b}),$$

---

[2]If the equality holds, the constraint is active (tight), otherwise inactive.

Algorithm 1: Penalty-based Algorithm for Solving (45)

Input : $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$ and set $\mu > 0$.

while
$\left\| \boldsymbol{\Theta}^{\mathrm{D}} - \mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}}, \ldots, \rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}} \right) \right\|_{\mathrm{F}}^2 > \epsilon_1$
do

   while *The fractional increase of the objective value of (47) is above* ϵ_2 **do**

      Update θ_m^{D} by solving (48) with fixed $\boldsymbol{\Theta}^{\mathrm{D}}$ and y.

      Update $\boldsymbol{\Theta}^{\mathrm{D}}$ and y by solving (49) with fixed θ_m^{D}.

   end

   Update the penalty coefficient $\mu \leftarrow \varrho\mu$.

end

$$\left(\boldsymbol{\Theta}^{\mathrm{D}} \right)_{i,j} = 0, \quad \forall i \neq j, \tag{47b}$$

where $\mu > 0$ is a penalty parameter for controlling the constraint violation of (46c) [50] and constraints (47b) is to restrict the non-diagonal entries of $\boldsymbol{\Theta}^{\mathrm{D}}$ are to be zero. It is observed that $\{\theta_m^{\mathrm{D}}\}$ can be updated with fixed $\boldsymbol{\Theta}^{\mathrm{D}}$ and y. On the other hand, $\boldsymbol{\Theta}^{\mathrm{D}}$ and y can be updated in parallel when $\{\theta_m^{\mathrm{D}}\}$ are fixed. This motivates us to apply the AM method to solve (47) efficiently.

For any given $\boldsymbol{\Theta}^{\mathrm{D}}$ and y, the subproblem for optimizing $\{\theta_m^{\mathrm{D}}\}$ is

$$\max_{\{\theta_m^{\mathrm{D}}\}} \; -\mu \left\| \boldsymbol{\Theta}^{\mathrm{D}} - \mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}}, \ldots, \rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}} \right) \right\|_{\mathrm{F}}^2, \tag{48}$$

which is an unconstrained optimization problem with a differentiable objective function, and it can be readily solved by the gradient descent method. On the other hand, for any given $\{\theta_m^{\mathrm{D}}\}$, the subproblem for optimizing $\boldsymbol{\Theta}^{\mathrm{D}}$ and y is

$$\max_{\boldsymbol{\Theta}^{\mathrm{D}}, y} \; y - \mu \left\| \boldsymbol{\Theta}^{\mathrm{D}} - \mathrm{diag}\left(\rho(\theta_1^{\mathrm{D}})e^{j\theta_1^{\mathrm{D}}}, \ldots, \rho(\theta_M^{\mathrm{D}})e^{j\theta_M^{\mathrm{D}}} \right) \right\|_{\mathrm{F}}^2,$$
$$\text{s.t.} \quad (46b), (47b). \tag{49}$$

Noticing that (46b) are convex constraints since they are SOC constraints that composite with affine function. This makes the optimization problem (49) a convex problem and can be directly solved via the interior-point method. The entire procedure for solving (45) is summarized in Algorithm 2, where we employ an increasing penalty strategy to successively enforce the constraint of (45b). Initially, μ is set to be a sufficiently small value. By gradually increasing the value of μ by a factor $\varrho > 1$, we can maximize the original objective function and obtain a solution that satisfies the equality constraints in (45b) within a predefined accuracy.

VI. FINDING INITIAL FEASIBLE POINTS AND COMPLEXITY ANALYSIS

Sections IV and V describe how to solve P_{2w} for a fixed w. To enforce the log barrier, the penalty weight w should be increased progressively after P_{2w} is solved with a particular w.

This leads to Algorithm 2 for solving the overall problem P_2, with the last line corresponding to the increase of w. Once we solve P_{2w} for a particular w, the solution would act as the initialization for the next P_{2w} with a larger w. However, a natural question arises regarding the initialization for P_{2w} for the first selected w. Since a feasible point for P_2 is also a feasible point for P_{2w} with any w, the following section will find a feasible point of P_2.

A. Finding Initial Feasible Points

Given the complicated constraints of P_2, it is not easy to find a feasible point. Inspired by [51], [52], we propose an effective heuristic method to address this challenge. The idea is to address a modified feasibility problem. This involves allowing violations of the selected constraints in the original problem while penalizing them within the objective function of the modified feasibility problem. Subsequently, these violations gradually decrease through the optimization process and fade away, ultimately leading to a feasible solution for the original problem. Particularly, let us consider the modification of the feasibility problem of P_2 as in (50), where we penalized the downlink SINR constraints (13d). Maximizing the objective function in (50) enhances downlink SINR and reduces constraint violations in (13d). If the objective function value of (50) surpasses zero, it signifies the violations disappear, and the solution of (50) is a feasible solution for P_2.

Solving (50) involves a straightforward process. Initially, we generate random values for $\{\theta_m^{\mathrm{U}}\}$, $\{\theta_m^{\mathrm{D}}\}$, and $\{a_k\}$. Following this, we generate $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$ and adjust their scaling to satisfy the conditions in (13e). By utilizing $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$, $\{\theta_m^{\mathrm{D}}\}$, $\{\theta_m^{\mathrm{U}}\}$ and $\{a_k\}$, $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}$, $\{R_k\}$, $\{f_{nk}\}$, $\{f_k^{\min}\}$ and $\{t_k\}$ are determined to adhere to the equalities stipulated in equations (13i), (14b), (15), (16), (17) and (18b). With this initial set of optimization variables established, we can initiate an iterative AM procedure to tackle the optimization problem (50). To begin with, when the other variables are held fixed, each individual variable in (50) exhibits convex behavior in the objective function, which is a summation of SOCs. Moreover, concerning the nonconvex constraints, the strategies for transforming them into convex constraints have been given in Sections IV-A through V. As a result, all subproblems of the variables can be transformed into convex forms and subsequently solved. This enables the utilization of the AM procedure to solve (50) efficiently, and the details of these steps are omitted for brevity.

B. Complexity Analysis of the Proposed Algorithm

The computational complexity of Algorithm 2 primarily comes from solving problems (28) and (21), which aim to design $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}$ and $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$, respectively, as well as the downlink RIS reflecting coefficients $\{\theta_m^{\mathrm{D}}\}$, and problem (35) for optimizing the power partition parameters $\{a_k\}$. The complexity associated with resolving (28) and (21) is estimated to be $\mathcal{O}((NKL)^{3.5})$, based on the complexities of solving convex problems using the interior point method [53]. Similarly, the task of optimizing $\{a_k\}$ is estimated to have a computational complexity of $\mathcal{O}(K^{3.5})$, following the same computational approach. The complexity of employing Algorithm 1 via the

$$
\max_{\substack{\{\boldsymbol{v}_{nk}^{\mathrm{U}}\},\{\boldsymbol{v}_{nk}^{\mathrm{D}}\},\{R_k\},\{f_{nk}\}, \\ \{\theta_m^{\mathrm{U}}\},\{\theta_m^{\mathrm{D}}\},\{a_k\},\{f_k^{\min}\},\{t_k\}}} \sum_{k\in\mathcal{K}} \|a_k\|_2 \left[\frac{1}{\sqrt{\gamma_k^{\mathrm{D}}}} \Re \left(\sum_{n\in\mathcal{N}} \left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}} \right)^{\mathrm{H}} \boldsymbol{v}_{nk}^{\mathrm{D}} \right) \right.
$$
$$
\left. - \sqrt{ \sum_{l\neq k} \left| \sum_{n\in\mathcal{N}} \left(\boldsymbol{h}_{\mathrm{d},nk}^{\mathrm{D}} + \left(\boldsymbol{G}_n^{\mathrm{D}}\right)^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{D}} \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}} \right)^{\mathrm{H}} \boldsymbol{v}_{nl}^{\mathrm{D}} \right|^2 + \sigma_k^2 } \, \right]^+, \tag{50}
$$
$$
\text{s.t.} \quad (13\mathrm{b}), (13\mathrm{e}), (13\mathrm{g}), (14\mathrm{b}), (13\mathrm{i}), (15), (16), (17), (18\mathrm{b})
$$

Algorithm 2: Algorithm for Solving P_2

Input : Initial $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$, $\{a_k\}$,$\{R_k\}$, $\{t_k\}$, $\{\theta_m^{\mathrm{U}}\}$, $\{\theta_m^{\mathrm{D}}\}$, $\{f_{nk}\}$,$\{f_k^{\min}\}$ and $w > 0$.

while *Stopping criterion is not satisfied* **do**
    while *Stopping criterion is not satisfied* **do**
        Update $\boldsymbol{s}$ using equation (31).
        Update $\{\boldsymbol{v}_{nk}^{\mathrm{U}}\}$ by solving (29) with fixed $\boldsymbol{s}$.
    end
    while *Stopping criterion is not satisfied* **do**
        Update $\boldsymbol{o}$ using equation (39).
        Update $\{a_k\}$ by solving (37) with fixed $\boldsymbol{o}$.
    end
    while *Stopping criterion is not satisfied* **do**
        Update $\{\theta_m^{\mathrm{U}}\}$ using equation (33).
    end
    Update $\{\boldsymbol{v}_{nk}^{\mathrm{D}}\}$ by solving (26).
    Update $\{f_{nk}\}$ and $\{f_k^{\min}\}$ by solving (27).
    while *Stopping criterion is not satisfied* **do**
        Update z using equation (42).
        Update $\{R_k\}$ by solving (44) with fixed z and $\{t_k\}$.
        Update $\{t_k\}$ by solving (43) with fixed z and $\{R_k\}$.
    end
    Update $\{\theta_m^{\mathrm{D}}\}$ by Algorithm 1.
    Update the penalty coefficient $w \leftarrow \vartheta w$.
end

---

penalty-based method is approximated as $\mathcal{O}(M^{3.5})$. If we denote the number of iterations for solving problems (28), (21), and (35) as I_1, I_2, and I_3, respectively, and those for Algorithm 1 as I_4, then the total computational complexity is given as $\mathcal{O}(I((I_1+I_2)(NKL)^{3.5}+I_3K^{3.5}+I_4M^{3.5}))$, where I represents the total number of iterations within Algorithm 2.

VII. SIMULATION RESULTS

In this section, we present simulation results to verify the effectiveness of our proposed algorithm. Under a three-dimensional Cartesian coordinate system, we consider a system with 10 APs and 20 users uniformly and randomly distributed in a square region of 200 m $\times$ 200 m. An RIS with 20 reflecting elements is located at the 3-dimensional coordinate $(100, 0, 15)$. In addition, the APs are with height 30 m, while the users are with height 1 m.

Rician fading channel model is considered for all channels to account for both the line-of-sight (LoS) and non-LoS (NLoS) components [54]. For example, the AP-RIS channel can be expressed as $\boldsymbol{G}_n^x = \sqrt{L_{\mathrm{AR}}(d)}\left(\sqrt{\frac{\kappa_{\mathrm{AR}}}{1+\kappa_{\mathrm{AR}}}}\boldsymbol{G}_n^{\mathrm{LoS}} + \sqrt{\frac{1}{1+\kappa_{\mathrm{AR}}}}\boldsymbol{G}_n^{\mathrm{NLoS}}\right)$, where κ_{AR} is the Rician factor representing the ratio of power between the LoS path and the scattered paths, $\boldsymbol{G}_n^{\mathrm{LoS}}$ is the LoS component modeled as the product of the unit spatial signature of the AP-RIS link [54], $\boldsymbol{G}_n^{\mathrm{NLoS}}$ is the Rayleigh fading components with entries distributed as $\mathcal{CN}(0,1)$, $L_{\mathrm{AR}}(d)$ is the distance-dependent path loss of the AP-RIS channel, and $x \in \{\mathrm{U}, \mathrm{D}\}$. We consider the following distance-dependent path loss model $L_{\mathrm{AR}}(d) = 10^{0.3}E_0\left(\frac{d}{d_0}\right)^{-\alpha_{\mathrm{AR}}}$, where E_0 is the constant path loss at the reference distance $d_0 = 1$ m, d is the Euclidean distance between the transceivers, α_{AR} is the path loss exponent, and $10^{0.3}$ accounts for a 3 dBi gain at each element of the RIS since it reflects signal only in its front half-space [55]. Since the RIS can be practically deployed in LoS with the AP, we set $\alpha_{\mathrm{AR}} = 2$ and $\kappa_{\mathrm{AR}} = 30$ dB [45]. In addition, other channels are similarly generated with $\alpha_{\mathrm{AU}} = 3.67$ and $\kappa_{\mathrm{AU}} = 0$ (i.e., Rayleigh fading to account for rich scattering) for the AP-user channel, $\alpha_{\mathrm{RU}} = 2.5$ and $\kappa_{\mathrm{RU}} = 3$ for the RIS-user channel. We consider a system with a bandwidth 10 MHz and $E_0 = -30$ dB. The effective noise power for the APs and users are $\sigma_n^2 = -60$ dBm and $\sigma_k^2 = -50$ dBm, respectively. Unless specified otherwise, other parameters are set as follows: $P_{n,\max}^{\mathrm{D}} = 1$ W, $P_k^c = 0.5$ W, $C_k = 200$ cycles/bit, $\kappa_k = \kappa_n = 10^{-25}$, $f_n = 1.2 \times 10^9$ bits/s, $T = 0.5$ s, $\chi = 0.4$ s and $U_k = 350$ Kb [4], [56]. For the proposed algorithm, we set $\mu = 10^{-3}$, $\varrho = 1.003$, $\epsilon_1 = 0.1$, $\epsilon_2 = 0.01$, and the convergence threshold in terms of the relative increment in the objective value as 10^{-3}. The simulation results are obtained by averaging over 100 simulation trials.

We compare the proposed algorithm with the following benchmarks.

- Without-RIS: Without the deployment of an RIS, the equivalent channels in (2) and (7) contain only the direct link, i.e., $\boldsymbol{h}_{\mathrm{r},k}^{\mathrm{U}} = \boldsymbol{h}_{\mathrm{r},k}^{\mathrm{D}} = \boldsymbol{0}$, $\forall k$.
- Without cooperative transmission (denoted as Without-CT): In this case, we assume that each user is served by only one AP. For each user, the AP with the best channel condition is selected.
- Without RIS and cooperative transmission (denoted as Without-RIS-CT): In this case, we assume that each user is served by only one AP without the assistance of an RIS.

- Proposed algorithm based on solving Feasibility Problem of $\{\theta_m^D\}$ (denoted as AM-FP): Instead of optimizing $\{\theta_m^D\}$ by our algorithm, in this case, we only find a feasible solution for $\{\theta_m^D\}$ that satisfied (13d). This benchmark is designed to reveal the necessity of the proposed criterion for optimizing $\{\theta_m^D\}$ in (45).
- Proposed algorithm without l_{12}-norm (denoted as AM-Exhaustive Search (AM-ES)): In this case, we do not eliminate the variable $\{A_k\}$ by exploiting the group sparsity structure of the beamforming vectors but search over all the possibilities of $\{A_k\}$. This benchmark is designed to reveal the efficiency of our proposed algorithm.

We first study the relationship between the feasibility ratio [46] of the problem (13) and the target SINR γ_k^D with different benchmarks. The feasibility ratio of the problem (13) is defined as $\frac{\text{number of feasible simulations for (13)}}{\text{the total number of simulations}}$. As the target SINR requirements become more stringent, i.e., larger values of γ_k^D, the feasibility ratio of the problem (13) is expected to decline. As shown in Fig. 2a, we observe that Without-RIS-CT and Without-CT almost fail to maintain feasibility in those settings with a target SINR higher than 10 dB, while Without-RIS, AM-FP, and proposed algorithm can still maintain feasibility with a high probability. Comparing cooperative transmission and RIS, we observed that including cooperative transmission leads to a better feasibility ratio than including an RIS. This can be seen from the fact that Without-RIS can provide more than a 10 dB gain compared with Without-RIS-CT, while the improvement of Without-CT from Without-RIS-CT is less than 3 dB. Obviously, including both RIS and cooperative transmission (proposed algorithm and AM-FP) has the highest feasibility ratio, with the proposed algorithm having an even higher feasibility ratio than the AM-FP. This shows the importance of optimizing $\{\theta_m^D\}$, rather than simply finding a feasible solution for it.

Next, the superiority of our proposed algorithm in terms of computation efficiency is shown in Fig. 2b. Under a wide range of SINR requirements, it is observed that the system without cooperative transmission achieves much worse computation efficiency than other schemes. The Without-RIS scheme has a 40% performance improvement compared with the Without-RIS-CT scheme. A similar amount of gain can also be observed when comparing AM-FP and Without-CT. The feasibility ratio and computation efficiency results demonstrate that cooperative transmission in wireless communication systems can dramatically boost the SINR and, in turn, increase the overall computation efficiency. For the schemes with an RIS, the proposed algorithm provides at least a 14% improvement in computation efficiency over AM-FP. Such a performance gain comes from the fact that the AM-FP only finds a feasible solution of $\{\theta_m^D\}$, while the proposed algorithm optimizes $\{\theta_m^D\}$ to improve the computation efficiency further. To make this more explicit, Fig. 3 shows the achieved computation efficiency in the first 16 iterations of the proposed algorithm and AM-FP under a specific channel realization. It is observed that the proposed algorithm outperforms AM-FP in every iteration. This shows the proposed algorithm's effectiveness in inactivating the active constraints that impede an increase
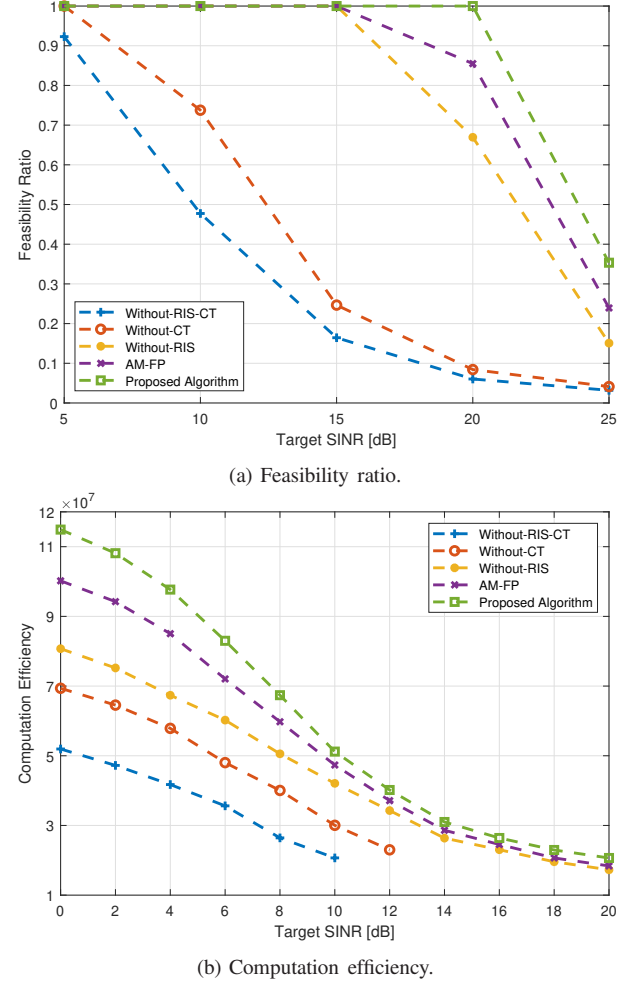

(a) Feasibility ratio.

(b) Computation efficiency.

Fig. 2. Performance comparison versus target SINR γ_k^D.

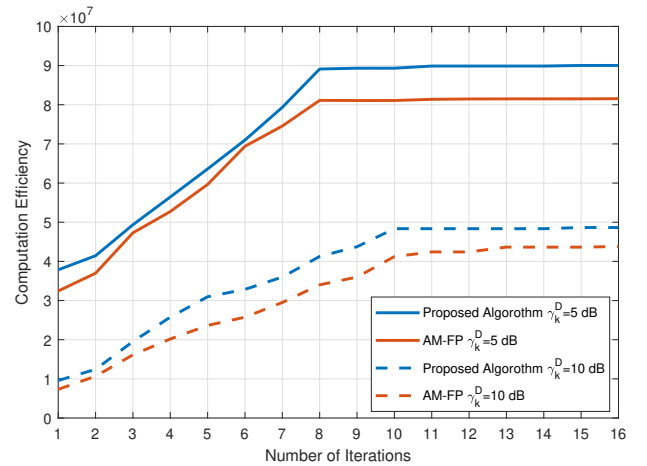in the objective value, as discussed in Section V.

Fig. 3. Convergence behaviors of both proposed algorithm and AM-FP.

To further show the complexity advantage of the proposed algorithm, we compare it with AM-ES. As shown in Fig. 4a and Fig. 4b, with the number of APs and users increase, the proposed algorithm saves the computation time to a large

extent compared with the AM-ES (e.g., more than 10 times differences with 8 APs and 12 users, in Fig. 4a and Fig. 4b, respectively), and the advantage becomes more prominent as K or N increases. On the other hand, Fig. 4c and Fig. 4d show that the proposed algorithm achieves almost the same computation efficiency as the AM-ES, and the performance gap is less than 5%. It is also noted that in Fig. 4c and 4d, the total computation efficiency increases as the number of users and APs increases, but the increase diminishes. If we divide the total computation efficiency by the number of users and APs, the average computation efficiency will decrease as the number of APs and users increases. The diminishing returns on average computation efficiency suggest that the number of APs and users in the proposed system need not be arbitrarily large to obtain a favorable computation efficiency. A similar phenomenon on the system's efficiency is also exhibited in [57], [58].

In Figure 5, we present a comparison of the sparsity structure of the downlink beamformer obtained by the proposed method and the AM-ES algorithm. Each subfigure represents the downlink beamforming vectors of a specific AP, where the user index is on the horizontal axis and the antenna index is on the vertical axis. The heatmap is generated by taking the absolute value of the beamforming vectors in a particular channel realization, and the color intensity shows the magnitude of the absolute value. For brevity, we only illustrate the results for 6 APs. It can be observed that both methods yield a sparse structure in the downlink beamforming vectors, which aligns with the intuition described in Section III. Furthermore, the solution of AM-ES has a more sparse structure than our proposed method since it reflects the optimal sparsity structure by enumerating the associations of APs and users, while our proposed method induces the sparsity structure by the group norm. The above observation explains the performance gap between our proposed method and AM-ES.

To visualize the division of computation load between offloading and local computation, we investigate how the optimal power partition parameter varies with the distance between the user and RIS in Fig. 6. In our scenario, we consider two users and two APs, with one user located at a fixed distance of 20 m from the RIS and the other from 20 to 60 m away. For the user with high power budgets ($P_k^c = 1$ or 0.5 W), we observe that the optimal strategy is to allocate more power to computation offloading as the distance increases. This is because computation offloading can provide a higher computation efficiency than local computation, and as the distance increases, the user needs to spend more power on data offloading. In contrast, for a user with a low power budget ($P_k^c = 0.3$ W), the optimal power partition parameter first increases as the distance increases. However, as it reaches the maximum value of 1, it starts to decline as the distance increases. The rise in the optimal power partition parameter at the first segment is due to the same reason as users with a high power budget. However, as the distance increases further, a limited power budget cannot support high-speed data transmission in computation offloading, and computation offloading becomes less computation efficient than local

computing. Thus, we observe a decline in the optimal power partition parameter at the second segment. As for the case of $P_k^c = 0.1$ W, the optimal power partition parameter declines as the distance increases. The simulation results demonstrate that the optimal power allocation strategy for maximizing the computation efficiency depends on the user's power budget and the distance between the RIS and the user. In general, the optimal power partition parameter first increases as the distance increases before reaching its maximum value of 1 and then decreases as the distance increases. The optimal power partition parameter curve will shift to the right if the power budget is abundant, while the curve will shift to the left if the budget is small.

In Fig. 7, we show how the number of APs (N), users (K), and RIS's reflecting elements (M) would affect the average number of APs serving each user in user association. In Fig. 7(a), it is evident that as the number of users increases, there is an upward trend in the average number of APs serving each user. This can be attributed to the increased downlink interference caused by more users. Therefore, increasing the number of APs per user can reduce downlink power consumption and enhance the overall system computation efficiency. In Fig. 7(b), we can observe that the increasing number of reflecting elements reduces the average number of APs serving each user. This is because the increased elements of RIS improved the channel condition, allowing the system to satisfy the user's downlink QoS with fewer associated APs. In Fig. 7(c), as the number of APs increases, the average number of APs serving each user increases from 2.45 to 3.4 initially, then levels off after reaching 25 APs. This phenomenon occurs because having more APs per user reduces the downlink's power consumption at first. However, as the number of APs per user keeps growing, this power-saving effect becomes less pronounced, and it is outweighed by the power consumption resulting from the additional computation load of extra APs. As a result, the proposed algorithm would not further increase the association of APs and users.

Finally, in Fig. 8, we compare the computation efficiency versus the number of reflecting elements M. As M increases, the computation efficiency increases moderately in both the proposed algorithm and AM-FP. It is worth mentioning that the performance gaps between the proposed algorithm and AM-FP are getting more prominent as the number of reflecting elements increases, indicating that the proposed algorithm is especially appealing when the RIS is equipped with a large number of elements.

VIII. CONCLUSION

In this paper, an RIS-aided mobile edge computing system with a cooperative transmission framework was proposed. Specifically, the computation efficiency was maximized via the joint design of user association, receive/downlink beamforming vectors, power partition parameters, and uplink/downlink phase-shift matrices. For an efficient algorithm design, the alternating maximization framework was employed. To handle the discrete user association variables, group l_{12}-norm was adopted to enforce group sparsity and merge user association
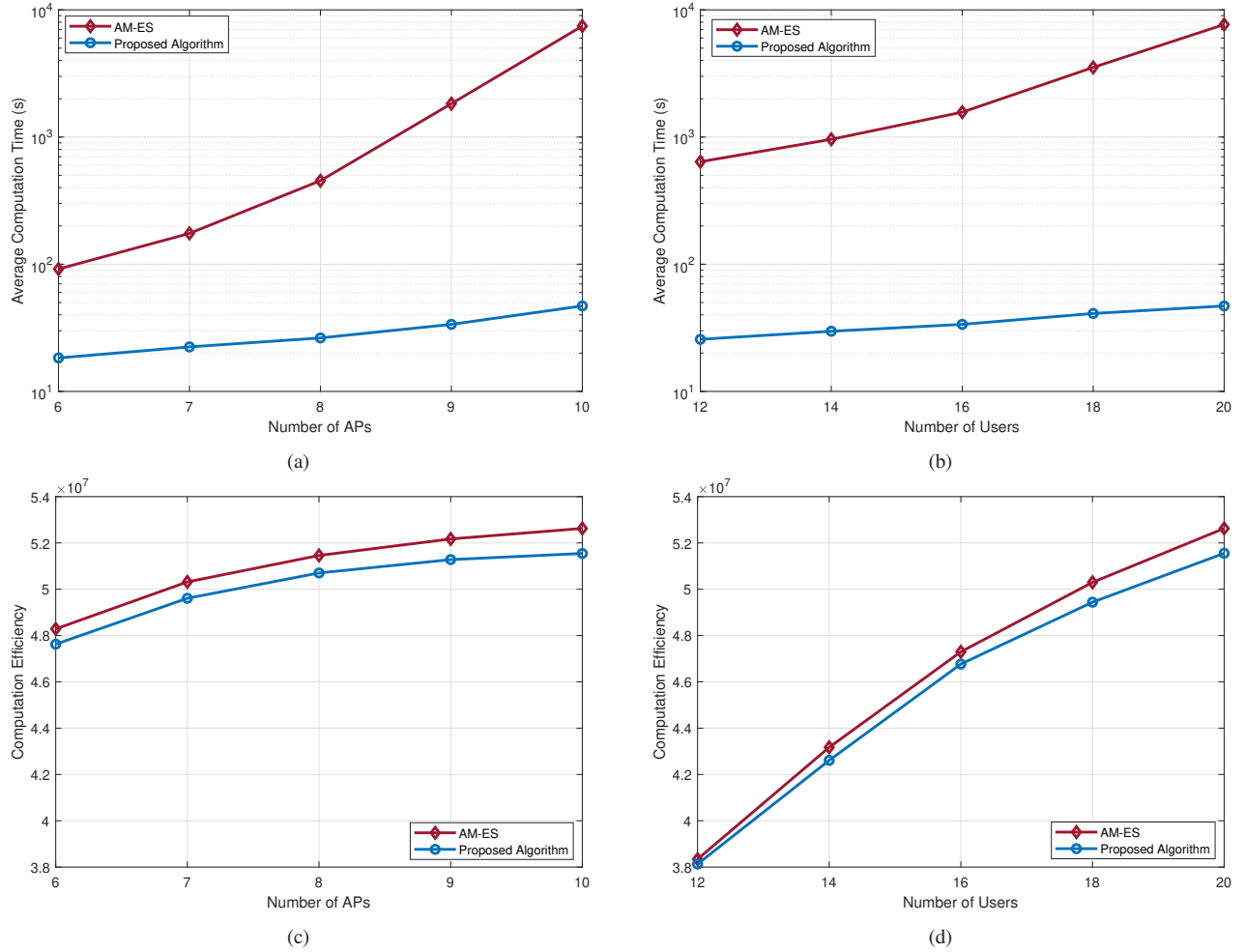
Fig. 4. Performance comparison with the AM-ES with target SINR equals 10 dB. The number of users in (a) and (c) is 20, while the number of APs in (b) and (d) is 10.

into the receive/downlink beamforming vectors. Furthermore, although the objective function does not explicitly depend on the downlink phase-shift matrix, we leverage their hidden relationship to convert it into an explicit form for optimization. This approach enables us to fully exploit the potential of the RIS rather than simply finding a feasible solution of the downlink phase-shift matrix that could be more optimal. Numerical results demonstrated that cooperative transmission and RIS could significantly improve the computation efficiency and feasibility ratio. At the same time, the proposed approach for optimizing downlink phase shifts can further promote this improvement.

REFERENCES

[1] 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, Fourth Quart. 2017.

[2] Y. Qi, Y. Zhou, Y.-F. Liu, L. Liu, and Z. Pan, "Traffic-aware task offloading based on convergence of communication and sensing in vehicular edge computing," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17 762–17 777, Dec. 2021.

[3] Y. Peng, X. Tang, Y. Zhou, J. Li, Y. Qi, L. Liu, and H. Lin, "Computing and communication cost-aware service migration enabled by transfer reinforcement learning for dynamic vehicular edge computing networks," *IEEE Trans. Mob. Comput.*, vol. 23, no. 1, pp. 257–269, Jan. 2024.

[4] X. Hu, C. Masouros, and K.-K. Wong, "Reconfigurable intelligent surface aided mobile edge computing: From optimization-based to location-only learning-based solutions," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3709–3725, Jun. 2021.

[5] F. Zhou and R. Q. Hu, "Computation efficiency maximization in wireless-powered mobile edge computing networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3170–3184, May 2020.

[6] H. Chen, D. Zhao, Q. Chen, and R. Chai, "Joint computation offloading and radio resource allocations in small-cell wireless cellular networks," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 3, pp. 745–758, Sep. 2020.

[7] Z. Chu, P. Xiao, M. Shojafar, D. Mi, J. Mao, and W. Hao, "Intelligent reflecting surface assisted mobile edge computing for internet of things," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 619–623, Mar. 2021.

[8] H. Xie, M. Xia, P. Wu, S. Wang, and H. V. Poor, "Edge learning for large-scale internet of things with task-oriented efficient communication," *IEEE Trans. Wireless Commun.*, pp. 1–1, Early Access 2023.

[9] B. L. Ng, J. S. Evans, S. V. Hanly, and D. Aktas, "Distributed downlink beamforming with cooperative base stations," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5491–5499, Dec. 2008.

[10] L. Liu, Y. Zhou, W. Zhuang, J. Yuan, and L. Tian, "Tractable coverage analysis for hexagonal macrocell-based heterogeneous udns with adaptive interference-aware comp," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 503–517, Jan. 2019.

[11] L. Liu, Y. Zhou, V. Garcia, L. Tian, and J. Shi, "Load aware joint comp clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2741–2755, Mar. 2018.

[12] Q. Cai, Y. Zhou, L. Liu, Y. Qi, Z. Pan, and H. Zhang, "Collaboration of heterogeneous edge computing paradigms: How to fill the gap between
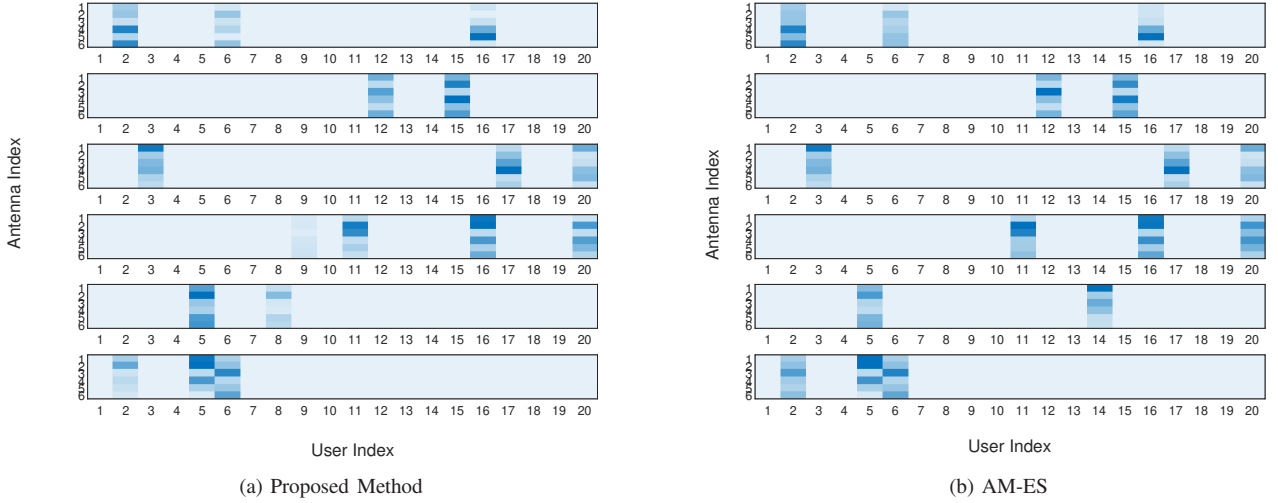
(a) Proposed Method

(b) AM-ES

Fig. 5. Comparison of a heatmap of downlink beamformers with the AM-ES. The number of antennas at APs and the number of users are 6 and 20, respectively.
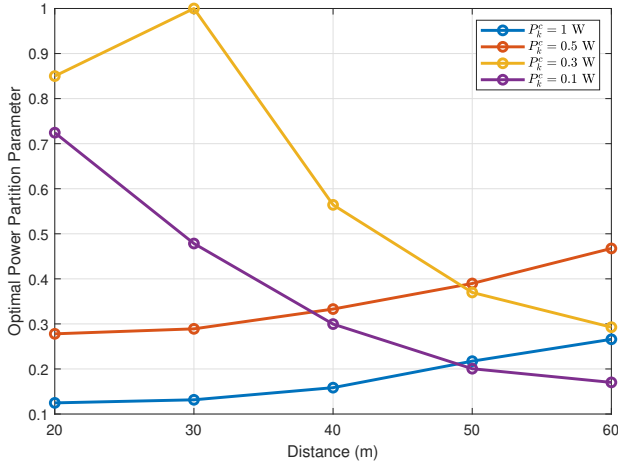

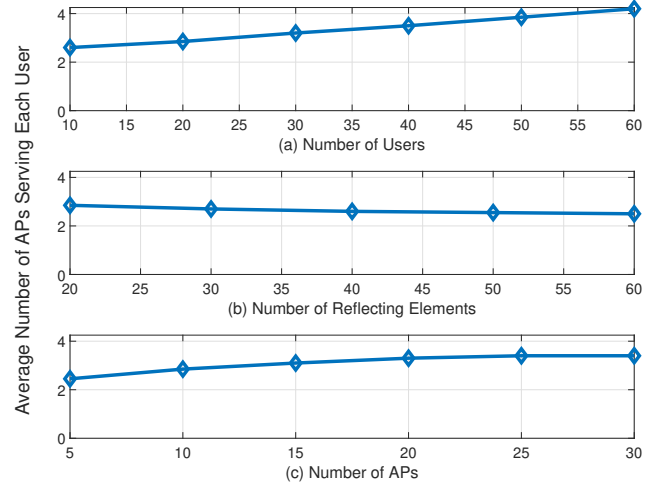Fig. 6. Optimal power partition parameter versus distance.

Fig. 7. Average number of APs serving each user under different numbers of APs, users, and reflecting elements, with 10 APs in (a) and (b), 20 users in (b) and (c), and 20 reflecting elements in (a) and (c).

theory and practice," *IEEE Wireless Commun.*, pp. 1–9, Early Access 2023.

[13] J. T. Chapman, J. Andreoli-Fang, M. Chauvin, E. C. Reyes, Z. Lu, D. Liu, J. Padden, and A. Bernstein, "Low latency techniques for mobile backhaul over DOCSIS," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.

[14] L. Su, C. Yang, and S. Han, "The value of channel prediction in CoMP systems with large backhaul latency," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4577–4590, Nov. 2013.

[15] K. Raaen and I. Kjellmo, "Measuring latency in virtual reality systems," in *Proc. Int. Conf. Entertainment Comput.*, 2015, pp. 457–462.

[16] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016.

[17] Z. Chu, P. Xiao, M. Shojafar, D. Mi, W. Hao, J. Shi, and F. Zhou, "Utility maximization for IRS assisted wireless powered mobile edge computing and caching (WP-MECC) networks," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 457–472, Jan. 2023.

[18] 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.

[19] 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, May 2020.

[20] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

[21] X. Pei, H. Yin, L. Tan, L. Cao, Z. Li, K. Wang, K. Zhang, and E. Björnson, "RIS-Aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8627–8640, Dec. 2021.

[22] "5G NR testbed 3.5 GHz coverage results," https://www.ericsson.com/en/reports-and-papers/research-papers/5g-nr-testbed-3.5-ghz-coverage-results, 2018, accessed: August 2, 2023.

[23] "White paper: 5G NR millimeter wave network coverage simulation," https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/5g_nr_millimeter_wave_network_coverage_simulation_studies_for_global_cities.pdf, 2017, accessed: August 2, 2023.

[24] S. Hua, Y. Zhou, K. Yang, Y. Shi, and K. Wang, "Reconfigurable intelligent surface for green edge inference," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 2, pp. 964–979, 2021.
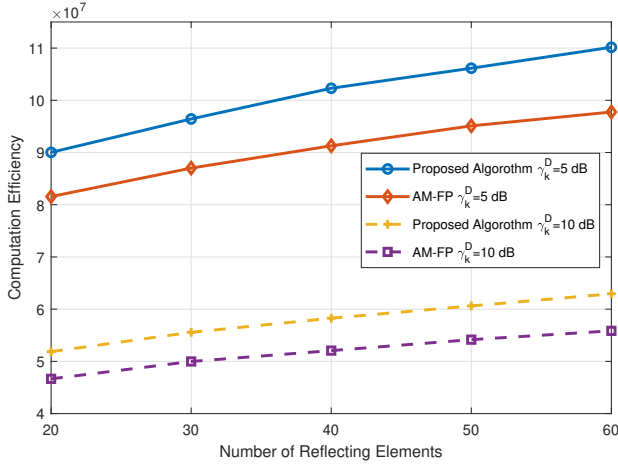
Fig. 8. Computation efficiency versus the number of reflecting elements.

[25] W. He, D. He, X. Ma, X. Chen, Y. Fang, and W. Zhang, "Joint user association, resource allocation, and beamforming in ris-assisted multi-server mec systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, Early Access 2023.

[26] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, May 2022.

[27] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.

[28] S. Bi and Y. J. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.

[29] Y. Ye, L. Shi, X. Chu, R. Q. Hu, and G. Lu, "Resource allocation in backscatter-assisted wireless powered MEC networks with limited MEC computation capacity," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10 678–10 694, Dec. 2022.

[30] 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, Mar. 2020.

[31] L. Lovasz, "Randomized algorithms in combinatorial optimization," *Combinatorial Optimization, volume 20 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 20, pp. 153–179, 1995.

[32] F. Neumann, "Combinatorial optimization and the analysis of randomized search heuristics," Ph.D. dissertation, Christian-Albrechts Universität Kiel, 2006.

[33] C. Park and J. Lee, "Mobile edge computing-enabled heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1038–1051, Feb. 2021.

[34] Y. Qu, H. Dai, F. Wu, D. Lu, C. Dong, S. Tang, and G. Chen, "Robust offloading scheduling for mobile edge computing," *IEEE Trans. Mob. Comput.*, vol. 21, no. 7, pp. 2581–2595, Jul. 2022.

[35] 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.

[36] E. Visotsky and U. Madhow, "Optimum beamforming using transmit antenna arrays," in *Proc. IEEE Veh. Technol. Conf*, vol. 1, May 1999, pp. 851–856.

[37] L. Du, S. Shao, G. Yang, J. Ma, Q. Liang, and Y. Tang, "Capacity characterization for reconfigurable intelligent surfaces assisted multiple-antenna multicast," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6940–6953, Oct. 2021.

[38] Z. Liu, Z. Li, M. Wen, Y. Gong, and Y.-C. Wu, "STAR-RIS-aided mobile edge computing: Computation rate maximization with binary amplitude coefficients," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 4313–4327, Jul. 2023.

[39] A. Andresen and V. Spokoiny, "Convergence of an alternating maximiza-

[40] tion procedure," *The J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2229–2281, Apr. 2016.

[40] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[41] S. Akila, "l_0 sparse signal processing and model selection with applications," Ph.D. dissertation, UNSW Sydney, 2012.

[42] I. Pólik and T. Terlaky, "Interior point methods for nonlinear optimization," in *Nonlinear optimization*. Springer, Berlin, Heidelberg, 2010, pp. 215–276.

[43] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[44] Z. Li, S. Wang, Q. Lin, Y. Li, M. Wen, Y.-C. Wu, and H. V. Poor, "Phase shift design in RIS empowered wireless networks: from optimization to AI-based methods," *Network*, vol. 2, no. 3, pp. 398–418, 2022.

[45] 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, 2019.

[46] 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, Nov. 2019.

[47] Y. Chen, M. Wen, E. Basar, Y.-C. Wu, L. Wang, and W. Liu, "Exploiting reconfigurable intelligent surfaces in edge caching: Joint hybrid beamforming and content placement optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7799–7812, Dec. 2021.

[48] M. T. Heath, *Scientific computing: an introductory survey, revised second edition*. Philadelphia, PA, USA: SIAM, 2018.

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

[50] Y. Li, M. Xia, and Y.-C. Wu, "First-order algorithm for content-centric sparse multicast beamforming in large-scale C-RAN," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5959–5974, Sep. 2018.

[51] T. Lipp and S. Boyd, "Variations and extension of the convex-concave procedure," *Optim. Eng.*, vol. 17, pp. 263–287, 2016.

[52] K.-G. Nguyen, Q.-D. Vu, L.-N. Tran, and M. Juntti, "Energy efficiency fairness for multi-pair wireless-powered relaying systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 357–373, Feb. 2019.

[53] Z.-q. Luo, W.-k. Ma, A. M.-c. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[54] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[55] 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.

[56] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2666–2682, Nov. 2020.

[57] R. Pinto Antonioli, I. M. Braga, G. Fodor, Y. C. B. Silva, A. L. F. de Almeida, and W. C. Freitas, "On the energy efficiency of Cell-Free systems with limited fronthauls: Is coherent transmission always the best alternative?" *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8729–8743, Oct. 2022.

[58] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of Cell-Free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.