Exploiting Reconfigurable Intelligent Surfaces in Edge Caching: Joint Hybrid Beamforming and Content Placement Optimization

Yingyang Chen, *Member, IEEE*, Miaowen Wen, *Senior Member, IEEE*, Ertugrul Basar, *Senior Member, IEEE*, Yik-Chung Wu, *Senior Member, IEEE*, Li Wang, *Senior Member, IEEE*, and Weiping Liu, *Member, IEEE*

***Abstract*— Edge caching can effectively reduce backhaul burden at core network and increase quality-of-service at wireless edge nodes. However, the beneficial role of edge caching cannot be fully realized when the offloading link is in deep fade. Fortunately, the impairments induced by wireless propagation environments could be renovated by a reconfigurable intelligent surface (RIS). In this paper, a new RIS-aided edge caching system is proposed, where a network cost minimization problem is formulated to optimize content placement at cache units, active beamforming at base station and passive phase shifting at RIS. After decoupling the content placement subproblem with the hybrid beamforming design, we propose an alternating optimization algorithm to tackle the active beamforming and passive phase shifting. For active beamforming, we transfer the problem into a semidefinite programming (SDP) and prove that the optimal solution of SDP is always rank-one. For passive phase shifting, we introduce the block coordinate descent method to alternately optimize the auxiliary variables and the RIS phase shifts. Further, a conjugate gradient algorithm based on manifold optimization is proposed to deal with the non-convex unit-modulus constraints. Numerical results show that our RIS-aided edge caching design can effectively decrease the network cost by improving the quality of offloading links.**

***Index Terms*— Reconfigurable intelligent surface (RIS), edge caching, network cost, beamforming, manifold optimization.**

Manuscript received November 2, 2020; revised March 16, 2021; accepted June 5, 2021. Date of publication June 16, 2021; date of current version December 10, 2021. This work was supported in part by the National Nature Science Foundation of China under Grant 61871190, Grant 61871416, Grant 62001191, and Grant U2066201; in part by the Natural Science Foundation of Guangdong Province under Grant 2018B030306005; in part by the Fundamental Research Funds for the Central Universities under Grant 21620351; in part by the National Key Research and Development Program of China under Grant 2020YFC1511801; and in part by the Beijing Municipal Natural Science Foundation under Grant L192030. The work of Ertugrul Basar was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 120E401. This article was presented in part at the IEEE Wireless Communications and Networking Conference (WCNC) 2021 [1]. The associate editor coordinating the review of this article and approving it for publication was C.-K. Wen. *(Corresponding author: Miaowen Wen.)*

Yingyang Chen and Weiping Liu are with the Department of Electronic Engineering, College of Information Science and Technology, Jinan University, Guangzhou 510632, China (e-mail: chenyy@jnu.edu.cn; wpl@jnu.edu.cn).

Miaowen Wen is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: eemwwen@scut.edu.cn).

Ertugrul Basar is with the Communications Research and Innovation Laboratory (CoreLab), Department of Electrical and Electronics Engineering, Koç University, 34450 Istanbul, Turkey (e-mail: ebasar@ku.edu.tr).

Yik-Chung Wu is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: ycwu@eee.hku.hk).

Li Wang is with the School of Electronic Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: liwang@bupt.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3087912.

Digital Object Identifier 10.1109/TWC.2021.3087912

I. Introduction

A. Motivation and Scope

NOWADAYS, the driving forces of the exponential growth in mobile network traffic have been fundamentally shifted from the steady increase in demand for conventional *connection-centric* communications to the explosion of *content-centric* ones [2]. Considering the characteristics of cache-able content as well as skewed content popularity, there is consensus today that caching can increase network performance, reduce expenditures for operators, and improve quality-of-service for users [3], [4].

Exploiting the concept of caching to support content delivery over wireless networks is referred to as **edge caching**, i.e., caching at a base station (BS) or mobile devices [5]. Caching also facilitates edge computing capabilities by pre-storing necessary computing datasets at the wireless edge nodes [6]. However, edge caching is fundamentally different from caching in a content delivery network (CDN), since it disperses content files at the wireless edge. Predominantly, the content transmission link over the wireless medium is far from perfect, which leads to uncertain caching performance. For example, the mobile devices located at the cell edge typically suffer from a low delivery rate, and their received content is prone to a low successful probability. Besides, the unpredictable user mobility deteriorating the propagation environments may heavily affect the caching strategies and complicate the content delivery process. Therefore, the edge caching designs naturally induce a coupling between wireless communications and caching strategies, and it is necessary to enhance the performance from a perspective of communications.

With the theoretical and experimental breakthrough in micro-electro-mechanical systems and meta-materials,

communications with reconfigurable intelligent surfaces (RISs), also called intelligent reflecting surfaces (IRSs), have recently been proposed as a powerful solution to enhancing the spectral efficiency (SE) and energy efficiency (EE) of wireless networks [10]–[12]. In particular, an RIS comprises an array of low-cost reflecting elements to proactively configure the end-to-end wireless propagation channel. Compared to legacy relaying systems, the RIS shapes the impinging signal by controlling the phase shift of each reflecting element instead of employing a power amplifier, and is capable of supporting full-duplex and full-band transmissions inherently. Furthermore, as electromagnetic materials, RISs are easy to be coated on existing structures such as building facades, vehicle windows, and indoor walls, which largely reduces the complexity of deployment. By shaping wireless propagation environments proactively, an RIS could be exploited to improve the quality of offloading links, especially on the wireless edge or in mobility. Hence, RIS-empowered communication could be a key enabler for improving caching performance.

When integrating RIS-aided communications into the caching system, it is imperative to investigate the overall system from a content-centric perspective. First, a number of new appropriate performance metrics should be characterized as a function of multi-dimensional resources, including cache storage, power consumption, and spectrum resource. Besides, joint optimization of cache placement and physical layer transmissions is required, and involves mixed time-scale optimization. To exploit the content popularity, the signals from multiple BSs dispersing identical files can be exploited through clustering and cooperative transmission. Hence, RISs supporting content-centric phase shifting are envisioned to improve the delivery quality and efficiency.

B. Related Works

1) Edge Caching in Wireless Networks: The design of wireless transmission techniques changes significantly in the presence of edge caching. Xu *et al.* [13] showed that caching at both transmitters and receivers can turn an interference channel to a cooperative X-multicast channel. In [9], the authors used zero-forcing (ZF) and interference alignment (IA) techniques in the delivery phase to exploit the presence of identical file portions at multiple transmitters and hence reap multiple benefits from caching. In [8], the authors introduced a content-centric multicast beamforming design for content delivery in a cache-enabled radio access network, and [4] further proposed a low-complexity algorithm tailored for large-scale systems. Poularakis *et al.* [14] optimized caching policies based on multicast transmissions to reduce energy costs. The caching performance analysis and optimization in stochastic wireless networks have also attracted great attention [7], [15]–[17], where the node locations are generally modeled as independent spatial random processes, e.g., Poisson Point Process, and advanced stochastic geometry tools can be resorted to [18]. Wen *et al.* [7] introduced cooperative multiple point (CoMP) into small BS (SBS) caching in a downlink large-scale heterogeneous network (HetNet). In [15], the authors exploited full-duplex relaying to boost wireless caching in a two-tier HetNet, where the success probability was derived in closed-form. Liu *et al.* [16] investigated the optimal caching policy to maximize the success probability and area SE in a cache-enabled HetNet. In [17], the authors investigated probabilistic content placement to control cache-based channel selection diversity and network interference. There are some works devoted to exploring the interplay between edge caching and other emerging wireless technologies and services, e.g., full duplex [19]–[21], non-orthogonal transmission [22], [23], unmanned aerial vehicles (UAVs) [24], and vehicular communications [25]. Against this background, there is a paucity of literature for enhancing the caching performance by invoking RISs.

2) RIS Empowered Wireless Networks: To exploit the gains provided by RISs, there is a number of works applying signal enhancement with the reflection path [26]–[32]. Hybrid beamforming consisting of active transmit/receive beamforming at transceiver and passive phase shifting at RIS was often developed to fully reap the multiple-antenna gain, where one of the difficulties lies in the non-convex unit-modulus constraints induced by the phase shifts of RISs [10]. In [26], Wu *et al.* minimized the transmit power for an RIS-aided multiple-input single-output (MISO) system by jointly optimizing the transmit beamforming and the reflect pattern, where the passive beamforming was designed by invoking the semidefinite relaxation (SDR) approach. Ning *et al.* [27] maximized the SE of an RIS-assisted multiple-input multiple-output (MIMO) system. The passive beamforming was solved by alternating direction method of multipliers (ADMM) after taking the amplitude of each reflection coefficient into consideration. Huang *et al.* [28] maximized the sum rate in an RIS-aided MISO downlink communication, where the non-convexity in the RIS matrix was tackled with the aid of a majorization-minimization (MM) method. Further, an energy-efficient design of an RIS for downlink multi-user communications was studied in [29], where the classic gradient descent was employed for obtaining the RIS phase coefficients. Di *et al.* [30] investigated hybrid beamforming in a downlink RIS multi-user system where the discrete feature of phase shifts were considered. Yu *et al.* [31] proposed to reformulate the RIS phase shift deign problem into an equivalent rank-constrained problem, which was further solved by a difference of convex (DC) method. More recently, low complexity RIS phase adjustment algorithms were also explored in [32].

Besides, there is another group of researches align the reflected signals of RISs for signal cancellation at certain terminals, which are often applied in physical layer security (PLS) and interference cancellation [33]–[35]. For example, Lyu *et al.* [33] investigated an RIS jamming scenario, where RISs act as jammers for attacking a legitimate communication without using internal energy. In [34], the authors invoked RISs at the cell boundary to assist the downlink transmission to cell-edge users whilst mitigating the inter-cell interference. Hou *et al.* [35] designed a passive beamforming weight in RIS-aided MIMO non-orthogonal multiple access (NOMA) networks, where the inter-cluster interference can be eliminated.

There are also some contributions devoted to integrating RISs with diverse wireless networks and other emerging technologies, such NOMA networks [35]–[37], mobile edge computing (MEC) [38], simultaneous wireless information and power transfer (SWIPT) [39], [40], UAV wireless networks [41], vehicular communications [42], and machine learning [43]. All above impressive contributions motivate us to exploit the benefits of RISs in edge caching systems.

C. Contributions and Organizations

Our main contributions in this paper are detailed as follows.

- Firstly, we develop a new RIS-aided edge caching design for the first time in literature, and formulate a network cost minimization problem. Specifically, we propose an RIS-aided edge caching system to assist content offloading for mobile users. The network cost minimization problem is formulated to optimize hybrid beamforming and content placement, where the network cost is characterized by both the backhaul capacity and the transmission power. The hybrid beamforming consists of active beamforming at BS and passive phase shifting at RIS. Owing to the non-convex unit-modulus constraints and the coupling of multiple optimization variables, the network-cost-minimization problem cannot be solved in a straightforward manner. By analyzing the problem structure, we decouple the content placement subproblem with the hybrid beamforming design.
- Secondly, we propose an alternating optimization algorithm to decouple active beamforming at BS with passive beamforming at RIS. When fixing the passive beamformer to optimize the active one, we transfer the problem into a semidefinite programming (SDP) by applying SDR. We prove that the optimal solution of the SDP is the solution of the primal active beamforming problem exactly. When fixing the active beamformer to optimize the passive one, we are confronted with a feasibility problem with non-convex unit-modulus constraints. By introducing auxiliary variables and the a penalty function method, we transfer the signal-to-interference-noise-ratio (SINR) constraints into the objective function of the feasibility problem. A block coordinate descent method is exploited to alternately optimize the auxiliary variables and the RIS phase shifts.
- Thirdly, we introduce an effective conjugate gradient algorithm based on manifold optimization to deal with the non-convex unit-modulus constraints. We show that the unit-modulus constraints of all phase shifters constitute a complex circle manifold, i.e., a Riemannian manifold. Then, we optimize the passive phase shifts in a Riemannian manifold, where the original unit-modulus constraints can be easily guaranteed. The optimized results can be obtained through the predefined retraction and mapping operations between the Riemmannian manifold and the Euclidean space. Goldstein criterion line search and Flecther-Reeves equation are employed to guarantee convergence to a locally optimal solution.
- Finally, we present numerical validations and evaluations. Numerical results show that our RIS-aided edge caching design effectively decreases the network cost in terms of backhaul capacity and power consumption. Specifically, the power cost can be decreased by installing more passive reflecting elements on RIS. Meanwhile, our proposed content placement design achieves lower backhaul cost than other existing caching strategies. Furthermore, simulation results show that the RIS location should be carefully chosen, and the path loss exponent of RIS-related links has prominent impact on the achievable performance. It is shown that the RIS is better to be deployed in an open area with a relatively low path loss exponent.

The rest of this paper is organized as follows. In Section II, we establish the system model and formulate the network cost minimization problem. An alternating optimization method is developed in Section III. In Section IV, we propose a block coordinate descent method for passive beamforming. Our numerical results are discussed in Section V. Finally, the conclusion of this article is presented in Section VI.

Notation: Uppercase and lowercase bold-faced letters indicate matrices and vectors, respectively. Calligraphic letters denote sets. Let $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^H$ refer to the transpose, conjugate and conjugate-transpose operations, respectively. The l_2-norm of a vector $\mathbf{a}$ is shown by $\|\mathbf{a}\|_2$, and $(\mathbf{a})_i$ represents the i-th element of $\mathbf{a}$. Notation $\mathbf{A}\succeq\mathbf{0}$ represents that $\mathbf{A}$ is a positive semidefinite matrix. The ring of complex numbers is denoted as $\mathbb{C}$, whilst $\mathbf{A}\in\mathbb{C}^{M\times N}$ indicates that $\mathbf{A}$ is a complex-element matrix with dimensions $M\times N$. An $N\times N$ dimensional identity matrix is denoted as $\mathbf{I}_N$, and $\mathrm{diag}\,(\cdot)$ denotes the diagonalization operation. The range space and null space of matrix $\mathbf{A}$ are denoted as $\mathcal{R}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A})$, respectively. The real part of a complex argument is noted by $\mathrm{Re}\,\{\cdot\}$, and $\circ$ denote the Hadamard (element-wise) product between two matrices. Finally, $x\sim\mathcal{CN}\left(\mu,\sigma^2\right)$ indicates that the random variable x obeys a complex Gaussian distribution with mean μ and variance σ^2.

II. System Model and Problem Formulation

A. System Model

As illustrated in Fig. 1, we consider the downlink transmission of a cache-enabled radio access network with an RIS. There exist an M multiple-antenna BS, K single-antenna mobile users, and an RIS equipped with N passive reflecting elements. In practice, the RIS is configured by the BS through a control link. The BS has local cache with a limited storage size. Meanwhile, the BS is also connected to the server via a high-capacity backhaul link and can access a database that contains a total number of F files with equal size. Let $\mathcal{K}=\{1,\cdots,K\}$ denote the set of mobile users, and $\mathcal{N}=\{1,\cdots,N\}$ represent the set of RIS elements. The channels between BS and RIS, the user-k and BS, the user-k and RIS are respectively noted as $\mathbf{G}\in\mathbb{C}^{N\times M}$, $\mathbf{h}_{d,k}\in\mathbb{C}^{M\times 1}$, and $\mathbf{h}_{r,k}\in\mathbb{C}^{N\times 1}$. We consider a time-division duplexing (TDD) mode for uplink and downlink, and assume that the channel state information (CSI) are obtained at BS through channel reciprocity and some sophisticated channel estimation algorithms [44], [45]. Further, we assume that all

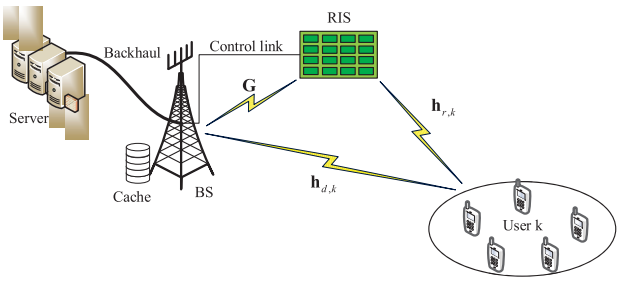


Fig. 1. RIS-aided edge caching system.

CSI acquired at BS are perfect, since we aim to characterize the maximum performance gain brought by the optimization design. Developing robust optimization designs under imperfect or statistic CSI acquisition constitutes a promising research direction for practical RIS implementation, and will be considered as our future work.

B. Cache Model

Let $\mathcal{F}=\{1,\cdots,F\}$ denote the set of F files in the content database, where each file is assumed to have normalized size of 1. We assume that all files in the content database $\mathcal{F}$ are sequenced according to their popularities, with the most popular one as the 1-st file and the least one as the F-th file. Particularly, the popularity of each content file obeys the Zipf distribution with skewness factor ε. The local storage size of the BS is denoted as S_0 ($S_0 < F$), which represents the maximum number of files it can cache. We assume that the BS applies a probabilistic caching strategy, that is, the BS caches a content randomly with deterministic probability. More specifically, we define a content placement vector $\mathbf{c} = \{c_1,\cdots,c_f,\cdots,c_F\}$, where $c_f \in [0,1]$ indicates the probability that the f-th content file is cached in the BS. Due to the limited cache size, $\sum_{f\in\mathcal{F}} c_f \le S_0$ should be satisfied. For the users, b_f^k denotes the probability that user k requests file f. For simplicity,[1] we assume that the request distributions for different users are uniform, i.e., $b_f^1 = b_f^2 = \cdots = b_f^K \triangleq b_f$. Further, we assume that users request content files according to their popularity, and the request probability follows a Zipf distribution with skewness factor ε, i.e., $b_f = \frac{f^{-\varepsilon}}{\sum_{i=1}^{F} i^{-\varepsilon}}$. In general, a large value of ε means more user requests are concentrated on fewer popular files.

During the period of content fetching, if content f has been cached in BS's local storage, users can access the content directly. Otherwise, it needs to fetch this content from BS via backhaul. Since the data rate of fetching a content needs to be as large as the content-delivery target rate R_k^0, we model the total backhaul cost as $\sum_{f=1}^{F}\sum_{k=1}^{K}(1-c_f)\,b_f R_k^0$.

C. Communication Model

Let $\mathbf{p}_k \in \mathbb{C}^{M\times 1}$ denote the precoding vector at the BS for user k, and $\mathbf{P} = [\mathbf{p}_1,\cdots,\mathbf{p}_k,\cdots,\mathbf{p}_K] \in \mathbb{C}^{M\times K}$ denote the active precoding matrix at the BS. Further, let $\mathbf{\Theta} = \mathrm{diag}\left\{e^{j\theta_1},\cdots,e^{j\theta_n},\cdots,e^{j\theta_N}\right\}$ denote the reflection coefficient matrix at the RIS, where $\theta_n \in [0,2\pi)$ is the phase shift of the n-th reflecting element. The signal received at user k from both of the BS-user and BS-RIS-user channels can be expressed as

$$y_k = \left(\mathbf{h}_{d,k}^H + \mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}\right)\mathbf{P}\mathbf{s} + z_k \tag{1}$$

$$= \sum_{l=1}^{K}\left(\mathbf{h}_{d,k}^H + \mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}\right)\mathbf{p}_l s_l + z_k \tag{2}$$

where $z_k \sim \mathcal{CN}\left(0,\sigma_0^2\right)$ is the i.i.d Gaussian noise at the receiver of user k, $\mathbf{s} = [s_1,\cdots,s_l,\cdots,s_K]^T \in \mathbb{C}^K$ is the transmit signal vector, s_l is the signal for user l, and σ_0^2 is the noise power spectral density. Therefore, the SINR of user k is given by

$$\gamma_k = \frac{\left|\left(\mathbf{h}_{d,k}^H + \mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}\right)\mathbf{p}_k\right|^2}{\sum\limits_{l=1,l\neq k}^{K}\left|\left(\mathbf{h}_{d,k}^H + \mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}\right)\mathbf{p}_l\right|^2 + \sigma_0^2 B} \tag{3}$$

where B is the system bandwidth.

D. Problem Formulation

In this paper, we aim to optimize the active beamformer $\mathbf{P}$, passive beamformer $\mathbf{\Theta}$ and content placement vector $\mathbf{c}$ to minimize the total network cost, which consists of both the backhaul capacity and the transmission power. To this end, we formulate the optimization problem as

$$\mathcal{P}_0: \min_{\{c_f\},\{\mathbf{p}_k\},\{\theta_n\}} \eta\sum_{f=1}^{F}\sum_{k=1}^{K}(1-c_f)\,b_f R_k^0 + \sum_{k=1}^{K}\|\mathbf{p}_k\|_2^2 \tag{4a}$$

$$s.t.\quad \gamma_k \ge \gamma_k^0, \quad \forall k\in\mathcal{K}, \tag{4b}$$

$$c_f \in [0,1], \quad \forall f\in\mathcal{F} \tag{4c}$$

$$\sum_{f=1}^{F} c_f \le S_0 \tag{4d}$$

$$0 \le \theta_n < 2\pi, \quad \forall n\in\mathcal{N} \tag{4e}$$

where $\eta > 0$ is the pricing factor to trade the backhaul capacity for the power consumption, and $\gamma_k^0 = 2^{R_k^0/B} - 1$ is the SINR requirement with respect to the content-delivery target rate of user k.

Constraint (4b) is a non-convex minimum SINR requirement constraint, while constraint (4d) considers the local storage limit at the BS, and constraint (4e) defines the interval of phase shifts, which is a non-convex unit-modulus constraint. Hence the formulated problem is non-convex and cannot be solved in a straightforward manner.

III. Alternating Optimization Development

In this section, we propose an alternating optimization method to deal with the formulated problem $\mathcal{P}_0$. The key idea is to decouple the content placement optimization with the beamforming design, and then decouple the active beamforming with the passive phase shifts as so to alternately update each other till convergence.

[1] In fact, our model is applicable for a general case, where diverse request distributions can be assumed for different users. Here b_f^k acts essentially as a weighting factor, and has no impact on the subsequent problem formulation and solution.

A. Content Placement Optimization

For the content placement part in the alternating optimization, we assume both of the passive and active beamformers are fixed. Hence $\mathcal{P}_0$ is transformed into

$$\begin{aligned}\mathcal{P}_1:\ &\min_{\{c_f\}} \sum_{f=1}^{F}\left(1-c_f\right)b_f\\ &s.t.\ c_f\in[0,1],\quad \forall f\in\mathcal{F}\\ &\sum_{f=1}^{F} c_f\le S_0\end{aligned}\tag{5}$$

Above program is convex with respect to the content placement vector $\mathbf{c}=\{c_1,\cdots,c_f,\cdots,c_F\}$, and can be solved by KKT optimality conditions [46].

B. Beamforming Optimization

On the other hand, when fixing the content placement, $\mathcal{P}_0$ is transformed into

$$\begin{aligned}\mathcal{P}_2:\ &\min_{\{\mathbf{p}_k\},\{\theta_n\}} \sum_{k=1}^{K}\|\mathbf{p}_k\|_2^2\\ &s.t.\ \gamma_k\ge\gamma_k^0,\quad \forall k\in\mathcal{K}\\ &0\le\theta_n\le 2\pi,\quad \forall n\in\mathcal{N}.\end{aligned}\tag{6}$$

In this sequel, we also turn to an alternating methodology to deal with $\mathcal{P}_2$.

1) Active Beamforming: When fixing the passive beamformer $\mathbf{\Theta}$ to optimize the active beamformer $\mathbf{P}$, $\mathcal{P}_2$ becomes

$$\begin{aligned}\mathcal{P}_{2-\mathrm{I}}:\ &\min_{\{\mathbf{p}_k\}} \sum_{k=1}^{K}\|\mathbf{p}_k\|_2^2\\ &s.t.\ \gamma_k\ge\gamma_k^0,\quad \forall k\in\mathcal{K}.\end{aligned}\tag{7}$$

This problem in the present form is nonconvex since the inequality constraint function apparently is nonconvex. Next, let us reformulate it into an SDP.

By denoting $\mathbf{f}_k^H=\mathbf{h}_{d,k}^H+\mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}$, the inequality constraints in (7) can be re-expressed as

$$\frac{1}{\gamma_k^0}\left|\mathbf{f}_k^H\mathbf{p}_k\right|^2\ge\sum_{l=1,l\ne k}^{K}\left|\mathbf{f}_k^H\mathbf{p}_l\right|^2+\sigma_0^2,\quad k=1,\cdots,K.\tag{8}$$

To cope with $\mathcal{P}_{2-\mathrm{I}}$, we reformulate the complex-variable problem into an SDP by applying SDR. Specifically, we replace the rank-one matrix $\mathbf{p}_k\mathbf{p}_k^H$ by a general-rank positive semidefinite (PSD) matrix $\mathbf{P}_k$ of $M\times M$ dimensions. Then the following SDP problem is yielded

$$\begin{aligned}\mathcal{P}_{2-\mathrm{I}}^{SDR}:\ &\min_{\{\mathbf{P}_k\}} \sum_{k=1}^{K}\mathrm{Tr}\left(\mathbf{P}_k\right)\\ &s.t.\ \frac{1}{\gamma_k^0}\mathbf{f}_k^H\mathbf{P}_k\mathbf{f}_k\ge\sum_{l=1,l\ne k}^{K}\mathbf{f}_k^H\mathbf{P}_l\mathbf{f}_k+\sigma_0^2,\quad \forall k\in\mathcal{K}\\ &\mathbf{P}_k\underline{\succeq}\mathbf{0},\quad \forall k\in\mathcal{K}.\end{aligned}\tag{9}$$

Lemma 1: The obtained optimal $\left\{\mathbf{P}_k^{(*)}\right\}$ of problem (9) is of rank one, i.e., $\mathbf{P}_k^{(*)}=\mathbf{p}_k^{(*)}\left(\mathbf{p}_k^{(*)}\right)^H$; thus $\left\{\mathbf{p}_k^{(*)}\right\}$ must be an optimal solution of problem (7).

Proof: The Lagrangian of problem (9) is shown as in (10) on the bottom of the next page, where $\lambda_k\in\mathbb{R}$ and $\mathbf{Z}_k\in\mathbb{H}^M$ are the dual variables associated with the constraints in (9). Let

$$\mathbf{A}_k=\mathbf{I}-\frac{\lambda_k}{\gamma_k^0}\mathbf{f}_k\mathbf{f}_k^H+\sum_{\substack{l=1\\ l\ne k}}^{K}\lambda_l\mathbf{f}_l\mathbf{f}_l^H-\mathbf{Z}_k,\tag{11}$$

the dual function of problem (9) can be easily seen to be

$$\begin{aligned}&g\left(\{\lambda_k,\mathbf{Z}_k\}\right)\\ &=\inf_{\mathbf{P}_k\in\mathbb{C}^M}\mathcal{L}\left(\{\mathbf{P}_k,\lambda_k,\mathbf{Z}_k\}\right)\\ &=\inf_{\mathbf{P}_k\in\mathbb{C}^M}\sum_{k=1}^{K}\frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{A}_k\mathbf{P}_k\right)+\mathrm{Tr}\left(\mathbf{A}_k^*\mathbf{P}_k^*\right)\right]+\sum_{k=1}^{K}\lambda_k\sigma_0^2\\ &=\begin{cases}\sum_{k=1}^{K}\lambda_k\sigma_0^2, & \mathbf{A}_k=\mathbf{0},\quad\forall k\\ -\infty, & \text{otherwise.}\end{cases}\end{aligned}\tag{12}$$

Thus, the dual of problem (9) is given by

$$\begin{aligned}&\max_{\{\lambda_k,\mathbf{Z}_k\}}\sum_{k=1}^{K}\lambda_k\sigma_0^2\\ &s.t.\ \mathbf{A}_k=\mathbf{0},\ \mathbf{Z}_k\underline{\succeq}\mathbf{0},\ \lambda_k\ge 0,\quad\forall k.\end{aligned}\tag{13}$$

Since problem (9) is convex and Slater's condition holds true, KKT conditions are the sufficient and necessary optimality conditions. Some KKT conditions needed in the proof are as follows

$$\frac{1}{\gamma_k^0}\mathbf{f}_k^H\mathbf{P}_k^{(*)}\mathbf{f}_k\ge\sum_{l=1,l\ne k}^{K}\mathbf{f}_k^H\mathbf{P}_l^{(*)}\mathbf{f}_k+\sigma_0^2,\quad k=1,\cdots,K\tag{14}$$

$$\mathbf{Z}_k^{(*)}=\mathbf{I}-\frac{\lambda_k^{(*)}}{\gamma_k^0}\mathbf{f}_k\mathbf{f}_k^H+\sum_{l=1,l\ne k}^{K}\lambda_l^{(*)}\mathbf{f}_l\mathbf{f}_l^H,\quad k=1,\cdots,K\tag{15}$$

$$\mathbf{Z}_k^{(*)}\mathbf{P}_k^{(*)}=\mathbf{0},\quad k=1,\cdots,K.\tag{16}$$

Note that $\lambda_k\ge 0$, $\mathbf{Z}_k^{(*)}\underline{\succeq}\mathbf{0}$ and $\mathbf{P}_k^{(*)}\underline{\succeq}\mathbf{0}$ are also KKT conditions.

Since $\mathbf{P}_k^{(*)}\ne\mathbf{0}$ (by (14)), the rank of $\mathbf{Z}_k^{(*)}$ must be less than or equal to $M-1$ to ensure that at least one dimension of null space exists (by (16)), i.e.,

$$\mathrm{rank}\left(\mathbf{Z}_k^{(*)}\right)\le M-1.\tag{17}$$

By defining

$$\mathbf{B}=\mathbf{I}+\sum_{\substack{l=1\\ l\ne k}}^{K}\lambda_l\mathbf{f}_l\mathbf{f}_l^H=\left(\mathbf{B}^{1/2}\right)^2\succ\mathbf{0}\tag{18}$$

where $\mathbf{B}^{1/2}=\left(\mathbf{B}^{1/2}\right)^H \succ \mathbf{0}$, the rank of $\mathbf{Z}_k^{(*)}$ can be further inferred as follows

$$\begin{aligned}
&\operatorname{rank}\left(\mathbf{Z}_k^{(*)}\right)\\
&=\operatorname{rank}\left(\mathbf{B}-\frac{\lambda_k^{(*)}}{\gamma_k^0}\mathbf{f}_k\mathbf{f}_k^H\right)\\
&=\operatorname{rank}\left(\mathbf{B}^{1/2}\left[\mathbf{I}-\frac{\lambda_k^{(*)}}{\gamma_k^0}\mathbf{B}^{-1/2}\mathbf{f}_k\mathbf{f}_k^H\mathbf{B}^{-1/2}\right]\mathbf{B}^{1/2}\right)\\
&=\operatorname{rank}\left(\mathbf{I}-\frac{\lambda_k^{(*)}}{\gamma_k^0}\mathbf{B}^{-1/2}\mathbf{f}_k\mathbf{f}_k^H\mathbf{B}^{-1/2}\right)\\
&\geq M-1. \qquad (19)
\end{aligned}$$

From (17) and (19), it can be inferred that $\operatorname{rank}\left(\mathbf{Z}_k^{(*)}\right)=M-1$. Then, by (16), we have

$$\begin{aligned}
\operatorname{rank}\left(\mathbf{P}_k^{(*)}\right) &\leq \dim\left(\mathcal{N}\left(\mathbf{Z}_k^{(*)}\right)\right)=M-\operatorname{rank}\left(\mathbf{Z}_k^{(*)}\right)=1\\
&\Rightarrow \operatorname{rank}\left(\mathbf{P}_k^{(*)}\right)=1 \quad (\text{since } \mathbf{P}_k^{(*)}\neq\mathbf{0}).
\end{aligned}$$

Therefore, the optimal solution $\left\{\mathbf{P}_k^{(*)}\right\}$ of the SDR problem (9) must yield the optimal solution $\left\{\mathbf{p}_k^{(*)}\right\}$ of the active beamforming problem (7) via the rank-one decomposition $\mathbf{P}_k^{(*)}=\mathbf{p}_k^{(*)}\left(\mathbf{p}_k^{(*)}\right)^H$. ■

Till now, by solving (9) efficiently via off-the-self convex solvers and then applying Lemma 1, we can obtain optimized active beamformers $\left\{\mathbf{p}_k^{(*)}\right\}$ of program $\mathcal{P}_{2-I}$ in (7).

2) Passive Beamforming: When fixing the active beamformer $\mathbf{P}$ to optimize the passive beamformer $\mathbf{\Theta}$, $\mathcal{P}_2$ reduces to

$$\begin{aligned}
\mathcal{P}_{2-\mathrm{II}}:\ &\min_{\{\theta_n\}}\sum_{k=1}^{K}\|\mathbf{p}_k\|_2^2\\
&s.t.\ \gamma_k\geq\gamma_k^0,\quad \forall k\in\mathcal{K}\\
&\quad 0\leq\theta_n\leq 2\pi,\quad \forall n\in\mathcal{N}. \qquad (20)
\end{aligned}$$

Apparently, $\mathcal{P}_{2-\mathrm{II}}$ is a feasibility problem, where the objective function is independent of the variables [46]. It is clear that the difficulty of the problem solving mainly lies in the non-convex unit modulus constraints induced by the phase shifts. In the following section, we resort to a block coordinate descent algorithm to solve the feasibility problem above.

IV. Block Coordinate Descent Method for Passive Beamforming

In the following, we propose a block coordinate descent algorithm to deal with passive beamforming problem above. Based on a penalty function method, we transfer the SINR constraints into the objective function of the feasibility problem by introducing auxiliary variables. Then, the block coordinate descent method is proposed to alternately optimize the auxiliary variables and the RIS phase shifts.

A. Problem Reformulation

First, we introduce auxiliary variables to transfer the SINR constraints in (20) into the objective function. To this end, let

$$\mathbf{f}_k^H\mathbf{p}_l=x_{k,l} \qquad (21)$$

and

$$\mathbf{f}_k^H\mathbf{p}_k=x_{k,k}. \qquad (22)$$

Then the SINR constraints in (20) can be expressed as

$$\begin{aligned}
\mathcal{P}_{2-\mathrm{II}}:\ &\min_{\{\theta_n\},\{x_{k,j}\}}\sum_{k\in\mathcal{K}}\|\mathbf{p}_k\|_2^2 && (23a)\\
&s.t.\ \frac{|x_{k,k}|^2}{\sum_{l\in\mathcal{K},l\neq k}|x_{k,l}|^2+\sigma_0^2}\geq\gamma_k^0,\quad \forall k\in\mathcal{K} && (23b)\\
&\quad \mathbf{f}_k^H\mathbf{p}_j=x_{k,j},\quad \forall k,j\in\mathcal{K} && (23c)\\
&\quad 0\leq\theta_n\leq 2\pi,\quad \forall n\in\mathcal{N} && (23d)
\end{aligned}$$

To associate the objective function with the variables, we integrate equality constraints into the objective function of (23) by introducing a penalty parameter [40], yielding the following optimization problem

$$\begin{aligned}
\mathcal{P}'_{2-\mathrm{II}}:\ &\min_{\{\theta_n\},\{x_{k,l}\}}\sum_{k\in\mathcal{K}}\|\mathbf{p}_k\|_2^2+\frac{1}{2\rho}\left(\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{K}}\left|\mathbf{f}_k^H\mathbf{p}_j-x_{k,j}\right|^2\right) && (24a)\\
&s.t.\ \frac{|x_{k,k}|^2}{\sum_{l\in\mathcal{K},l\neq k}|x_{k,l}|^2+\sigma_0^2}\geq\gamma_k^0,\quad \forall k\in\mathcal{K} && (24b)\\
&\quad 0\leq\theta_n\leq 2\pi,\quad \forall n\in\mathcal{N} && (24c)
\end{aligned}$$

where $\rho>0$ denotes the penalty coefficient used for penalizing the violation of equality constraints.

$$\begin{aligned}
\mathcal{L}\left(\{\mathbf{P}_k,\lambda_k,\mathbf{Z}_k\}\right)&=\sum_{k=1}^{K}\mathrm{Tr}\left(\left[\mathbf{I}-\frac{\lambda_k}{\gamma_k^0}\mathbf{f}_k\mathbf{f}_k^H-\mathbf{Z}_k\right]\mathbf{P}_k\right)+\sum_{k=1}^{K}\sum_{\substack{l=1\\l\neq k}}^{K}\mathrm{Tr}\left(\lambda_k\mathbf{f}_k\mathbf{f}_k^H\mathbf{P}_l\right)+\sum_{k=1}^{K}\lambda_k\sigma_0^2\\
&=\sum_{k=1}^{K}\mathrm{Tr}\left(\left[\mathbf{I}-\frac{\lambda_k}{\gamma_k^0}\mathbf{f}_k\mathbf{f}_k^H-\mathbf{Z}_k\right]\mathbf{P}_k\right)+\sum_{l=1}^{K}\sum_{\substack{k=1\\k\neq l}}^{K}\mathrm{Tr}\left(\lambda_l\mathbf{f}_l\mathbf{f}_l^H\mathbf{P}_k\right)+\sum_{k=1}^{K}\lambda_k\sigma_0^2\\
&=\sum_{k=1}^{K}\mathrm{Tr}\left(\left[\mathbf{I}-\frac{\lambda_k}{\gamma_k^0}\mathbf{f}_k\mathbf{f}_k^H+\sum_{\substack{l=1\\l\neq k}}^{K}\lambda_l\mathbf{f}_l\mathbf{f}_l^H-\mathbf{Z}_k\right]\mathbf{P}_k\right)+\sum_{k=1}^{K}\lambda_k\sigma_0^2 \qquad (10)
\end{aligned}$$

B. Block Coordinate Descent Algorithm

It can be found in (24) that, for given $\rho > 0$, ${\mathcal{P}'}_{2-\mathrm{II}}$ is still a non-convex optimization problem due to the non-convex objective function as well as the non-convex constraints. However, it is observed that each optimization variable in ${\mathcal{P}'}_{2-\mathrm{II}}$ is involved in at most one constraint. This observation motivates us to apply the block coordinate descent method to solve ${\mathcal{P}'}_{2-\mathrm{II}}$ efficiently by properly partitioning the optimization variables into a pair of blocks. Besides, in the following, we may find that the objective of each subproblem has no connection with the penalty coefficient ρ, which simplifies the subsequent problem solving. Hence we do not need to initialize ρ to an efficient value and update it iteratively.

1) Subproblem With Respect to $\{x_{k,j}\}$: On the one hand, for any given phase shifts $\{\theta_n\}$, the auxiliary variables can be optimized by solving the following problem

$$\begin{aligned} &\min_{\{x_{k,j}\}} \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{K}} \left|\bar{x}_{k,j}-x_{k,j}\right|^2 \\ &s.t.\ \frac{\left|x_{k,k}\right|^2}{\sum\limits_{l\in\mathcal{K},l\neq k}\left|x_{k,l}\right|^2+\sigma_0^2} \geq \gamma_k^0, \quad \forall k\in\mathcal{K} \end{aligned} \tag{25}$$

where $\bar{x}_{k,j} := \mathbf{f}_k^H\mathbf{p}_j$. In (25), the optimization variables with respect to different receiving users are separable in both objective function and constraints, we can solve the resultant problem by solving K independent subproblems in parallel, each with only one single constraint (i.e., $\frac{\left|x_{k,k}\right|^2}{\sum\limits_{l\in\mathcal{K},l\neq k}\left|x_{k,l}\right|^2+\sigma_0^2} \geq \gamma_k^0$ with k fixed). Specifically, for user k, (25) is reduced to

$$\begin{aligned} &\min_{\{x_{k,j}\}} \sum_{j\in\mathcal{K}} \left|\bar{x}_{k,j}-x_{k,j}\right|^2 \\ &s.t.\ \frac{\left|x_{k,k}\right|^2}{\sum\limits_{l\in\mathcal{K},l\neq k}\left|x_{k,l}\right|^2+\sigma_0^2} \geq \gamma_k^0. \end{aligned} \tag{26}$$

The problem above is a non-convex quadratically constrained quadratic program (QCQP). It has been shown in the literature that strong duality holds for this problem [46]. Accordingly, the Lagrangian function can be expressed as

$$\begin{aligned} \mathcal{L}\left(\{x_{k,j}\},\lambda_k\right) &= \left(1-\lambda_k\right)\left|x_{k,k}\right|^2 \\ &+ \sum_{l\in\mathcal{K},l\neq k}\left(1+\lambda_k\gamma_k^0\right)\left|x_{k,l}\right|^2 - 2\sum_{j\in\mathcal{K}}\mathrm{Re}\left(\bar{x}_{k,j}x_{k,j}^*\right). \end{aligned} \tag{27}$$

Hence, the dual function is given by

$$g\left(\lambda_k\right) = \min_{\{x_{k,j}\}} \mathcal{L}\left(\{x_{k,j}\},\lambda_k\right). \tag{28}$$

It is easy to find that $0 < \lambda_k < 1$; otherwise we have $g\left(\lambda_k\right) \to -\infty$ when $\left|x_{k,k}\right| \to \infty$. By exploiting the first-order optimality condition, the optimal solution of the dual function is given by

$$x_{k,k}^{(*)} = \frac{\bar{x}_{k,k}}{1-\lambda_k} \tag{29}$$

$$x_{k,l}^{(*)} = \frac{\bar{x}_{k,l}}{1+\lambda_k\gamma_k^0}, l\in\mathcal{K}, l\neq k. \tag{30}$$

Substituting above optimal solutions into the equality constraint in (26), we obtain

$$f\left(\lambda_k\right) \triangleq \frac{\left|\bar{x}_{k,k}\right|^2}{\left(1-\lambda_k\right)^2} - \sum_{l\in\mathcal{K},l\neq k}\frac{\gamma_k^0\left|\bar{x}_{k,l}\right|^2}{\left(1+\lambda_k\gamma_k^0\right)^2} - \gamma_k^0\sigma_0^2 = 0. \tag{31}$$

It can be observed that $f\left(\lambda_k\right)$ is a monotonically increasing function of λ_k for $0 < \lambda_k < 1$. Hence, the optimal variables can be obtained by resorting to a bisection search.

2) Subproblem With Respect to $\{\theta_n\}$: On the other hand, for any given auxiliary variables $\{x_{k,l}\}$, the phase shifts $\{\theta_n\}$ can be optimized by solving the following problem

$$\begin{aligned} {\mathcal{P}''}_{2-\mathrm{II}}:\ &\min_{\{\theta_n\}} \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{K}}\left|\mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}\mathbf{p}_j - x_{k,j}\right|^2 \\ &s.t.\ 0\leq\theta_n<2\pi, \quad \forall n\in\mathcal{N}. \end{aligned} \tag{32}$$

By letting $\varphi_n = e^{j\theta_n}$, $\mathbf{x} = \left[\varphi_1,\cdots,\varphi_n,\cdots,\varphi_N\right]^T$, and $\mathbf{g}_j = \mathbf{G}\mathbf{p}_j$, above program is equivalent to

$$\begin{aligned} {\mathcal{P}''}_{2-\mathrm{II}}:\ &\min_{\mathbf{x}} \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{K}}\left|\mathbf{h}_{r,k}^H\mathrm{diag}\left(\mathbf{x}\right)\mathbf{g}_j - x_{k,j}\right|^2 \\ &s.t.\ \left|\varphi_n\right| = 1, \quad \forall n\in\mathcal{N}. \end{aligned} \tag{33}$$

In (33), the unit modulus constraints are still included, which are intrinsically non-convex, and there is no general approach to solve the optimization problem with these constraints optimally to the best of our knowledge.

C. Conjugate Gradient Algorithm Based on Manifold Optimization

In this subsection, we introduce a conjugate gradient algorithm based on manifold optimization to find a sub-optimal solution to problem (33). Manifold optimization provides a powerful alternative to constrained optimization problem [48]. In literature, the methodology of manifold optimization has been proved to be effective in handling non-convex unit-modulus constraints [34], [49].

To solve (33), we start with some related definitions and terminologies in manifold optimization. A manifold $\mathcal{M}$ is a topological space that resembles a Euclidean space near individual point. The tangent space $T_x\mathcal{M}$ at a given point x on the manifold $\mathcal{M}$ is composed of the tangent vectors ξ_x of the curves γ through the point x. In most applications, manifolds fall into a special category of topological manifold, namely, a *Riemannian manifold*. A Riemannian manifold is equipped with an inner product, which is defined on the tangent spaces $T_x\mathcal{M}$, and allows one to measure distances and angles on manifolds. More importantly, optimization over a Riemannian manifold is locally analogous to that over an Euclidean space with smooth constraints.

Specifically, the complex circle manifold of $x\in\mathbb{C}$, which is defined by

$$\mathcal{M}_{cc} = \left\{x\in\mathbb{C}\,|x^*x = 1\right\}. \tag{34}$$

The complex circle manifold $\mathcal{M}_{cc}$ is a Riemannian submanifold of $\mathbb{C}$. Note that the Euclidean metric over the complex plane $\mathbb{C}$ for $\forall x_1, x_2\in\mathbb{C}$ is defined as $\langle x_1, x_2\rangle = \mathrm{Re}\left\{x_1^*x_2\right\}$.

Hence, the tangent space at the point $x \in \mathcal{M}_{cc}$ can be represented by

$$T_x\mathcal{M}_{cc} = \{z \in \mathbb{C} \,|\langle x, z\rangle = 0\}. \tag{35}$$

Now we expand from a one-dimensional manifold to a multiple dimensional one. Let $\varphi_n = e^{j\theta_n}$ and $\mathbf{x} = [\varphi_1, \cdots, \varphi_n, \cdots, \varphi_N]^T$. Observing the unit modulus constraint $|\varphi_n| = 1$, we may find that $\mathbf{x}$ forms an N-dimensional complex circle manifold $\mathcal{M}_{cc}^N = \{\mathbf{x} \in \mathbb{C}^N \,|{x_1}^* x_1 = \cdots = {x_N}^* x_N = 1\}$. Intrinsically, the complex circle manifold $\mathcal{M}_{cc}^N$ is a Riemannian submanifold of the N-dimensional complex space $\mathbb{C}^N$. Therefore, the feasible region of the optimization problem (33) is over the manifold $\mathcal{M}_{cc}^N$. Further, the tangent space at the point $\mathbf{x} \in \mathcal{M}_{cc}^N$ is expressed as

$$T_{\mathbf{x}}\mathcal{M}_{cc}^N = \{\mathbf{z} \in \mathbb{C}^N \,|\mathrm{Re}\{\mathbf{z} \circ \mathbf{x}^*\} = \mathbf{0}_N\}. \tag{36}$$

In the following we introduce some basic operations in manifold optimization.

Riemannian gradient: Among all the tangent vectors in $T_{\mathbf{x}}\mathcal{M}_{cc}^N$, similar to the Euclidean space, one of them is related to the negative Riemannian gradient, representing the direction of the greatest decrease of a function at the point $\mathbf{x} \in \mathcal{M}_{cc}^N$. It can be computed as the projection from the Euclidean gradient $\nabla f(\mathbf{x})$ to the tangent space using the orthogonal projector. Explicitly, the Riemannian gradient at point $\mathbf{x}$ is a tangent vector $\mathrm{grad} f(\mathbf{x})$, which is given by the orthogonal projection of the Euclidean gradient $\nabla f(\mathbf{x})$ onto the tangent space $T_{\mathbf{x}}\mathcal{M}_{cc}^N$ at point $\mathbf{x} \in \mathcal{M}_{cc}^N$, and can be written as

$$\mathrm{grad} f(\mathbf{x}) = \nabla f(\mathbf{x}) - \mathrm{Re}\{\nabla f(\mathbf{x}) \circ \mathbf{x}^*\} \circ \mathbf{x}. \tag{37}$$

Denoting the cost function in (33) as $f(\mathbf{x})$, which can be written equivalently as

$$f(\mathbf{x}) = \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{K}} \left| \sum_{n\in\mathcal{N}} \left(\mathbf{h}_{r,k}^H\right)_n \varphi_n (\mathbf{g}_j)_n - x_{k,j} \right|^2, \tag{38}$$

the Euclidean gradient of $f(\mathbf{x})$ is given by

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial \varphi_1}, \cdots, \frac{\partial f(\mathbf{x})}{\partial \varphi_n}, \cdots, \frac{\partial f(\mathbf{x})}{\partial \varphi_N} \right]^T$$

where

$$\frac{\partial f(\mathbf{x})}{\partial \varphi_n} = \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{K}} \left(\mathbf{h}_{r,k}^H \mathrm{diag}(\mathbf{x})\, \mathbf{g}_j - x_{k,j}\right)^* \left(\mathbf{h}_{r,k}^H\right)_n (\mathbf{g}_j)_n. \tag{39}$$

Solving the Euclidean gradient involves some techniques on complex-valued matrix derivatives. Interested readers can refer to the details in [50].

Retraction: Retraction is to map a tangent vector from the tangent space back into the manifold. It determines the next point on the manifold when descending along the negative Riemannian gradient direction. The retraction of a tangent vector $\mathbf{d}$ at point $\mathbf{x}_k$ from $T_{\mathbf{x}_k}\mathcal{M}_{cc}^N$ into $\mathcal{M}_{cc}^N$ is expressed as

$$\mathrm{Rtrctn}_{\mathbf{x}_k}(\mathbf{d}) = \mathrm{vec}\left[\frac{(\mathbf{x}_k + \mathbf{d})_i}{|(\mathbf{x}_k + \mathbf{d})_i|} \right]. \tag{40}$$

Algorithm 1 Conjugate Gradient Algorithm for Passive Beamforming Based on Manifold Optimization

Input: $\mathbf{h}_{r,k}, \mathbf{g}_j, x_{k,j}$
Output: $\mathbf{x}$
1 Randomly initialize $\mathbf{x}_0 \in \mathcal{M}_{cc}^m$; Calculate $\mathbf{d}_0 = -\mathrm{grad} f(\mathbf{x}_0)$; Set $k = 0$.
2 **repeat**
3 Choose line search step size α_k according to **Goldstein criterion**.
4 Find the next point $\mathbf{x}_{k+1}$ using retraction by (40): $\mathbf{x}_{k+1} = \mathrm{Rtrctn}_{\mathbf{x}_k}(\alpha_k \mathbf{d}_k)$.
5 Determine the Riemannian gradient at $\mathbf{x}_{k+1}$ by (37) and (38): $\mathbf{g}_{k+1} = \mathrm{grad} f(\mathbf{x}_{k+1})$.
6 Calculate vector transports $\bar{\mathbf{g}}_k$ of gradient $\mathbf{g}_k$, and $\bar{\mathbf{d}}_k$ of conjugate direction $\mathbf{d}_k$ by (41): $\bar{\mathbf{g}}_k = \mathrm{Trnsprt}_{\mathbf{x}_k \to \mathbf{x}_{k+1}}(\mathbf{g}_k)$, $\bar{\mathbf{d}}_k = \mathrm{Trnsprt}_{\mathbf{x}_k \to \mathbf{x}_{k+1}}(\mathbf{d}_k)$
7 Calculate β_k according to **Flecther-Reeves equation**: $\beta_k = (\mathbf{g}_{k+1}^H \mathbf{g}_{k+1})/(\bar{\mathbf{g}}_k^H \bar{\mathbf{g}}_k)$.
8 Compute conjugate direction $\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \bar{\mathbf{d}}_k$.
9 Update $k \leftarrow k + 1$.
10 **until** *convergence*;
11 Return $\mathbf{x} = \mathbf{x}_{k+1}$.

Transport: Vectors in different tangent spaces cannot be combined directly. To determine the step parameter of the following algorithm, a mapping between a tangent vector in different tangent spaces is need. We call such mapping operation as transport. The transport of a tangent vector $\mathbf{d}$ from the tangent space $T_{\mathbf{x}_k}\mathcal{M}_{cc}^N$ into $T_{\mathbf{x}_{k+1}}\mathcal{M}_{cc}^N$ is expressed as

$$\mathrm{Trnsprt}_{\mathbf{x}_k \to \mathbf{x}_{k+1}}(\mathbf{d}) = \mathbf{d} - \mathrm{Re}\{\mathbf{d} \circ \mathbf{x}_{k+1}^*\} \circ \mathbf{x}_{k+1}. \tag{41}$$

Bearing the Riemannian gradient, retraction and transport in mind, we are able to develop a conjugate gradient method in Riemannian space. The procedures of obtaining the passive beamformer $\{\theta_n\}$ based on manifold optimization as well as conjugate gradient method are listed in Algorithm 1. Algorithm 1 utilizes the well-known Goldstein criterion and Flecther-Reeves equation to guarantee the objective function to be non-increasing in each iteration. According to Theorem 4.3.1 in [48], Algorithm 1 is guaranteed to converge to a critical point, i.e., at this point the gradient of the objective function is zero.

D. Overall Algorithm and Complexity Analysis

Based on the above analysis, the overall alternating optimization algorithm proposed in this paper is summarized in Algorithm 2. At first, the content placement problem is solved in Step 2. Step 3 to Step 11 are performed to optimize active and passive beamforming alternately. Step 5 to Step 8 invoke the block coordinate descent method to deal with the passive beamforming integrally, where in each iteration, Step 6 exploits bisection search to solve the auxiliary variables, and

Algorithm 2 Alternating Optimization Algorithm

Input: $\{\mathbf{h}_{d,k}\}$, $\{\mathbf{h}_{r,k}\}$, $\mathbf{G}$ and $\left\{R_k^0\right\}$
Output: $\{c_f\}$, $\{\mathbf{p}_k\}$ and $\{\theta_n\}$

Randomly construct the initial point $\left\{\mathbf{p}_k^{(0)}\right\}$ and $\left\{\theta_k^{(0)}\right\}$; Set the convergence tolerance $\epsilon > 0$ and iteration index $t = 0$ for the outer layer.
Find $\{c_f\}$ by solving problem (5).
repeat
  Solve (9) to obtain $\mathbf{p}_k^{(t+1)}$ for $\theta_n = \theta_n^{(t)}$.
  repeat
    Update the auxiliary variables $\{x_{k,j}\}$ by solving (25).
    Update the complex circle manifold vector $\mathbf{x}$ by running Algorithm 1.
  until *The decrease of the objective value of* (24) *is below a threshold* $\delta > 0$;
  $\theta_n^{(t+1)} = -j \ln \{\mathbf{x}\}_n$.
  Update $t \leftarrow t + 1$.
until *The decrease of the objective value of* (4) *is below the threshold* ϵ;
Return $\{c_f\}$, $\left\{\mathbf{p}_k = \mathbf{p}_k^{(t+1)}\right\}$ and $\left\{\theta_n = \theta_n^{(t+1)}\right\}$.

Step 7 resorts to conjugate gradient algorithm to compute the present phase shifts.

Let us analyze the complexity of the overall algorithm. The SDP problem (9) can be solved by iterative optimization techniques, e.g., interior-point method (IPM). A worst-case complexity result to solve the SDP is given by [51]

$$\mathcal{O}\Big(\sqrt{\sum_{i=1}^{N_{sdp}} n_i^{sdp} + m} \times \left(\sum_{i=1}^{N_{sdp}} \left(n_i^{sdp}\right)^3 + \sum_{i=1}^{N_{sdp}} \left(n_i^{sdp}\right)^2 m + m^3\right) \log\left(1/\varepsilon_1\right)\Big) \tag{42}$$

where N_{sdp} is the number of SDP cone constraints, n_i^{sdp} is the dimension of the i-th SDP cone, m is the number of constraints, and ε_1 is the accuracy of the convex optimization solution. With regards to the SDP considered in this paper, its worst-cast complexity is correspondingly given by $\mathcal{O}\left(\sqrt{KM+K}\left(KM^3 + K^2M^2 + K^3\right)\log\left(1/\varepsilon_1\right)\right)$, i.e., the complexity of Step 4 in Algorithm 2. For the Step 6 in Algorithm 2, it can be shown that the complexity of solving (25) is $\mathcal{O}\left(K^2 \log\left(1/\varepsilon_2\right)\right)$, where ε_2 is the accuracy for the bisection search. For the conjugate gradient algorithm based on manifold optimization, i.e., Algorithm 1, for each iteration, the complexity mainly depends on the calculation of the Riemannian gradient, which is given by $\mathcal{O}\left(N^2\right)$. Denote the iteration number required by the conjugate gradient algorithm as T. Then the complexity of Step 7 in Algorithm 2 is given by $\mathcal{O}\left(TN^2\right)$. Thus, the overall complexity of Algorithm 2 can be written as

$$\mathcal{O}\big(\, I_{out}\sqrt{KM+K}\left(KM^3 + K^2M^2 + K^3\right) \times \log\left(1/\varepsilon_1\right) I_{in}\left(K^2 \log\left(1/\varepsilon_2\right) + TN^2\right)\big)$$

where I_{in} and I_{out} denote the outer and inner iteration numbers of Algorithm 2 required for convergence, respectively.

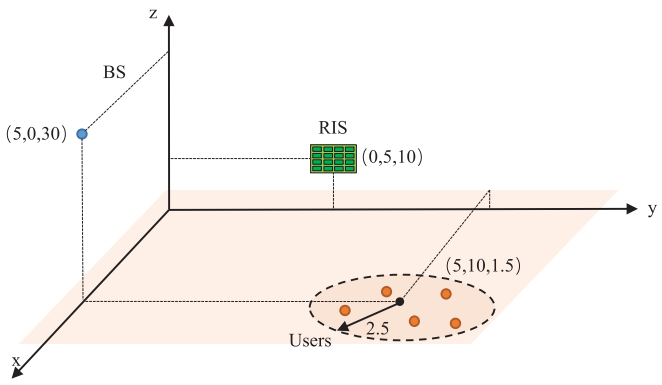


Fig. 2. Simulation scenario.

V. Simulation Analysis

A. Simulation Settings

We consider a system operating on a carrier frequency of 2.4 GHz. The noise power spectral density is given by $\sigma_0^2 = -150$ dBm/Hz, and bandwidth is $B = 10$ MHz. As shown in Fig. 2, a three-dimensional (3D) coordinate system is considered, where the BS is equipped with a uniform linear array (ULA) located on the x-axis, and the RIS is configured with a uniform planar array (UPA) located on the $y-z$ plane, respectively. The reference locations of the BS and the RIS are set at $(5, 0, 30)$ meters and $(0, 5, 10)$ meters, respectively. The antenna spacing is half wavelength. The served users $k \in \mathcal{K}$ are randomly and uniformly distributed in a circle region centered at $(5, 10, 1.5)$ with a radius of 2.5 m. Here the z-coordinates indicate that the heights of BS, RIS, and users are assumed to be 30 m, 10 m, and 1.5 m, respectively. The number of transmit antennas at the BS is $M = 16$, and the number of serving users is $K = 5$. We consider $F = 1000$ files in the database, and $S_0 = 100$ for the local storage size at the BS. Unless otherwise stated, the number of the reflection elements is $N = 50$, and Zipf parameter is set to $\varepsilon = 1$.

The distances for the direct BS-user link, the BS-RIS link and the RIS-user link are denoted by $d_{d,k}$, d_G and $d_{r,k}$, respectively. The distance-dependent path loss for all channels is modeled as $PL(d) = \rho_0\left(\frac{d}{d_0}\right)^{-\alpha}$, where $\rho_0 = -30$ dB denotes the path loss at the reference distance $d_0 = 1$ m [34], d denotes the link distance, and α denotes the path loss exponent. For small scale fading, the Rayleigh fading and the Rician fading models are assumed for the direct BS-user link and the BS-RIS/RIS-user links, respectively [36]. Then,

the corresponding channel coefficients can be expressed as

$$\mathbf{h}_{d,k} = \sqrt{PL(d_{d,k})}\mathbf{h}_{d,k}^{NLOS} \tag{43}$$

$$\mathbf{G} = \sqrt{\frac{PL(d_G)}{K_G+1}}\left(\sqrt{K_G}\mathbf{G}^{LOS} + \mathbf{G}^{NLOS}\right) \tag{44}$$

$$\mathbf{h}_{r,k} = \sqrt{\frac{PL(d_{r,k})}{K_{r,k}+1}}\left(\sqrt{K_{r,k}}\mathbf{h}_{r,k}^{LOS} + \mathbf{h}_{r,k}^{NLOS}\right) \tag{45}$$

where K_G and $K_{r,k}$ denote the related Rician factors, $\mathbf{G}^{LOS}$ and $\mathbf{h}_{r,k}^{LOS}$ denote the deterministic line-of-sight (LoS) components, $\mathbf{h}_{d,k}^{NLOS}$, $\mathbf{G}^{NLOS}$ and $\mathbf{h}_{r,k}^{NLOS}$ denote the non-LoS fading components. More specifically, the non-LoS component is modeled as Rayleigh fading, while the LoS component is modeled as the product of the array response vectors of the transceivers [34], [36]. For instance, $\mathbf{G}^{LOS}$ is given by

$$\mathbf{G}^{LOS} = \mathbf{a}_{\mathrm{G}}(\theta)\,\mathbf{a}_{\mathrm{BS}}^{H}(\phi)$$

with

$$\mathbf{a}_{\mathrm{G}}(\theta) = \left[1, e^{j\frac{2\pi}{\lambda}d\sin(\theta)}, \cdots, e^{j\frac{2\pi}{\lambda}(N-1)d\sin(\theta)}\right]^T$$

$$\mathbf{a}_{\mathrm{BS}}(\phi) = \left[1, e^{j\frac{2\pi}{\lambda}d\sin(\phi)}, \cdots, e^{j\frac{2\pi}{\lambda}(M-1)d\sin(\phi)}\right]^T$$

where θ and ϕ are the angels of arrival and departure (AoA/AoD), respectively, λ is the signal wavelength, and d is the distance between antenna elements. In this paper, without otherwise specified, the path loss exponents for the direct link, BS-RIS and RIS-user link are set to be $\alpha_{d,k} = 3.5$, $\alpha_G = 2.2$ and $\alpha_{r,k} = 2.2$, respectively. The Rician factors are $K_G = K_{r,k} = 3$ dB. The target SINR requirements for different users are all set to $\gamma_0 = 30$ dB, which is related to the content-delivery rate of 100 Mbps. The convergence tolerance is set to $\epsilon = 10^{-2}$. All simulation results are obtained by averaging over 1000 independent realizations.

The following heuristic caching strategies are considered in simulations.

- *Uniform random caching (URC)*: In each realization, BS caches the content files randomly with equal probabilities regardless of their popularity distribution.
- *File popularity based probabilistic caching (FPPC)*: In each realization, BS caches a content randomly with probability depending on the content popularity, and the more popular the content is, the more likely it will be cached.
- *Optimized caching (OC)*: In each realization, the content placement is optimized by solving the program $\mathcal{P}_1$.

The following three communication schemes are considered for comparisons.

- *Optimized hybrid beamforming*: The active beamforming at BS and the phase shifts of the reflecting elements on RIS are optimized by invoking Algorithms 1 & 2.
- *Random phase at RIS*: The phase shifts of the reflecting elements on RIS are randomly generated, whilst the active beamforming is designed by solving the program $\mathcal{P}_{2-\mathrm{I}}$.
- *Without RIS*: The reflecting path $\mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G}$ is set to zero. The active beamforming is designed by solving the program $\mathcal{P}_{2-\mathrm{I}}$.

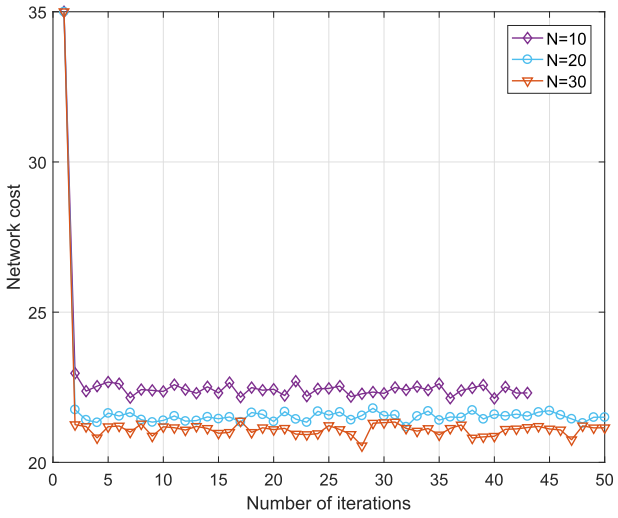


Fig. 3. Convergence performance of Algorithm 2 under different numbers of phase shift elements.

B. Simulation Results

First we disclose the convergence performance of Algorithm 2 in a single channel realization. The initial active beamforming at BS is set with equal coefficients while satisfying the maximum transmit power of 40 dBm. The initial phase shifts of RIS are generated uniformly between $[0, 2\pi)$. Besides, at the beginning, no caching policy is employed in the system. To trade the backhaul capacity for power consumption, we set the pricing factor to $\eta = 10^{-7}$. Figure 3 shows the total network cost versus the number of iterations under different number of RIS elements. It can be observed that our proposed algorithm is capable of realizing a significant decrease on the network cost within two iterations, and then keeps iterating to convergence to satisfy the precision requirement.

Figure 4 compares the network cost performance of three communication schemes versus the number of reflecting elements N, where the OC strategy is applied in these schemes. Specifically, it is seen that the RIS scheme with optimized hybrid beamforming has the lowest network cost, since the quality of the offloading links has been improved by the tailored phase shifts. In contrast, the scheme without RIS endures the highest network cost. For the pair of schemes with RIS, the network cost can be reduced by increasing N. Notice that installing more passive reflecting elements is practical, and both energy and cost efficient, since RISs do not need radio frequency chains, and are compatible with the hardware of existing wireless networks. Besides, it is observed that the performance of random phase is inferior to that of the RIS scheme with optimized phase shifts. This is because the reflected signals have not been constructively added towards the target receivers. By contrast, with aid of our proposed hybrid beamforming design, the network cost has been decreased dramatically.

Figure 5 compares the backhaul capacity of caching strategies versus the Zipf exponent ε. Clearly, compared with the benchmark of no caching, all caching schemes achieve much lower backhaul capacity cost. Notice that with the increase of the Zipf exponent, both of the content popularity and the

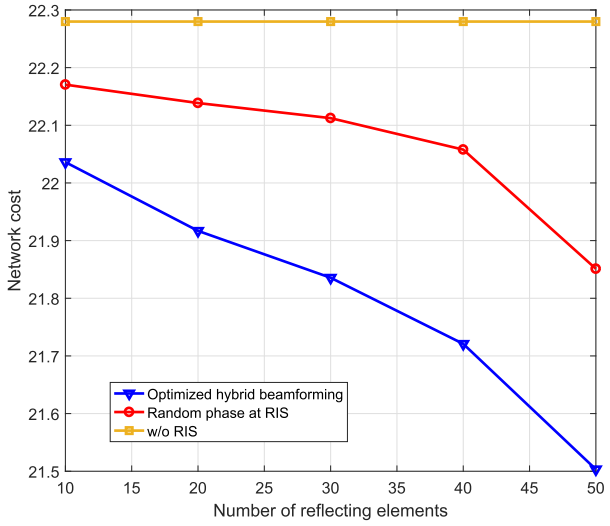


Fig. 4. Network cost versus the number of reflecting elements.

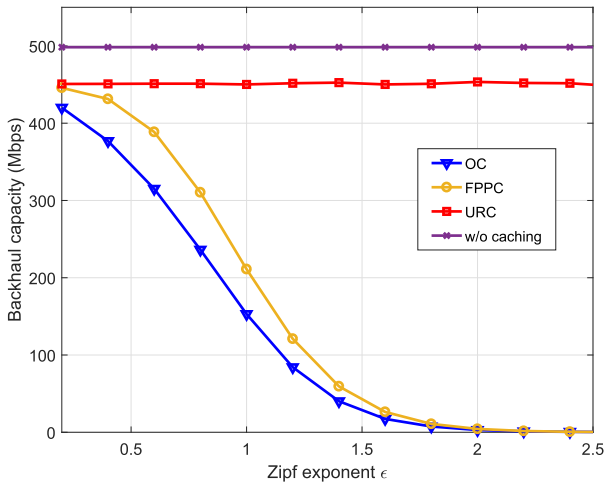


Fig. 5. Backhaul capacity versus Zipf exponent.

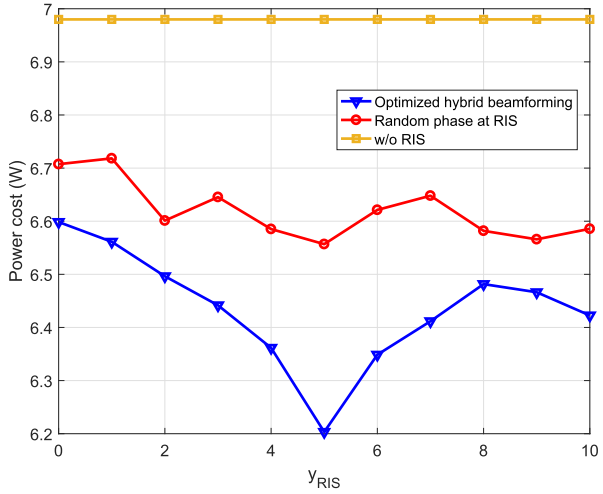


Fig. 6. Power cost versus the RIS location.

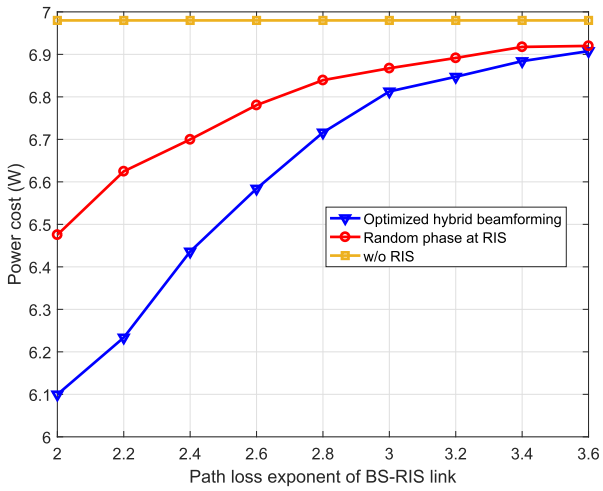


Fig. 7. Power cost versus the path loss exponent of BS-RIS link.

user requests are concentrated on fewer files. It is observable from Fig. 5 that, upon increasing Zipf exponent, the backhaul capacity of URC strategy stays the same value. This is because in URC strategy, BS caches content files randomly with equal probabilities regardless of the content popularity or the request distribution. By contrast, the backhaul capacities of FPPC and OC keep shrinking with the increase of the Zipf exponent. Notably, our proposed OC strategy achieves the lowest backhaul cost, demonstrating that our proposed content placement design is capable of reaping the benefit of skewed content popularity and user request distribution.

Denote the coordinate in y-axis of RIS as y_{RIS}. In Fig. 6, we study the impact of the RIS location by moving the RIS from $y_{\mathrm{RIS}} = 0$ m to $y_{\mathrm{RIS}} = 10$ m. It is observed that the power consumed by the RIS scheme first decreases with y_{RIS}, and then increases for $y_{\mathrm{RIS}} > 5$ m. This is because the large scale path loss of BS-RIS-user links is $PL(d_G)PL(d_{r,k})$, which achieves its minimum around the middle of the y-coordinate and can be calculated accurately in the 3D coordinate system. This observation can be also explained intuitively as the specular reflection achieved among transmitter, RIS and receivers. Besides, as RIS is very close to the destination in y-axis, the power cost decreases slightly, since the RIS-user link becomes much better. By contrast, moving the RIS location along with the y-axis fails to bring an obvious variation trend to the random phase scheme. This is because, with randomly generated phase shifts, the RIS cannot reflect signals to beam effectively towards the destination, not to mention adapting to the RIS location. Nevertheless, both RIS schemes achieve better power cost performance than the no RIS one.

In previous simulations, the path loss exponents of the RIS-related links are set to $\alpha_G = 2.2$ and $\alpha_{r,k} = 2.2$, which are close to the propagation conditions in free space. However, it may become impractical in some specific scenarios with certain obstacles. Hence, we are motivated to investigate the impact of the RIS-related path loss exponents on the achievable performance. In Figs. 7 and 8, the impacts of the path loss exponents of BS-RIS and RIS-user links are depicted, respectively. It is observed that the power cost achieved by the proposed scheme increases upon increasing α_G and $\alpha_{r,k}$,

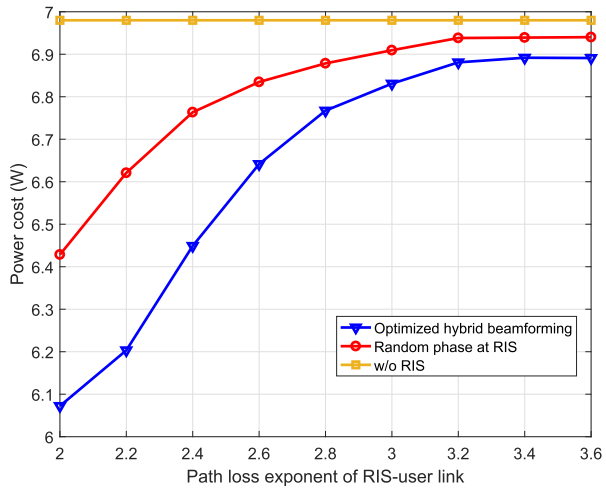


Fig. 8. Power cost versus the path loss exponent of RIS-user link.

and finally approaches the power cost as achieved by the no RIS scheme. This is because, upon increasing α_G and $\alpha_{r,k}$, the signal attenuation associated with the RIS-related links becomes larger, leading to more power consumption at the transmitter. However, when α_G and $\alpha_{r,k}$ are small, a significant performance gain can be achieved by our proposed scheme compared with random phase and no RIS schemes. This observation reminds us that an RIS had better be deployed in a relatively open scenario with low density of obstacles and low fraction of energy each object absorbs, so that a relatively small path loss exponent is confronted [52].

VI. Conclusion

In this paper, an RIS-aided edge caching system has been considered, where a network cost minimization problem has been formulated to optimize the content placement and hybrid beamformer. After decoupling the content placement subproblem with the hybrid beamforming design, we have proposed an alternating optimization algorithm to tackle the active beamforming and passive phase shifting. For active beamforming, we have transferred the problem into an SDP by applying SDR. For passive phase shifting, we have introduced block coordinate descent method to alternately optimize the auxiliary variables and the RIS phase shifts. Further, a conjugate gradient algorithm based on manifold optimization has been proposed to deal with the non-convex unit-modulus constraints in the passive phase shifting design. Numerical results have showed that our RIS-aided edge caching design can effectively decrease the network cost in terms of backhaul capacity and power consumption, compared with existing caching strategies, random phase shifting scheme, and no RIS counterpart.

References

[1] Y. Chen, M. Wen, E. Basar, Y.-C. Wu, L. Wang, and W. Liu, "Network cost minimization for reconfigurable intelligent surface aided edge caching," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Mar. 2021, pp. 1–6.

[2] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[3] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "The role of caching in future communication systems and networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1111–1125, Jun. 2018.

[4] Y. Li, M. Xia, and Y.-C. Wu, "Caching at base stations with multi-cluster multicast wireless backhaul via accelerated first-order algorithms," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 2920–2933, May 2020.

[5] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[6] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 80–87, Jun. 2018.

[7] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2809–2825, Jul. 2018.

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

[9] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1714–1724, Mar. 2019.

[10] Y. Liu *et al.*, "Reconfigurable intelligent surfaces: Principles and opportunities," 2020, *arXiv:2007.03435*. [Online]. Available: http://arxiv.org/abs/2007.03435

[11] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[12] M. D. Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, Dec. 2019.

[13] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[14] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[15] Y. Chen, L. Wang, R. Ma, W. Liu, M. Wen, and A. Fei, "Performance analysis of heterogeneous networks with wireless caching and full duplex relaying," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2429–2442, Oct. 2020.

[16] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2699–2714, Jun. 2017.

[17] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Oct. 2016.

[18] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[19] Y. Chen, L. Wang, R. Ma, B. Jiao, and L. Hanzo, "Cooperative full duplex content sensing and delivery improves the offloading probability of D2D caching," *IEEE Access*, vol. 7, pp. 29076–29084, 2019.

[20] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "D2D-aware device caching in mmWave-cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2025–2037, Sep. 2017.

[21] E. Chen, M. Tao, and N. Zhang, "User-centric joint access-backhaul design for full-duplex self-backhauled wireless networks," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7980–7993, Nov. 2019.

[22] E. Chen, M. Tao, and Y.-F. Liu, "Joint base station clustering and beamforming for non-orthogonal multicast and unicast transmission with backhaul constraints," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6265–6279, Sep. 2018.

[23] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4854–4876, Oct. 2018.

[24] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[25] B. Hu, L. Fang, X. Cheng, and L. Yang, "In-vehicle caching (IV-cache) via dynamic distributed storage relay (D^2SR) in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 843–855, Jan. 2019.

[26] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[27] B. Ning, Z. Chen, W. Chen, and J. Fang, "Beamforming optimization for intelligent reflecting surface assisted MIMO: A sum-path-gain maximization approach," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 1105–1109, Jul. 2020.

[28] C. Huang, A. Zappone, M. Debbah, and C. Yuen, "Achievable rate maximization by passive intelligent mirrors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 3714–3718.

[29] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[30] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[31] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[32] Z. Yigit, E. Basar, and I. Altunbas, "Low complexity adaptation for reconfigurable intelligent surface-based MIMO systems," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2946–2950, Dec. 2020.

[33] B. Lyu, D. T. Hoang, S. Gong, D. Niyato, and D. I. Kim, "IRS-based wireless jamming attacks: When jammers can attack without power," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1663–1667, Oct. 2020.

[34] C. Pan *et al.*, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

[35] T. Hou, Y. Liu, Z. Song, X. Sun, and Y. Chen, "MIMO-NOMA networks relying on reconfigurable intelligent surface: A signal cancellation-based design," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6932–6944, Nov. 2020.

[36] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, Oct. 2020.

[37] B. Zheng, Q. Wu, and R. Zhang, "Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA?" *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 753–757, Apr. 2020.

[38] 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.

[39] C. Pan *et al.*, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.

[40] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted SWIPT under QoS constraints," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1735–1748, Aug. 2020.

[41] L. Yang, F. Meng, J. Zhang, M. O. Hasna, and M. D. Renzo, "On the performance of RIS-assisted dual-hop UAV communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10385–10390, Sep. 2020.

[42] A. U. Makarfi *et al.*, "Reconfigurable intelligent surfaces-enabled vehicular networks: A physical layer security perspective," 2020, *arXiv:2004.11288*. [Online]. Available: http://arxiv.org/abs/2004.11288

[43] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," 2019, *arXiv:1904.10136*. [Online]. Available: http://arxiv.org/abs/1904.10136

[44] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, early access, Mar. 2, 2021.

[45] S. Gao, P. Dong, Z. Pan, and G. Y. Li, "Deep multi-stage CSI acquisition for reconfigurable intelligent surface aided MIMO systems," *IEEE Commun. Lett.*, early access, Mar. 2, 2021.

[46] C.-Y. Chi, W.-C. Li and C.-H. Lin. *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications*. Boca Raton, FL, USA: CRC Press, 2017.

[47] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[48] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

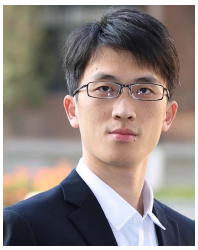
[49] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for secure IRS-assisted multiuser MISO systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2019, pp. 1–6.

[50] A. Hjorungnes, *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[51] P. Biswas *et al.*, "Semidefinite programming based algorithms for sensor network localization," *ACM Trans. Sensor Netw.*, vol. 2, no. 2, pp. 188–220, 2006.

[52] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

Yingyang Chen (Member, IEEE) received the B.Eng. degree in electronic engineering from the Yingcai Honors College, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2014, and the Ph.D. degree in signal and information processing from Peking University, Beijing, China, in 2019. From March 2018 to September 2018, she worked as a Visiting Student with the Next Generation Wireless Group, University of Southampton, supervised by Prof. Lajos Hanzo. She is currently an Assistant Professor with the College of Information Science and Technology, Jinan University, Guangzhou, China. Her research interests mainly focus on the performance analysis and optimization in wireless communications, mobile edge caching and computing, and signal processing.



Miaowen Wen (Senior Member, IEEE) received the Ph.D. degree from Peking University, Beijing, China, in 2014.

From 2019 to 2021, he was with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, as a Post-Doctoral Research Fellow. He is currently an Associate Professor with the South China University of Technology, Guangzhou, China. He has published two books and more than 120 journal articles. His research interest includes a variety of topics in the areas of wireless and molecular communications.

Dr. Wen was a recipient of the IEEE Asia-Pacific (AP) Outstanding Young Researcher Award in 2020, and four Best Paper Awards from the IEEE ITST'12, the IEEE ITSC'14, the IEEE ICNC'16, and the IEEE ICCT'19. He was the winner in data bakeoff competition (Molecular MIMO) at IEEE Communication Theory Workshop (CTW) 2019, Selfoss, Iceland. He served as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is currently serving as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL, AND MULTI-SCALE COMMUNICATIONS, and IEEE COMMUNICATIONS LETTERS, and a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (Special Issue on Advanced Signal Processing for Local and Private 5G Networks).

Ertugrul Basar (Senior Member, IEEE) received the B.S. degree (Hons.) from Istanbul University, Istanbul, Turkey, in 2007, and the M.S. and Ph.D. degrees from Istanbul Technical University, Istanbul, in 2009 and 2013, respectively. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Koç University, Istanbul, and the Director of Communications Research and Innovation Laboratory (CoreLab). His primary research interests include MIMO systems, index modulation, intelligent surfaces, waveform design, visible light communications, and signal processing for communications. He is the Academic Chair of IEEE ComSoc Emerging Technologies Initiative on Reconfigurable Intelligent Surfaces. He currently serves as a Senior Editor for IEEE COMMUNICATIONS LETTERS and the Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and *Frontiers in Communications and Networks*.

Yik-Chung Wu (Senior Member, IEEE) received the B.Eng. (EEE) and M.Phil. degrees from The University of Hong Kong (HKU), in 1998 and 2001, respectively, and the Ph.D. degree from Texas A&M University, College Station, USA, in 2005. From 2005 to 2006, he was with the Thomson Corporate Research, Princeton, NJ, USA, as a member of Technical Staff. Since 2006, he has been with HKU, where he is currently an Associate Professor. He was a Visiting Scholar with Princeton University in Summer 2015 and in Summer 2017. His research interests include general areas of signal processing, machine learning, and communication systems. He served as an Editor for IEEE COMMUNICATIONS LETTERS and IEEE TRANSACTIONS ON COMMUNICATIONS. He is currently an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Editor of *Journal of Communications and Networks*.

Li Wang (Senior Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2009.

She is currently a Full Professor with the School of Computer Science National Pilot Software Engineering School, BUPT, where she is also an Associate Dean and the Head of the High Performance Computing and Networking Laboratory. She is also a member of the Key Laboratory of the Universal Wireless Communications, Ministry of Education, Beijing. She also held a visiting positions with the School of Electrical and Computer Engineering, Georgia Tech, Atlanta, GA, USA, from December 2013 to January 2015, and Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, from August 2015 to November 2015 and July 2018 to August 2018. She has authored or coauthored almost 50 journal articles and two books. Her current research interests include wireless communications, distributed networking and storage, vehicular communications, social networks, and edge AI.

Dr. Wang was a recipient of the 2013 Beijing Young Elite Faculty for Higher Education Award, Best Paper Awards from several IEEE conferences, including IEEE ICCC 2017, IEEE GLOBECOM 2018, and IEEE WCSP 2019. She was also a recipient of the Beijing Technology Rising Star Award in 2018. She has served on TPC of multiple IEEE conferences, including IEEE INFOCOM, GLOBECOM, International Conference on Communications, IEEE Wireless Communications and Networking Conference, and IEEE Vehicular Technology Conference in recent years. She was the Symposium Chair of IEEE ICC 2019 on Cognitive Radio and Networks Symposium and a Tutorial Chair of IEEE VTC 2019-Fall. She also serves as the Vice Chair of Meetings and Conference Committee (MCC) for IEEE Communication Society (ComSoc) Asia Pacific Board (APB) for the term of 2020–2021, and chairs the special interest group (SIG) on Social Behavior Driven Cognitive Radio Networks for IEEE Technical Committee on Cognitive Networks. She currently serves on the Editorial Board for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, *Computer Networks*, IEEE ACCESS, and *China Communications*.

Weiping Liu (Member, IEEE) received the Ph.D. degree from South China Normal University in 2000. He has postdoctoral research with the University of Science and Technology of China from 2001 to 2003. He is currently a Professor with the Department of Electronic Engineer, Jinan University. His research interests include optical-wireless communications and the optical fiber sensor. He has published more than 30 papers in international conferences and articles in journals in the past five years.