Joint Computation Offloading and Resource Allocation for MEC-Enabled IoT Systems With Imperfect CSI

Jun Wang, Daquan Feng, *Member, IEEE*, Shengli Zhang, *Senior Member, IEEE*, An Liu, *Senior Member, IEEE*, and Xiang-Gen Xia, *Fellow, IEEE*

***Abstract*—Mobile-edge computing (MEC) is considered as a promising technology to reduce the energy consumption (EC) and task accomplishment latency of smart mobile user equipments (UEs) by offloading computation-intensive tasks to the nearby MEC servers. However, the Quality of Experience (QoE) for computation highly depends on the wireless channel conditions when computation tasks are offloaded to MEC servers. In this article, by considering the imperfect channel-state information (CSI), we study the joint offloading decision, transmit power, and computation resources to minimize the weighted sum of EC of all UEs while guaranteeing the probabilistic constraint in multiuser MEC-enabled Internet-of-Things (IoT) networks. This formulated optimization problem is a stochastic mixed-integer nonconvex problem and challenging to solve. To deal with it, we develop a low-complexity two-stage algorithm. In the first stage, we solve the relaxed version of the original problem to obtain offloading priorities of all UEs. In the second stage, we solve an iterative optimization problem to obtain a suboptimal offloading decision. As both stages include solving a series of nonconvex stochastic problems, we present a constrained stochastic successive convex approximation-based algorithm to obtain a near-optimal solution with low complexity. The numerical results demonstrate that the proposed algorithm provides comparable performance to existing approaches.**

***Index Terms*—Computation offloading, optimization, resource allocation, stochastic programming.**

Manuscript received February 15, 2020; revised June 20, 2020; accepted September 2, 2020. Date of publication September 8, 2020; date of current version February 19, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61701317; in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2018QNRC001; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515130003; in part by the Innovation Project of Guangdong Educational Department under Grant 2019KTSCX147; in part by the Shenzhen Science and Technology Planning Project under Grant JCYJ20170412104656685; in part by the Tencent "Rhinoceros Birds" Scientific Research Foundation for Young Teachers of Shenzhen University; and in part by the Start-Up Fund of Peacock Project. *(Corresponding author: Daquan Feng.)*

Jun Wang, Daquan Feng, and Shengli Zhang are with the Shenzhen Key Laboratory of Digital Creative Technology, Guangdong Province Engineering Laboratory for Digital Creative Technology, Shenzhen University, Shenzhen 518060, China (e-mail: johnwangqc@gmail.com; fdquan@gmail.com; zsl@szu.edu.cn).

An Liu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: anliu@zju.edu.cn).

Xiang-Gen Xia is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: xianggen@udel.edu).

Digital Object Identifier 10.1109/JIOT.2020.3022802

I. Introduction

With the rapid growth of the Internet-of-Things (IoT) devices, various new mobile applications, such as online gaming, augmented reality (AR), and face recognition, are emerging. These applications pose new challenges to the IoT systems since they not only consume a great number of computation resources but also have a stringent delay requirement [1]–[4]. However, IoT devices, i.e., smartphones and wearable devices, are, in general, only with limited battery energy and computation resources due to their small physical sizes. To tackle this problem, mobile cloud computing (MCC) has been put forward in the last decade. In MCC systems, IoT devices offload their computation tasks via wireless links to the resource-rich remote cloud servers, which can save battery energy of IoT devices [5].

However, in MCC, the execution delay may be very large because of long distance and heavy network congestion between IoT devices and cloud servers. This is unacceptable for delay-sensitive applications. To address this issue, mobile-edge computing (MEC) has been proposed to deliver the cloud functionalities, i.e., computing and storage, to the edge of wireless networks [6], [7]. The existing works [8], [9] have shown that the computation experience of IoT devices can be significantly improved when offloading the computation tasks to MEC servers.

Recently, the MEC paradigm has attracted significant attention from both academia and industry. Earlier works on MEC have been focused on offloading platforms, such as MAUI [10], CloneCloud [11], ThinkAir [12], and CONCERT [13]. Unlike traditional cloud servers with rich resources, MEC servers may be resource limited. Therefore, it is crucial to optimize computation offloading strategies and resource allocation for better Quality of Experience (QoE) of IoT devices. Thus, computation offloading in MEC systems has recently attracted increasing attention.

For a single-user MEC, computation offloading strategy has been widely considered [14]–[17]. In [14], a binary offloading framework has been proposed for minimizing the energy consumption (EC) under a stochastic wireless channel by dynamically adjusting the local CPU frequency for mobile execution and scheduling the data transmission rate for cloud execution. Considering the divisibility of computation tasks, partial offloading decision design has been studied in [15],

where the offloading ratio, local CPU frequency, and transmit power of smart mobile user equipment (UE) are jointly optimized. Dinh *et al.* [16] have proposed a binary offloading scheme to minimize the weighted sum of EC and execution latency by jointly optimizing offloading decision and UE's CPU frequency, where a UE has multiple computation tasks to offload simultaneously to multiple MEC servers. In [17], a dynamic Lyapunov optimization-based binary offloading algorithm has been proposed for MEC systems with an energy harvesting (EH) device.

For multiuser MEC, a large number of works have also been devoted to computation offloading [5], [18]–[31]. In [18], a joint optimization of partial offloading and resource allocation has been proposed to minimize the EC for multiuser MEC systems with TDMA. Different from the partial offloading, binary offloading designs in multiuser MEC are mixed-integer optimization problems that are usually nonconvex. In [5] and [19], binary offloading mechanisms based on game theory have been proposed. In [20], the problem of the total EC minimization has been formulated by optimizing transmit power for multiuser MIMO MEC systems. Note that in [5], [19], and [20], either offloading decision or resource allocation was optimized separately. In [21], joint optimization of offloading decision and resource allocation has been developed to minimize the total EC under latency requirements. In [22]–[24], a joint optimization framework of binary offloading decision and resource allocation in heterogeneous networks (HetNets) with MEC has been investigated. These works have been further extended to the multitask offloading in HetNets in [25], where the authors have considered the joint computation offloading, user association, and resource allocation problem and proposed an alternating optimization framework to solve it. Du *et al.* [26] have considered a hierarchical fog-cloud computing system and studied a min–max fairness problem under binary offloading decision constraint. Meanwhile, some works have also considered multiuser MEC systems with EH [27], [28]. In [27], a joint optimization framework of offloading ratio, energy beamforming, and computing resource has been developed to minimize the total EC. Different from [27], Bi and Zhang [28] have studied the weighted sum of computation rate maximization by jointly optimizing the binary offloading and the transmission time allocation. Besides, Mao *et al.* [29] have investigated stochastic resource allocation in multiuser MEC systems and proposed a low-complexity Lyapunov optimization-based online algorithm. More recently, some works [30], [31] have proposed an offloading scheme based on machine learning. A reinforcement learning-based offloading policy for an IoT device has been proposed in [30] while an after-state learning-based computation offloading scheme has been designed for MEC systems with EH in [31].

However, the aforementioned works assume that the perfect channel-state information (CSI) is known at the base station (BS). This assumption may be unrealistic in practical systems due to the estimation errors, limited CSI feedback quantization, delays, etc. On the other hand, the QoE of computation heavily relies on the wireless fading channel conditions since task offloading requires effective wireless transmission. Nguyen *et al.* [32] have studied the fairness problem for MEC systems with imperfect CSI with the goal to minimize the maximum weighted EC under the upper bound of the average latency constraint. Different from it, our goal is to minimize the weighted sum of EC of UEs while guaranteeing the probabilistic delay constraint for MEC systems with imperfect CSI, which seeks to provide "safe" performance guaranteeing for a certain probability (often high) of satisfying the delay requirements. Our formulated problem is the mixed-integer nonconvex stochastic optimization problem under the probabilistic constraint that is more difficult to solve. Note that Wang *et al.* [33] have presented a conservative approach to solve the robust transmit beamforming design problem under the probabilistic signal-to-interference-plus-noise ratio (SINR) constraint for multiuser wireless systems. Their idea is to replace the difficult probabilistic constraint by the worst case constraint with spherically bounded channel errors. However, this idea cannot be applied to our formulated problem, where the objective function involves the expectation and has no closed-form expression. To deal with this issue, we propose a low-complexity two-stage optimization framework based on constrained stochastic successive convex approximation (CSSCA) to solve the formulated problem. The main contributions of this article are summarized as follows.

1) By taking imperfect CSI into account, we formulate the average weighted sum of EC minimization problem under the probabilistic delay constraint and jointly optimize the offloading decision, UEs' transmit power, and computation resource allocation.
2) To deal with the mixed-integer nature of the formulated problem, we develop a novel low-complexity optimization framework. It includes two stages. In the first stage, we solve the relaxed version of the original problem to obtain offloading priorities of all UEs. In the second stage, we iteratively solve a stochastic optimization problem to obtain a suboptimal offloading decision and corresponding resource allocation.
3) Both stages include solving a series of nonconvex stochastic problems. One of the difficulties in the stochastic optimization problem is that the objective function involves the expectation and has no closed-form expression. To tackle this challenge, we propose a CSSCA-based iterative algorithm, in which a quadratic convex problem is solved at each iteration by a low-complexity algorithm based on the dual decomposition method. In addition, we have also analyzed the convergence and complexity of the proposed algorithm.
4) We perform extensive numerical simulations to evaluate the convergence of the proposed iterative algorithm and compare the performance of the proposed solution with the existing approaches.

The remainder of this article is organized as follows. In Section II, the system model is first described and the optimization problem is then formulated. In Section III, the low-complexity framework to solve the mixed-integer nonconvex stochastic optimization problem is presented. Convergence and complexity analysis of the proposed algorithm are shown

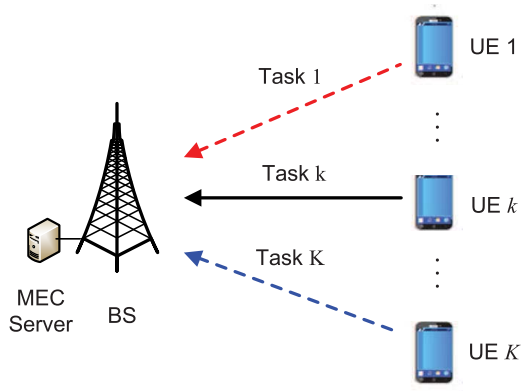


Fig. 1. Multiuser MEC system.

in Section IV. In Sections V and VI, simulation results and conclusion are, respectively, presented.

Notation: $\mathbb{E}[\cdot]$ presents the mathematical expectation. $\Pr\{\cdot\}$ denotes the probability operator. $[x]_a^b \triangleq \min\{b, \max\{a, x\}\}$.

II. System Model and Problem Transformation

In this section, we introduce the communication model and computation model of a multiuser MEC system, followed by the optimization problem formulation.

A. Communication Model

Consider a multiuser MEC system described in Fig. 1, where K UEs are served by a BS equipped with the MEC server. For simplicity, we will use the MEC server and BS interchangeably. Denote the set of UEs as $\mathcal{K} = \{1, 2, \ldots, K\}$. Without loss of generality, assume that each UE has only one computation-intensive task that is required to be accomplished within a certain time.

To avoid the mutual interference among UEs, we suppose that each UE occupies an orthogonal quasistatic wireless channel when offloading its task. Let h_k be the channel fading coefficient between UE k and BS. However, in practical wireless systems, the accurate CSI at BS is difficult to obtain owing to the channel estimation error or the feedback delay. Thus, in this article, the actual CSI h_k is modeled as

$$h_k = \widehat{h}_k + e_k \tag{1}$$

where $\widehat{h}_k$ is a channel estimate and e_k is the estimation error. Assume that e_k obeys a circularly symmetric complex Gaussian distribution, i.e., $e_k \sim \mathcal{CN}(0, \sigma_{e_k}^2)$. Then, the received signal associated with UE k is given by

$$y_k = \sqrt{p_k}\widehat{h}_k s_k + \underbrace{\sqrt{p_k} e_k s_k + n_k}_{n_k'} \tag{2}$$

where p_k is the transmit power of UE k, s_k is the transmitted symbol from UE k satisfying $\mathbb{E}[|s_k|^2] = 1$, n_k is the additive white Gaussian noise with power spectral density N_0, n_k' is the effective noise, which combines the additive noise and residual channel estimation error, and its variance is

$$\mathbb{E}\left[|n'_k|^2\right] = p_k|e_k|^2 + \sigma_k^2 = p_k|e_k|^2 + \theta_k B N_0 \tag{3}$$

where B is the total bandwidth, and θ_k is the normalized portion of the bandwidth to UE k.

As in [34]–[36], since the effective noise n_k' is neither independent nor Gaussian, the effective SINR for UE k is bounded below by

$$\mathrm{SINR}_k = \frac{p_k|\widehat{h}_k|^2}{p_k|e_k|^2 + \theta_k B N_0}. \tag{4}$$

Thus, the transmission rate for UE k in (2) is bounded below by

$$R_k(p_k, \theta_k, e_k) = \theta_k B \log_2(1 + \mathrm{SINR}_k). \tag{5}$$

B. Computation Model

The task of UE k to be processed can be characterized by a three-tuple $(L_k, C_k, T_k^{\max})$, where L_k is the size of task input data (in bits), C_k is the number of CPU cycles to accomplish one bit task (in CPU cycles/bit), and $T_k^{\max}$ presents the maximum tolerable delay for the task. The parameters L_k, C_k, and $T_k^{\max}$ can be measured by task profilers [37]. Assume that each task is atomic, i.e., cannot be further partitioned. That is, it can be either locally processed at the UE or transferred to the MEC server for remote processing. Let x_k be the binary offloading decision variable for UE k, where $x_k = 1$ implies that UE k chooses to offload its task to the MEC server, and $x_k = 0$ means that UE k chooses to locally process its task.

In what follows, we present the process delay and EC for both local and remote computing.

1) Local Computing: The processing time T_k^{loc} for local task executing can be written as

$$T_k^{\mathrm{loc}} = \frac{L_k C_k}{f_k^{\mathrm{loc}}} \tag{6}$$

where f_k^{loc} is the CPU computational capability of UE k (in CPU cycles/s).

As in [14] and [16], EC E_k^{loc} for local task executing can be expressed as

$$E_k^{\mathrm{loc}} = \beta_k \left(f_k^{\mathrm{loc}}\right)^3 T_k^{\mathrm{loc}} = \beta_k \left(f_k^{\mathrm{loc}}\right)^2 L_k C_k \tag{7}$$

where $\beta_k (f_k^{\mathrm{loc}})^3$ is the computational power of UE k, and β_k is a constant related to chip architecture [15], [16].

2) Remote Computing: When UE k decides to offload its task to the MEC server for processing, the offloading process includes the following three stages. First, UE k sends task input data to the MEC server over the wireless channel. Then, the MEC server processes task k by allocating part of the computation resources to it. Finally, the MEC server sends back the task output data to UE k after completing the task.

Accordingly, the uplink transmission time can be expressed by

$$T_k^{\mathrm{tr}} = \frac{L_k}{R_k(p_k, \theta_k, e_k)}. \tag{8}$$

The remote computation time can be given by

$$T_k^{\mathrm{exe}} = \frac{L_k C_k}{w_k F_s^{\max}} \tag{9}$$

where $w_k F_s^{\max}$ denotes the computation resources (in CPU cycles/s) allocated to UE k, w_k is the normalized portion of the computation resource assigned to UE k, and $F_s^{\max}$ presents the maximum computation resource of the MEC server.

As in [18], [27], and [28], the size of task output data is, in general, much smaller than that of task input data. Thus, we neglect the time consumed by returning the task output data to UEs. Therefore, the total delay for remote processing consists of two parts

$$T_k^{\text{mec}} = T_k^{\text{tr}} + T_k^{\text{exe}} = \frac{L_k}{R_k(p_k, \theta_k, e_k)} + \frac{L_k C_k}{w_k F_s^{\max}}. \tag{10}$$

The EC of UE k for remote processing is mainly the transmit energy [28], and thus can be written as

$$E_k^{\text{mec}} = p_k T_k^{\text{tr}} = \frac{p_k L_k}{R_k(p_k, \theta_k, e_k)}. \tag{11}$$

C. Problem Formulation

As introduced above, each task is either processed locally at UE or offloaded to the MEC server. Thus, the total delay for completing task k can be expressed as

$$T_k(x_k, p_k, \theta_k, w_k, e_k) = (1 - x_k) T_k^{\text{loc}} + x_k T_k^{\text{mec}}. \tag{12}$$

The total EC for completing a task can be expressed as

$$E_k(x_k, p_k, \theta_k, e_k) = (1 - x_k) E_k^{\text{loc}} + x_k E_k^{\text{mec}}. \tag{13}$$

Each UE requires its task to be completed during the maximum tolerable delay $T_k^{\max}$. However, in the presence of random CSI errors, the traditional deterministic delay requirements can no longer be guaranteed. Therefore, we model the delay requirement in the form of a chance constraint as follows:

$$\Pr_{e_k}\left\{T_k(x_k, p_k, \theta_k, w_k, e_k) \le T_k^{\max}\right\} \ge 1 - \xi_k^{\max} \tag{14}$$

where $\xi_k^{\max}$ is the maximum tolerable outage probability for task k. The inequality can be recognized as a *soft delay requirement*. To facilitate the analysis, we have the following lemma.

Lemma 1: The delay probability constraint in (14) is equivalently expressed by

$$\Theta_k(x_k, p_k, \theta_k, w_k) \le 0 \tag{15}$$

where $\Theta_k(x_k, p_k, \theta_k, w_k)$ is given in (16), shown at the bottom of the page.

Proof: See Appendix A. ∎

For a MEC system, the QoE of UE k is mainly determined by EC and its task execution delay. Considering that UEs are usually energy limited, our goal is to design an energy-efficient joint offloading decision and resource allocation scheme to minimize the *average weighted sum of EC* of UEs under the *probabilistic delay requirement*. The optimization problem can be stated as

$$\mathcal{P}_0: \min_{\mathbf{x},\mathbf{p},\boldsymbol{\theta},\mathbf{w}} \sum_{k=1}^{K} \alpha_k \mathbb{E}_{e_k}\left[E_k(x_k, p_k, \theta_k, e_k)\right] \tag{17}$$

$$\text{s.t.} \quad \Theta_k(x_k, p_k, \theta_k, w_k) \le 0 \quad \forall k \tag{17a}$$

$$x_k \in \{0, 1\} \quad \forall k \tag{17b}$$

$$0 \le p_k \le P_k^{\max} \quad \forall k \tag{17c}$$

$$0 \le \theta_k \le 1 \quad \forall k \tag{17d}$$

$$\sum_{k=1}^{K} \theta_k \le 1 \tag{17e}$$

$$0 \le w_k \le 1 \quad \forall k \tag{17f}$$

$$\sum_{k=1}^{K} w_k \le 1 \tag{17g}$$

where $\mathbf{x}$, $\mathbf{p}$, $\boldsymbol{\theta}$, and $\mathbf{w}$ denote the vectors of the offloading decision x_k, transmit power p_k, the normalized portion θ_k of bandwidth, and the normalized portion w_k of computation resource, respectively. α_k is the weight for UE k, where the larger the value of α_k, the higher the priority for energy saving of user k. In problem $\mathcal{P}_0$, (17a) gives the delay constraint for each UE. Equation (17b) implies that each task can be either executed locally or offloaded to the MEC server. Equation (17c) describes the transmit power constraint of UE k. Equations (17d) and (17e) are the bandwidth constraints. Equations (17f) and (17g) describe the computation resource constraints. It should be pointed out that the objective function is a function of $\mathbf{x}$, $\mathbf{p}$, and $\boldsymbol{\theta}$, which depend on the random state given by channel estimation error e_k.

Problem $\mathcal{P}_0$ is challenging to solve due to the following three reasons.

1) The objective function cannot be expressed in closed form.
2) There exists coupling between the optimization variables $\mathbf{x}$, $\mathbf{p}$, and $\boldsymbol{\theta}$, leading to the nonconvex objective function and constraint (17a).
3) The binary variable $\mathbf{x}$ makes problem $\mathcal{P}_0$ a mixed-integer nonconvex problem.

Therefore, problem $\mathcal{P}_0$ is a *stochastic mixed-integer nonconvex* optimization problem that is generally difficult to solve. In the next section, we will propose a low-complexity algorithm to obtain a suboptimal solution for problem $\mathcal{P}_0$.

III. Low-Complexity Framework for Problem $\mathcal{P}_0$

In this section, we first develop a CSSCA framework to solve the relaxed version of problem $\mathcal{P}_0$. Specifically, in the CSSCA framework, the problem at each iteration is solved by using the dual decomposition method [38]. Then, we propose

$$\Theta_k(x_k, p_k, \theta_k, w_k) = (1 - x_k)\left(\frac{L_k C_k}{f_k^{\text{loc}}} - T_k^{\max}\right) + x_k\left(\frac{1}{\sigma_{e_k}^2}\left(\frac{\theta_k B N_0}{p_k} - \frac{|\widehat{h}_k|^2}{2^{\frac{L_k}{\theta_k B\left(T_k^{\max} - \frac{L_k C_k}{w_k F_s^{\max}}\right)}} - 1}\right) - \ln \xi_k^{\max}\right) \tag{16}$$

a ranking-based algorithm to obtain a suboptimal solution for the original problem.

A. Problem Transformation

To handle the binary constraint (17b), we relax the binary x_k as a continuous variable, i.e., $x_k \in [0, 1]$. As a result, problem $\mathcal{P}_0$ can be reformulated as

$$\mathcal{P}_1\text{: } \min_{\mathbf{z}} \sum_{k=1}^{K} \Phi_k(\mathbf{z}_k) \tag{18}$$

$$\text{s.t. } \Theta_k(\mathbf{z}_k) \le 0 \quad \forall k \tag{18a}$$

$$0 \le x_k \le 1 \quad \forall k$$

$$(17c)-(17g) \tag{18b}$$

where $\mathbf{z}_k = [x_k, p_k, \theta_k, w_k]^T \in \mathbb{R}^{4\times 1}$ and $\mathbf{z} = [\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_K^T]^T \in \mathbb{R}^{4K\times 1}$. The functions $\Phi_k(\mathbf{z}_k)$ and $\varphi_k(\mathbf{z}_k, e_k)$ are defined as

$$\Phi_k(\mathbf{z}_k) = \mathbb{E}[\varphi_k(\mathbf{z}_k, e_k)] \tag{19}$$

$$\varphi_k(\mathbf{z}_k, e_k) = \alpha_k E_k(x_k, p_k, \theta_k, e_k). \tag{20}$$

It is noteworthy that problem $\mathcal{P}_1$ is still a nonconvex constrained stochastic optimization problem. There exist a few algorithms that can deal with nonconvex stochastic optimization problems, i.e., sample average approximation (SAA) [39]–[41] and the online primal–dual algorithm [42]. However, they either have a higher computational complexity or cannot guarantee the convergence. So it is necessary to develop a new effective method to handle it. In what follows, we will develop a CSSCA-based algorithm to solve problem $\mathcal{P}_1$, which is more efficient and provably convergent [43], [44]. The key idea of CSSCA is to convert the complicated nonconvex stochastic optimization problem into an iterative optimization problem. At each iteration, variable $\mathbf{z}$ is updated by solving a simple convex problem that is obtained by replacing the nonconvex functions $\Phi_k(\mathbf{z}_k)$ and $\Theta_k(\mathbf{z}_k)$ with their corresponding convex surrogate functions.

Let $\varphi_k^c(\mathbf{z}_k)$ and $\varphi_k^{\bar{c}}(\mathbf{z}_k, e_k)$ denote convex and nonconvex parts of $\varphi_k(\mathbf{z}_k, e_k)$, respectively. Then, $\varphi_k(\mathbf{z}_k, e_k)$ can be expressed as

$$\varphi_k(\mathbf{z}_k, e_k) = \varphi_k^c(\mathbf{z}_k) + \varphi_k^{\bar{c}}(\mathbf{z}_k, e_k) \tag{21}$$

where

$$\varphi_k^c(\mathbf{z}_k) \triangleq \alpha_k(1 - x_k)\beta_k\left(f_k^{\text{loc}}\right)^2 L_k C_k \tag{22}$$

and

$$\varphi_k^{\bar{c}}(\mathbf{z}_k, e_k) \triangleq \alpha_k x_k \frac{p_k L_k}{\theta_k B \log_2\left(1 + \frac{p_k|\widehat{h}_k|^2}{p_k|e_k|^2 + \theta_k B N_0}\right)}. \tag{23}$$

Let $\widehat{\Phi}_k^{(i)}(\mathbf{z}_k)$ denote the convex surrogate function of $\Phi_k(\mathbf{z}_k)$ at the ith iteration of problem $\mathcal{P}_1$. The same as in [43] and [45], $\widehat{\Phi}_k^{(i)}(\mathbf{z}_k)$ is expressed as

$$\begin{aligned}\widehat{\Phi}_k^{(i)}(\mathbf{z}_k) = {} & \left(1 - \rho^{(i)}\right)\Phi_k^{(i-1)} + \rho^{(i)}\varphi_k^c(\mathbf{z}_k) + \rho^{(i)}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) \\ & + \rho^{(i)}\nabla^T\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right)\left(\mathbf{z}_k - \mathbf{z}_k^{(i)}\right) + \left(1 - \rho^{(i)}\right) \\ & \times \left(\mathbf{u}_k^{(i-1)}\right)^T\left(\mathbf{z}_k - \mathbf{z}_k^{(i)}\right) + \tau_k^0\left\|\mathbf{z}_k - \mathbf{z}_k^{(i)}\right\|^2\end{aligned} \tag{24}$$

where τ_k^0 is any positive number, and $\Phi_k^{(i)}$ and $\mathbf{u}_k^{(i)}$ are *approximations* for $\mathbb{E}[\varphi_k(\mathbf{z}_k^{(i)}, e_k)]$ and $\nabla\mathbb{E}[\varphi_k(\mathbf{z}_k^{(i)}, e_k)]$, respectively, which are updated recursively according to

$$\Phi_k^{(i)} = \left(1 - \rho^{(i)}\right)\Phi_k^{(i-1)} + \rho^{(i)}\varphi_k\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right), \tag{25}$$

$$\mathbf{u}_k^{(i)} = \left(1 - \rho^{(i)}\right)\mathbf{u}_k^{(i-1)} + \rho^{(i)}\nabla\varphi_k\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) \tag{26}$$

with the initial values $\Phi_k^{(-1)} = 0$, $\mathbf{u}_k^{(-1)} = \mathbf{0}$, and $\rho^{(i)} \in (0, 1]$ is a given constant sequence.

In (26), $\nabla\varphi_k(\mathbf{z}_k^{(i)}, e_k^{(i)})$ is expressed as

$$\nabla\varphi_k\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) = \nabla\varphi_k^c\left(\mathbf{z}_k^{(i)}\right) + \nabla\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) \tag{27}$$

where $\nabla\varphi_k^c(\mathbf{z}_k^{(i)})$ and $\nabla\varphi_k^{\bar{c}}(\mathbf{z}_k^{(i)}, e_k^{(i)})$ are, respectively, given by (28a) and (28b), shown at the bottom of the page. The

$$\nabla\varphi_k^c\left(\mathbf{z}_k^{(i)}\right) = \begin{bmatrix} -\alpha_k L_k \beta_k\left(f_k^{\text{loc}}\right)^2 C_k & 0 & 0 & 0 \end{bmatrix}^T \tag{28a}$$

$$\nabla\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) = \begin{bmatrix} \frac{\partial}{\partial x_k}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) & \frac{\partial}{\partial p_k}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) & \frac{\partial}{\partial \theta_k}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) & 0 \end{bmatrix}^T \tag{28b}$$

$$\begin{aligned}\frac{\partial}{\partial x_k}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) &= \frac{p_k^{(i)}\alpha_k L_k \ln 2}{\theta_k^{(i)} m_k^{(i)} B} \\ \frac{\partial}{\partial p_k}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) &= \frac{x_k^{(i)}\alpha_k L_k \ln 2}{\theta_k^{(i)}\left(m_k^{(i)}\right)^2 B}\left(m_k^{(i)} - \frac{p_k^{(i)}\theta_k^{(i)}|\widehat{h}_k|^2 B N_0}{\left(p_k^{(i)}\left|e_k^{(i)}\right|^2 + \theta_k^{(i)} B N_0\right)\left(p_k^{(i)}|\widehat{h}_k|^2 + p_k^{(i)}\left|e_k^{(i)}\right|^2 + \theta_k^{(i)} B N_0\right)}\right) \\ \frac{\partial}{\partial \theta_k}\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) &= -\frac{x_k^{(i)} p_k^{(i)}\alpha_k L_k \ln 2}{\left(\theta_k^{(i)} m_k^{(i)}\right)^2 B}\left(m_k^{(i)} - \frac{p_k^{(i)}\theta_k^{(i)}|\widehat{h}_k|^2 B N_0}{\left(p_k^{(i)}\left|e_k^{(i)}\right|^2 + \theta_k^{(i)} B N_0\right)\left(p_k^{(i)}\left|e_k^{(i)}\right|^2 + \theta_k^{(i)} B N_0 + p_k^{(i)}|\widehat{h}_k|^2\right)}\right)\end{aligned} \tag{29}$$

first three components in (28b) are given by (29), shown at the bottom of the previous page, with

$$m_k^{(i)} = \ln\left(1 + \frac{p_k^{(i)}\left|\widehat{h}_k\right|^2}{p_k^{(i)}\left|e_k^{(i)}\right|^2 + \theta_k^{(i)}BN_0}\right).$$

Similarly, $\Theta_k(\mathbf{z}_k)$ is also divided into two parts as

$$\Theta_k(\mathbf{z}_k) = \Theta_k^c(\mathbf{z}_k) + \Theta_k^{\overline{c}}(\mathbf{z}_k) \tag{30}$$

where

$$\Theta_k^c(\mathbf{z}_k) \triangleq (1 - x_k)\left(\frac{L_k C_k}{f_k^{\text{loc}}} - T_k^{\max}\right) - x_k \ln \xi_k^{\max} \tag{31}$$

and

$$\Theta_k^{\overline{c}}(\mathbf{z}_k) \triangleq \frac{x_k}{\sigma_{e_k}^2}\left(\frac{\theta_k BN_0}{p_k} - \frac{\left|\widehat{h}_k\right|^2}{2^{\frac{L_k}{\theta_k B\left(T_k^{\max} - \frac{L_k C_k}{w_k F_s^{\max}}\right)}} - 1}\right) \tag{32}$$

are its convex and nonconvex parts, respectively.

Let $\widehat{\Theta}_k^{(i)}(\mathbf{z}_k)$ denote the convex surrogate function of $\Theta_k(\mathbf{z}_k)$ at the ith iteration of problem $\mathcal{P}_1$. Similar to $\widehat{\Phi}_k^{(i)}(\mathbf{z}_k)$, $\widehat{\Theta}_k^{(i)}(\mathbf{z}_k)$ can be given as

$$\begin{aligned}\widehat{\Theta}_k^{(i)}(\mathbf{z}_k) = {} & \left(1 - \rho^{(i)}\right)\Theta_k^{(i-1)} + \rho^{(i)}\Theta_k^c(\mathbf{z}_k) \\ & + \rho^{(i)}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) + \rho^{(i)}\nabla^T\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right)\left(\mathbf{z}_k - \mathbf{z}_k^{(i)}\right) \\ & + \left(1 - \rho^{(i)}\right)\left(\mathbf{v}_k^{(i-1)}\right)^T\left(\mathbf{z}_k - \mathbf{z}_k^{(i)}\right) + \tau_k^1\left\|\mathbf{z}_k - \mathbf{z}_k^{(i)}\right\|^2\end{aligned} \tag{33}$$

where τ_k^1 is any positive number, $\Theta_k^{(i)} = \Theta_k(\mathbf{z}_k^{(i)})$, and $\mathbf{v}_k^{(i)} = \nabla\Theta_k(\mathbf{z}_k^{(i)})$. Similar to the procedure in (27), $\mathbf{v}_k^{(i)}$ can be expressed as $\mathbf{v}_k^{(i)} = \nabla\Theta_k^c(\mathbf{z}_k^{(i)}) + \nabla\Theta_k^{\overline{c}}(\mathbf{z}_k^{(i)})$, where $\nabla\Theta_k^c(\mathbf{z}_k^{(i)})$ and $\nabla\Theta_k^{\overline{c}}(\mathbf{z}_k^{(i)})$ are, respectively, given by (34a) and (34b), shown at the bottom of the page. The components in (34b) are given by (35), shown at the bottom of the page, with

$$n_k^{(i)} = 2^{\frac{L_k}{\theta_k^{(i)}B\left(T_k^{\max} - \frac{L_k C_k}{w_k^{(i)} F_s^{\max}}\right)}}.$$

As a result, the optimization problem of the ith iteration can be expressed as

$$\mathcal{P}_2^{(i)}\colon \min_{\mathbf{z}} \sum_{k=1}^{K} \widehat{\Phi}_k^{(i)}(\mathbf{z}_k) \tag{36}$$

$$\text{s.t. } \widehat{\Theta}_k(\mathbf{z}_k) \le 0 \quad \forall k$$

$$\text{(17c)}-\text{(17g)},\ \text{(18b)}. \tag{36a}$$

Note that $\mathcal{P}_2^{(i)}$ is not necessarily feasible. Let $\widehat{\mathbf{z}}^{(i)}$ denote the optimal solution to $\mathcal{P}_2^{(i)}$. If $\mathcal{P}_2^{(i)}$ is infeasible, the following problem will be solved to minimize the constraints:

$$\mathcal{P}_3^{(i)}\colon \min_{\mathbf{z},\delta} \ \delta \tag{37}$$

$$\text{s.t. } \widehat{\Theta}_k(\mathbf{z}_k) \le \delta \quad \forall k \tag{37a}$$

$$-\delta \le x_k \le 1 + \delta \quad \forall k \tag{37b}$$

$$-\delta \le p_k \le P_k^{\max} + \delta \quad \forall k \tag{37c}$$

$$-\delta \le \theta_k \le 1 + \delta \quad \forall k \tag{37d}$$

$$\sum_{k=1}^{K} \theta_k \le 1 + \delta \tag{37e}$$

$$-\delta \le w_k \le 1 + \delta \quad \forall k \tag{37f}$$

$$\sum_{k=1}^{K} w_k \le 1 + \delta. \tag{37g}$$

Obviously, both problems $\mathcal{P}_2^{(i)}$ and $\mathcal{P}_3^{(i)}$ are convex since the surrogate functions in (24) and (33) are convex. Thus, $\mathcal{P}_2^{(i)}$ and $\mathcal{P}_3^{(i)}$ can be efficiently solved by interior-point methods [46].

$$\nabla\Theta_k^c\left(\mathbf{z}_k^{(i)}\right) = \left[T_k^{\max} - \tfrac{L_k C_k}{f_k^{\text{loc}}} - \ln\xi_k^{\max} \quad 0 \quad 0 \quad 0\right]^T \tag{34a}$$

$$\nabla\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) = \left[\tfrac{\partial}{\partial x_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) \quad \tfrac{\partial}{\partial p_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) \quad \tfrac{\partial}{\partial \theta_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) \quad \tfrac{\partial}{\partial w_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right)\right]^T \tag{34b}$$

$$\begin{aligned}\frac{\partial}{\partial x_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) &= \frac{1}{\sigma_{e_k}^2}\left(\frac{\theta_k BN_0}{p_k} - \frac{\left|\widehat{h}_k\right|^2}{n_k - 1}\right) \\ \frac{\partial}{\partial p_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) &= -\frac{x_k^{(i)}\theta_k^{(i)}BN_0}{\sigma_{e_k}^2\left(p_k^{(i)}\right)^2} \\ \frac{\partial}{\partial \theta_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) &= \frac{x_k^{(i)}}{\sigma_{e_k}^2}\left(\frac{BN_0}{p_k^{(i)}} - \frac{\left|\widehat{h}_k\right|^2 L_k \ln 2}{B\left(\theta_k^{(i)}\right)^2} \times \frac{n_k^{(i)}}{\left(n_k^{(i)} - 1\right)^2\left(T_k^{\max} - \frac{L_k C_k}{w_k^{(i)} F_s^{\max}}\right)}\right) \\ \frac{\partial}{\partial w_k}\Theta_k^{\overline{c}}\left(\mathbf{z}_k^{(i)}\right) &= -\frac{x_k^{(i)} L_k^2 C_k \left|\widehat{h}_k\right|^2 \ln 2}{BF_s^{\max}\sigma_{e_k}^2\theta_k^{(i)}\left(w_k^{(i)}\right)^2} \times \frac{n_k^{(i)}}{\left(n_k^{(i)} - 1\right)^2\left(T_k^{\max} - \frac{L_k C_k}{w_k^{(i)} F_s^{\max}}\right)^2}\end{aligned} \tag{35}$$

Algorithm 1 RODRA for Solving $\mathcal{P}_1$

Initialization:
Choose proper sequences $\{\gamma^{(i)}\}$ and $\{\rho^{(i)}\}$;
Choose an initial point $\mathbf{z}^{(0)}$ meeting all constraints in $\mathcal{P}_1$;
Set iteration index $i = 0$.
repeat
  Generate a channel error vector $\mathbf{e}^{(i)} = [e_1^{(i)}, ..., e_K^{(i)}]^T$.
  Calculate the surrogate functions $\widehat{\Phi}_k^{(i)}(\mathbf{z}_k)$, $\widehat{\Theta}_k^{(i)}(\mathbf{z}_k)$, $\forall k$ using (24) and (33).
  if $\mathcal{P}_2^{(i)}$ is feasible **then**
    Solve problem $\mathcal{P}_2^{(i)}$ to obtain $\widehat{\mathbf{z}}^{(i)}$.
  else
    Solve problem $\mathcal{P}_3^{(i)}$ to obtain $\widehat{\mathbf{z}}^{(i)}$.
  end if
  Update $\mathbf{z}^{(i+1)}$ according to (38).
  Set $i = i + 1$.
until the stopping criterion is met.
Output: The solution $\mathbf{z}^{\text{Rel}} = \mathbf{z}^{(i+1)}$ for problem $\mathcal{P}_1$.

After obtaining the optimal solution $\widehat{\mathbf{z}}^{(i)}$ for problems $\mathcal{P}_2^{(i)}$ or $\mathcal{P}_3^{(i)}$, $\mathbf{z}$ can be updated as follows:

$$\mathbf{z}^{(i+1)} = \left(1 - \gamma^{(i)}\right)\mathbf{z}^{(i)} + \gamma^{(i)}\widehat{\mathbf{z}}^{(i)} \tag{38}$$

where $\gamma^{(i)} \in (0, 1]$ is a given constant sequence.

The procedure to solve $\mathcal{P}_1$ is described in Algorithm 1. However, it is not necessary to guarantee that the optimal value of the offloading decision is in the integral form since we have relaxed the binary constraint. Thus, Algorithm 1 can be referred to as a relaxed offloading decision and resource allocation algorithm (RODRA).

Remark 1: Algorithm 1 is a centralized algorithm and executed at the BS, which has strong computing power and communication capability.

B. Low-Complexity Solution for Problem $\mathcal{P}_2^{(i)}$

Although problems $\mathcal{P}_2^{(i)}$ and $\mathcal{P}_3^{(i)}$ are convex and can be solved by adopting general interior-point methods [46], this method suffers from relatively high complexity since it does not utilize their special structures. Hence, we develop a low-complexity algorithm based on the dual decomposition theory [38] to obtain the optimal solution for problem $\mathcal{P}_2^{(i)}$ (or $\mathcal{P}_3^{(i)}$) with closed-form update at each iteration.

Due to the similar structures of problems $\mathcal{P}_2^{(i)}$ and $\mathcal{P}_3^{(i)}$, we only focus on solving $\mathcal{P}_2^{(i)}$ in the remainder of this section for brevity. In what follows, we illustrate how to exploit the Lagrange dual decomposition method to solve $\mathcal{P}_2^{(i)}$. The partial Lagrangian of $\mathcal{P}_2^{(i)}$ is written as

$$\mathcal{L}^{(i)}(\mathbf{z}, \boldsymbol{\mu}, \eta, \lambda) = \sum_{k=1}^{K} \widehat{\Phi}_k^{(i)}(\mathbf{z}_k) + \sum_{k=1}^{K} \mu_k \widehat{\Theta}_k^{(i)}(\mathbf{z}_k) + \eta\left(\sum_{k=1}^{K} \theta_k - 1\right) + \lambda\left(\sum_{k=1}^{K} w_k - 1\right) \tag{39}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]$, η, and λ are Lagrange multipliers associated to constraints (36a), (17e), and (17g), respectively.

The dual function is then written as

$$\mathcal{D}^{(i)}(\boldsymbol{\mu}, \eta, \lambda) = \min_{\mathbf{z}} \ \mathcal{L}^{(i)}(\mathbf{z}, \boldsymbol{\mu}, \eta, \lambda) \quad \text{s.t. (18b), (17c), (17d), and (17f).} \tag{40}$$

Therefore, the dual problem is expressed by

$$\max_{\boldsymbol{\mu}, \eta, \lambda} \ \mathcal{D}^{(i)}(\boldsymbol{\mu}, \eta, \lambda) \quad \text{s.t.} \ \ \boldsymbol{\mu} \succeq 0, \quad \eta \geq 0, \quad \lambda \geq 0. \tag{41}$$

Since problem $\mathcal{P}_2^{(i)}$ is convex and Slater's condition is satisfied, the strong duality holds between it and its dual problem. Thus, the optimal solution to problem $\mathcal{P}_2^{(i)}$ can be obtained by solving problem (41).

After some algebraic manipulations, (39) can be rewritten by

$$\mathcal{L}^{(i)}(\mathbf{z}, \boldsymbol{\mu}, \eta, \lambda) = \sum_{k=1}^{K} a_k \|\mathbf{z}_k\|^2 + (\mathbf{b}_k)^T \mathbf{z}_k + c_k \tag{42}$$

where a_k, $\mathbf{b}_k$, and c_k are given by (43), shown at the bottom of the page.

Clearly, for a given $(\boldsymbol{\mu}, \eta, \lambda)$, the dual function $\mathcal{D}^{(i)}(\boldsymbol{\mu}, \eta, \lambda)$ in (40) can be decomposed into K independent subproblems as

$$\min_{\mathbf{z}_k} \ a_k \|\mathbf{z}_k\|^2 + (\mathbf{b}_k)^T \mathbf{z}_k + c_k \quad \text{s.t. (18b), (17c), (17d), and (17f)} \tag{44}$$

for $k = 1, \ldots, K$.

$$\begin{aligned}
a_k &= \tau_k^0 + \mu_k \tau_k^1 \\
\mathbf{b}_k &= \rho^{(i)}\left(\nabla \varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) + \mu_k \nabla \Theta_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}\right)\right) + \left(1 - \rho^{(i)}\right)\left(\mathbf{u}_k^{(i-1)} + \mu_k \mathbf{v}_k^{(i-1)}\right) - 2\left(\tau_k^0 + \mu_k \tau_k^1\right)\mathbf{z}_k^{(i)} \\
&\quad - \rho^{(i)}\left(\ln \xi_k^{\max} + \alpha_k \beta_k \left(f_k^{\text{loc}}\right)^2 L_k C_k + \mu_k\left(\frac{L_k C_k}{f_k^{\text{loc}}} - T_k^{\max}\right)\right)[1, 0, 0, 0]^T + [0, 0, \eta, \lambda]^T \\
c_k &= -\frac{1}{K}(\eta + \lambda) + \left(1 - \rho^{(i)}\right)\left(\Phi_k^{(i-1)} + \mu_k \Theta_k^{(i-1)}\right) + \rho^{(i)}\left(\varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) + \mu_k \Theta_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}\right)\right) \\
&\quad - \left(1 - \rho^{(i)}\right)\left(\mathbf{u}_k^{(i-1)} + \mu_k \mathbf{v}_k^{(i-1)}\right)^T \mathbf{z}_k^{(i)} - \rho^{(i)}\left(\nabla \varphi_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}, e_k^{(i)}\right) + \mu_k \nabla \Theta_k^{\bar{c}}\left(\mathbf{z}_k^{(i)}\right)\right)^T \mathbf{z}_k^{(i)} \\
&\quad + \left(\tau_k^0 + \mu_k \tau_k^1\right)\left\|\mathbf{z}_k^{(i)}\right\|^2 + \rho^{(i)}\left(\alpha_k \beta_k \left(f_k^{\text{loc}}\right)^2 L_k C_k + \mu_k\left(\frac{L_k C_k}{f_k^{\text{loc}}} - T_k^{\max}\right)\right)
\end{aligned} \tag{43}$$

For the optimal solution of (44), we have the following lemma.

Lemma 2: For given $(\boldsymbol{\mu}, \eta, \lambda)$, the optimal solution $\mathbf{z}_k^* = [x_k^*, p_k^*, \theta_k^*, w_k^*]^T$ for subproblem (44) is

$$x_k^* = [\bar{x}_k]_0^1 \tag{45}$$

$$p_k^* = \left[\bar{p}_k\right]_0^{P_k^{\max}} \tag{46}$$

$$\theta_k^* = \left[\bar{\theta}_k\right]_0^1 \tag{47}$$

$$w_k^* = [\bar{w}_k]_0^1 \tag{48}$$

where $\bar{x}_k = -(b_{k,1}/2a_k)$, $\bar{p}_k = -(b_{k,2}/2a_k)$, $\bar{\theta}_k = -(b_{k,3}/2a_k)$, and $\bar{w}_k = -(b_{k,4}/2a_k)$.

Proof: See Appendix B. ∎

On the other hand, the dual variables $(\boldsymbol{\mu}, \eta, \lambda)$ can be updated by adopting the subgradient method, which are given as follows:

$$\mu_k(l+1) = \left[\mu_k(l) + \delta_{\mu,l}\widehat{\Theta}_k^{(i)}\left(\mathbf{z}_k^*\right)\right]^+ \tag{49}$$

$$\eta(l+1) = \left[\eta(l) + \delta_{\eta,l}\left(\sum_{k=1}^{K}\theta_k^* - 1\right)\right]^+ \tag{50}$$

$$\lambda(l+1) = \left[\lambda(l) + \delta_{\lambda,l}\left(\sum_{k=1}^{K}w_k^* - 1\right)\right]^+ \tag{51}$$

where l is the iteration index of the subgradient method; $\mathbf{z}_k^* = [x_k^*, p_k^*, \theta_k^*, w_k^*]^T$ is the optimal solution to (44) with $(\boldsymbol{\mu}(l), \eta(l), \lambda(l))$; $\delta_{\mu,l}$, $\delta_{\eta,l}$, and $\delta_{\lambda,l}$ are positive step sizes; and $[\,\cdot\,]^+ = \max\{0, \cdot\}$.

Since problem (40) can be decomposed as multiple independent subproblems, it is convenient to implement at the MEC server in a parallel way.

C. Ranking-Based Recovery of Binary Offloading Decision Variables

So far, we can solve problem $\mathcal{P}_1$ successfully and efficiently. However, the solution $\mathbf{x}^{\text{Rel}}$ obtained from solving problem $\mathcal{P}_1$ is not necessarily a vector with integer elements. If $\mathbf{x}^{\text{Rel}}$ is a vector with binary elements, $\mathbf{z} = \mathbf{z}^{\text{Rel}}$. If $\mathbf{x}^{\text{Rel}}$ is not a vector with integer elements, which is infeasible for problem $\mathcal{P}_0$. Thus, we need to recover the binary $\mathbf{x}$. In this section, we propose a ranking-based algorithm to obtain a suboptimal solution of $\mathcal{P}_0$, where $\mathbf{x}$ is a binary vector. This algorithm is composed of two stages summarized in Algorithm 2.

In the first stage, the MEC server computes the computation offloading priorities of all UEs, denoted by $\mathbf{x}^{\text{Rel}}$. The computation offloading priority of UE k can be defined as x_k^{Rel} obtained by Algorithm 1. Furthermore, we sort the elements of $\mathbf{x}^{\text{Rel}}$ in the decreasing order. Let $\mathcal{S}^{\text{off}} = \{k \in \mathcal{K} | x_k = 1\}$ denote the set of offloading UEs.

In the second stage, UE is added to $\mathcal{S}^{\text{off}}$ in a descending order of offloading priorities until the weighted sum of EC cannot be further decreased. For example, if $U(\mathbf{z}|\mathcal{S}^{\text{off}} \cup \{k_j\}) \geq U(\mathbf{z}|\mathcal{S}^{\text{off}})$, $\mathcal{S}^{\text{off}}$ is the final set of offloading UEs. Here, $U(\mathbf{z}|\mathcal{S}^{\text{off}})$ denotes the optimal value of $\mathcal{P}_1$ under given offloading variable $\mathbf{x}$ and can be obtained by solving the following problem:

$$\mathcal{P}_4\text{: } \min_{\mathbf{z}} \sum_{k=1}^{K}\Phi_k(\mathbf{z}_k) \tag{52}$$

$$\text{s.t. } x_k = 1 \quad \forall k \in \mathcal{S}^{\text{off}}$$

$$(17a),\ (17c)-(17g). \tag{52a}$$

Note that $\mathcal{P}_4$ is a special form of $\mathcal{P}_1$, and thus can be solved by Algorithm 1 by setting $x_k = 1\ \forall k \in \mathcal{S}^{\text{off}}$.

The whole procedure to solve $\mathcal{P}_0$ is summarized in Algorithm 2, which is nested by Algorithm 1.

Algorithm 2 RBORA for Solving $\mathcal{P}_0$

Stage 1:

1: Obtain the solution $\mathbf{z}^{\text{Rel}}$ for problem $\mathcal{P}_1$ by using Algorithm 1.
2: Order $\mathbf{x}^{\text{Rel}}$ from largest to smallest, i.e., $x_{k_1}^{\text{Rel}} \geq x_{k_2}^{\text{Rel}} \geq \ldots \geq x_{k_K}^{\text{Rel}}$.
3: Set $\mathcal{S}^{\text{off}} = \emptyset$.

Stage 2:

4: **for** $j = 1, ..., K$ **do**
5: Set $\mathcal{S}^{\text{temp}} = \mathcal{S}^{\text{off}} \cup \{k_j\}$ and calculate $U(\mathbf{z}|\mathcal{S}^{\text{temp}})$ by using Algorithm 1.
6: **if** $U(\mathbf{z}|\mathcal{S}^{\text{temp}}) \leq U(\mathbf{z}|\mathcal{S}^{\text{off}})$ **then**
7: $\mathcal{S}^{\text{off}} = \mathcal{S}^{\text{temp}}$.
8: **else**
9: Break.
10: **end if**
11: **end for**
12: **Output:** $\mathcal{S}^{\text{off}}$ and corresponding $\mathbf{p}$, $\boldsymbol{\theta}$, and $\mathbf{w}$.

IV. Convergence and Complexity Analysis

A. Convergence Analysis of Algorithm 1

The whole procedure to solve $\mathcal{P}_0$ is summarized in Algorithm 2, which includes two stages. In the first stage, $\mathcal{P}_1$ is solved by using Algorithm 1. In the second stage, the ranking-based method is adopted to search $\mathcal{S}^{\text{off}}$ and compute corresponding resource allocation variables. Specifically, each iteration of the second stage needs to solve $\mathcal{P}_4$ (namely, $\mathcal{P}_1$ under given $\mathbf{x}$ according to offloading priorities obtained from the first stage) by using Algorithm 1. Thus, we only need to check the convergence of Algorithm 1.

Note that Algorithm 1 includes a nested loop. In the inner loop (lines 5–10), convex problem $\mathcal{P}_2^{(i)}$ or $\mathcal{P}_3^{(i)}$ is solved by using the Lagrange dual decomposition method. For the outer loop of Algorithm 1, with the properly chosen sequences $\{\gamma^{(i)}\}$ and $\{\rho^{(i)}\}$, it *almost surely* converges to a stationary point of problem $\mathcal{P}_1$. In other words, it almost surely converges to a local optimal point. For a rigorous proof of this convergence, interested readers are referred to [43].

B. Complexity Analysis

As mentioned above, Algorithm 2 for solving $\mathcal{P}_0$ is composed of two stages. For the first stage, it needs to run Algorithm 1 once. For the second stage, it needs to run

TABLE I
SIMULATION PARAMETERS

Notation	Parameter	Value
r	Cell radius	500 m
B	Total bandwidth	10 MHz
K	Number of UEs	$\{6-20\}$
$P_k^{\max}$	Maximum transmit power	23 dBm
N_0	Noise power density	-174 dBm/Hz
$\sigma_{e_k}^2$	Variance of estimation error	0.002
L_k	Size of task input data	0.42 MB
C_k	Density of task computation	297.6 cycles/bit
$T_k^{\max}$	Maximum tolerable delay	$\{0.9-2.5\}$ s
$F^{\max}$	Maximum computation resource of MEC	$\{5-20\}$ GHz
f_k^{loc}	Local computation resource	1.2 GHz
β_k	Energy consumption coefficient	1×10^{-28}
$\xi_k^{\max}$	Maximum delay outage probability	10%

Algorithm 1 at most K times. Note that Algorithm 1 includes a nested loop. At its each iteration, a convex problem $\mathcal{P}_2^{(i)}$ or $\mathcal{P}_3^{(i)}$ is solved by using the Lagrange dual decomposition method. Thus, the complexity of Algorithm 1 is $\mathcal{O}(K)$ [47]. Supposing that Algorithm 1 requires $I_{\text{inner}}^{\max}$ iterations to converge, the complexity of Algorithm 1 is at most $\mathcal{O}(I_{\text{inner}}^{\max}K)$. Thus, the total complexity of Algorithm 2 is at most $\mathcal{O}(I_{\text{inner}}^{\max}K^2)$.

As one of the important benchmarks, the complexity of the robust method in [32] is $\mathcal{O}(N_2^{\text{OA}}(2K+N_1^{\text{bis}}K^{3.5}N_3^{\text{feas}}))$, where N_1^{bis}, N_2^{OA}, and N_3^{feas} are the number of bisection search, the number of alternating optimization of outer loop, and the number of iteration of feasibility verification, respectively. Obviously, the complexity of this algorithm is much higher than that of the proposed algorithm.

V. NUMERICAL SIMULATION

In this section, numerical simulations are presented to evaluate the performance of the proposed ranking-based binary offloading and resource allocation algorithm (RBORA). We consider a multiuser MEC system consisting of a BS and K UEs. The BS is in the center of a circle where UEs are randomly distributed. The large-scale fading model is $128.1+37.5\log_{10} r_{[\text{Km}]}$ (dB), and the log-normal shadowing standard deviation is 10 dB. For simplicity, we assume that the estimated channel $\widehat{h}_k \sim \mathcal{CN}(0,1)$ and the estimation error $e_k \sim \mathcal{CN}(0,\sigma_{e_k}^2)$. For comparison, the weight of each UE is $\alpha_k = 1$. Other default simulation parameters are shown in Table I, if not specified.

We compare our proposed RBORA algorithm with the following benchmarks.

1) *Exhaustive Search (ES):* The optimal offloading decision variable is obtained by ES. Given the offloading decision, we optimize other variables by using Algorithm 1. This provides a performance upper bound for practical schemes.
2) *Nonrobust Scheme (NonR):* When the estimated channels are used as the perfect CSI, the offloading decision and resource allocation are jointly optimized to minimize the weighted sum of EC, which is a combination of works [21] and [26]. Note that the nonrobust method is designed without considering the channel errors, this will result in the dissatisfaction of delay constraint (15).
3) *Nonrobust Scheme With Bisection Search (NonR-bisec):* Due to the dissatisfaction of the delay constraint in NonR, the bisection search procedure [33] is applied to the nonrobust method, which will make the delay constraint satisfied. The readers are referred to [33] for details.
4) *Robust Scheme (Robust):* The offloading decision and resource allocation (including transmit power and computation resource) are jointly optimized to minimize the maximum weighted EC under the upper bound of the average latency and resource constraints, as in [32].
5) *Local Execution Only (Local Only):* All tasks are always locally executed without any optimization.

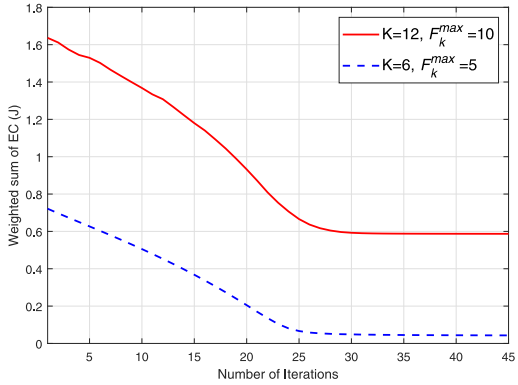


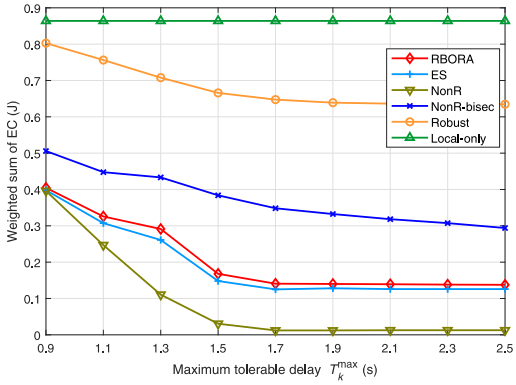
Fig. 2. Convergence of Algorithm 1.

A. Convergence Behavior of Algorithm 1

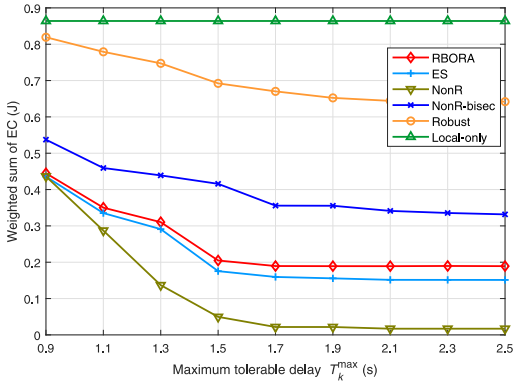
To verify the convergence of the proposed Algorithm 1 (RODRA), Fig. 2 plots the objective function in problem $\mathcal{P}_1$ versus the number of iterations for two cases $K=6$, $F_s^{\max}=5$ GHz, and $T_k^{\max}=1.5$ s and $K=12$, $F_s^{\max}=10$ GHz, and $T_k^{\max}=1.5$ s, respectively. Observe that Algorithm 1 converges to a stationary point at most 30 iterations.

B. Performance Comparison

Fig. 3(a) and (b) shows the weighted sum of EC comparison of the proposed RBORA and benchmarks versus different maximum tolerable delay $T_k^{\max}$ for two cases $\sigma_{e_k}^2=0.002$ and $\sigma_{e_k}^2=0.005$, respectively. From the figures, we can see that the weighted sum of EC of the RBORA is close to that of the ES, especially when $T_k^{\max}$ is smaller. This is because a tighter delay requirement leads to almost the same set of offloading UEs and so the weighted sum of EC is almost the same. But compared to ES, RBORA has a lower complexity especially when K is very large. In addition, compared to NonR, RBORA consumes more energy since it accounts for the channel estimation error and guarantees the delay outage requirement. However, the RBORA scheme outperforms NonR-bisec and the robust in terms of the weighted sum of EC. For the robust, it is because that this method aims to minimize the maximum weighted EC instead of the overall EC.



(a)



(b)

Fig. 3. Weighted sum of EC versus maximum tolerable delay. (a) $\sigma_{e_k}^2 = 0.002$. (b) $\sigma_{e_k}^2 = 0.005$.

Furthermore, comparing Fig. 3(a) and (b), it can be observed that when $\sigma_{e_k}^2$ increases, RBORA takes more energy to meet the delay outage requirement.

To highlight the advantages of the proposed RBORA algorithm, Fig. 4 depicts the cumulative distribution function (CDF) of the realistic task execution delay of all UEs for the case $K = 6$, $F_s^{\max} = 5$ GHz, and $T_k^{\max} = 1.5$ s. We can observe that for all schemes except NonR, all UEs obtain the prescribed task execution delay with a very high probability. In particular, for the Robust, the reason is that it is constrained by the upper bound of the average delay, which leads to that the realistic task execution delay of all UEs is almost smaller than $T_k^{\max}$. However, for NonR, only a few UEs meet the prescribed task execution delay with a very high probability. This implies that the lower the weighted sum of EC of NonR, as shown in Fig. 3, is at the cost of severe violation of the delay constraint. Moreover, the proposed RBORA algorithm has almost the same delay outage performance as ES.

Fig. 5(a) and (b) shows the impact of the maximum computation resource $F_s^{\max}$ of MEC on the weighted sum of EC of the proposed RBORA and benchmarks for the case $K = 6$. It can be seen that the larger $F_s^{\max}$, the less the weighted sum of EC for all schemes except the local only. Moreover, it can be observed that the weighted sum of EC of the RBORA scheme is fairly close to that of ES. In addition, when the delay requirement is looser, i.e., $T_k^{\max} = 2$ s, the weighted sum of EC obtained by all schemes except the local only and

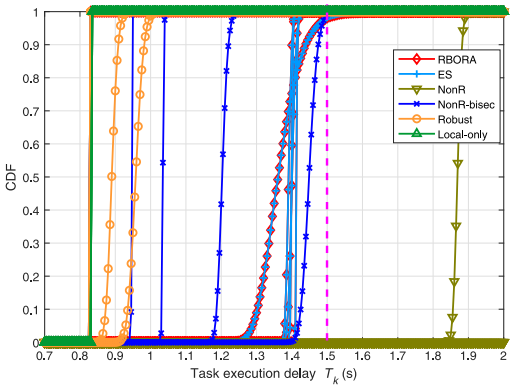
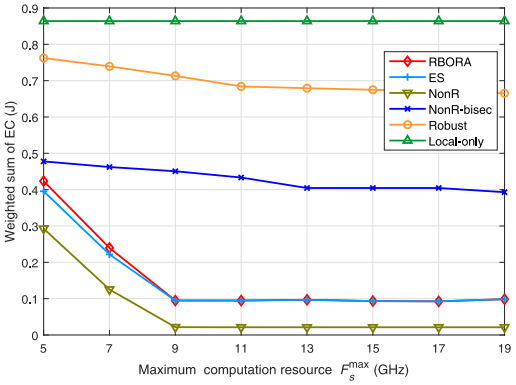
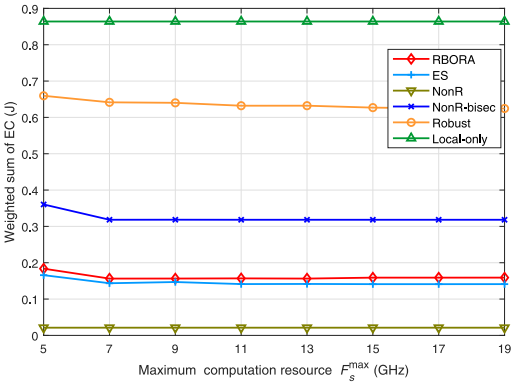


Fig. 4. CDF of realistic task execution delay T_k for $T_k^{\max} = 1.5$ s.



(a)



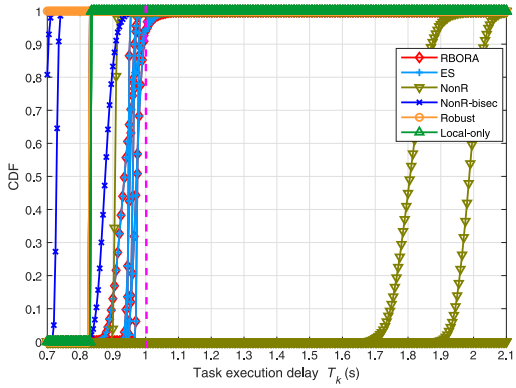
(b)

Fig. 5. Weighted sum of EC versus maximum computation resource of MEC. (a) $T_k^{\max} = 1$ s. (b) $T_k^{\max} = 2$ s.

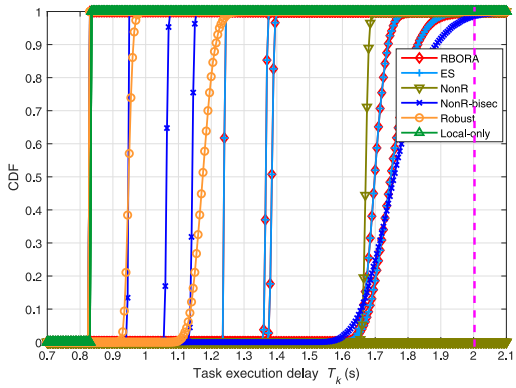
NonR will sooner stabilize. This means that under the looser $T_k^{\max}$, almost all UEs prefer to offload their tasks even when $F_s^{\max}$ is small, i.e., $F_s^{\max} = 7$ GHz.

Similarly, we also plot the CDF of the realistic task execution delay of all UEs in Fig. 6(a) and (b), which, respectively, corresponds to Fig. 5(a) and (b) for the case $F_s^{\max} = 5$ GHz. This phenomenon is similar to that of Fig. 4, that is, all UEs for all schemes except the NonR meet the prescribed task execution delay with a very high probability.

Fig. 7 shows the impact of the task input data size L_k on the weighted sum of EC for the case $K = 6$, $F_s^{\max} = 5$ GHz, and $T_k^{\max} = 1.6$ s. As L_k becomes large, the weighted sum of



(a)



(b)

Fig. 6. CDFs of realistic task execution delay T_k. (a) $T_k^{\max} = 1$ s. (b) $T_k^{\max} = 2$ s.

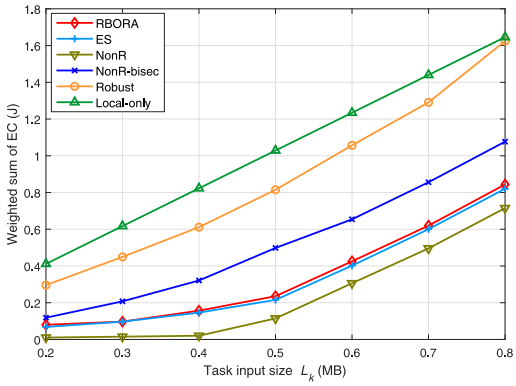


Fig. 7. Weighted sum of EC versus task input size.

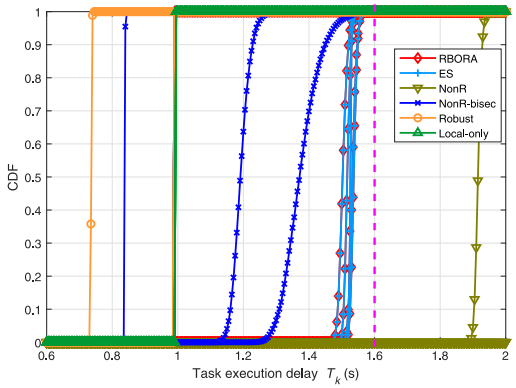
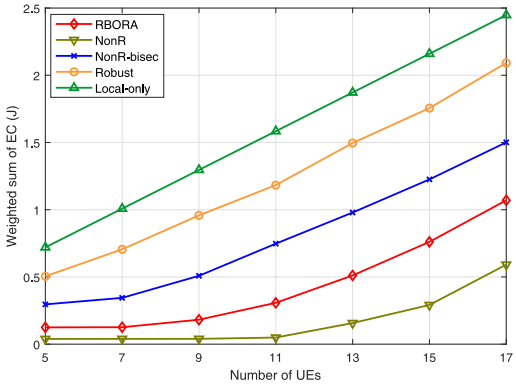
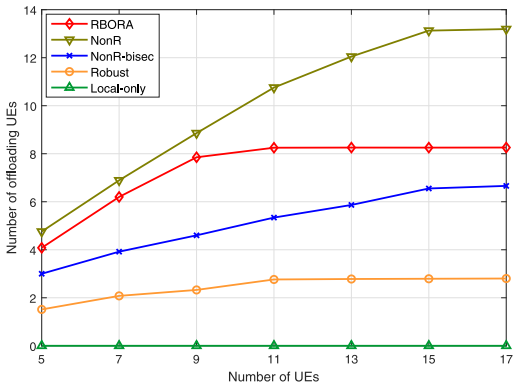


Fig. 8. CDF of realistic task execution delay T_k for $T_k^{\max}$ =1.6 s.



(a)



(b)

Fig. 9. Offloading performance versus the number of UEs. (a) Weighted sum of EC. (b) Number of offloading UEs.

EC of all schemes increases. Besides, we also note that the weighted sum of EC of the proposed RBORA is almost the same as that of ES and higher than that of NonR by at most 0.1 J, but lower than that of NonR-bisec, the robust, and the local only.

Fig. 8 depicts the CDF of the realistic task execution delay of all UEs, which corresponds to Fig. 7 for the case $L = 0.5$MB. We can observe the same result as in Figs. 4 and 6.

Fig. 9(a) and (b) shows the influence of the number of UEs on the offloading performance for the case $F_s^{\max} = 10$ GHz and $T_k^{\max} = 1.6$ s. Since the ES has high computational complexity in MEC systems with a large number of users, we ignore it here. From Fig. 9(a), we can see that the weighted sum of EC increases with the number of UEs for all schemes. Besides, the weighted sum of EC of all schemes except the local only first increases slowly and then increases rapidly. The reason can be explained by Fig. 9(b). As shown in Fig. 9(b), due to the stringent delay requirement, only some UEs prefer to offload their tasks to the MEC server with the limited computation resource, which implies the rest of UEs can only process their tasks locally and thus cause a rapid increase in terms of the weighted sum of EC.

Fig. 10 depicts the CDF of the realistic task execution delay of all UEs for the case $K = 11$, which corresponds to Fig. 9. The phenomenon is consistent with the observations in Figs. 4, 6, and 8.

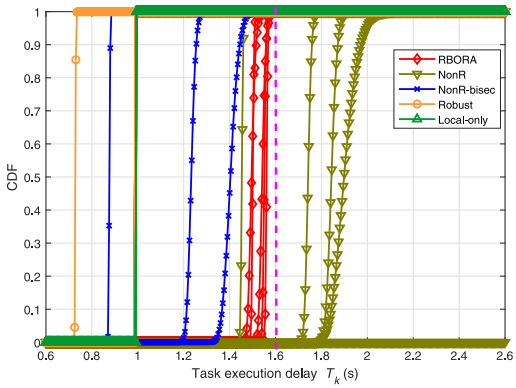


Fig. 10. CDF of realistic task execution delay T_k for $T_k^{\max}$ =1.6 s.

VI. Conclusion

In this article, we have studied the joint offloading decision, transmit power, and computation resource allocation problem to minimize the average weighted sum of EC while meeting the probabilistic delay constraint in the multiuser MEC-enabled IoT network with imperfect CSI. To solve the formulated mixed-integer nonconvex stochastic problem, we developed a novel low-complexity framework. This framework includes two stages. In the first stage, we solved the relaxed version of the original problem to obtain offloading priorities of all UEs. In the second stage, we iteratively solved a stochastic optimization problem to obtain a suboptimal offloading decision and corresponding resource allocation. We have also analyzed the convergence and complexity of the proposed algorithms. Our numerical results have confirmed the advantages of the proposed algorithms over existing methods in terms of the weighted sum of EC and the violation of the delay constraint.

Appendix A
Proof of Lemma 1

When $x_k = 0$, $T_k(x_k, p_k, \theta_k, w_k, e_k) = (L_k C_k)/f_k^{\text{loc}}$. Obviously, it does not contain the random variable e_k and thus the left-hand side of (14) can only be equal to 0 or 1. Therefore, if the inequality (14) holds, the left-hand side of it must be equal to 1, that is

$$\Pr_{e_k}\left\{T_k(x_k, p_k, \theta_k, w_k, e_k) \le T_k^{\max}\right\} = 1. \tag{53}$$

In this case, (14) is reduced as

$$\frac{L_k C_k}{f_k^{\text{loc}}} \le T_k^{\max}. \tag{54}$$

When $x_k = 1$

$$T_k(x_k, p_k, \theta_k, w_k, e_k) = \frac{L_k}{R_k(p_k, \theta_k, e_k)} + \frac{L_k C_k}{w_k F^{\max}}$$

thus, the probability in (14) is calculated as

$$\begin{aligned}
&\Pr_{e_k}\left\{T_k(x_k, p_k, \theta_k, w_k, e_k) \le T_k^{\max}\right\} \\
&= \Pr_{e_k}\left\{|e_k|^2 \le \frac{\left|\widehat{h}_k\right|^2}{2^{\frac{L_k}{B}\frac{1}{T_k^{\max} - \frac{L_k C_k}{w_k F_s^{\max}}}} - 1} - \frac{\theta_k B N_0}{p_k}\right\} \\
&= \Pr_{\widetilde{e}_k}\left\{|\widetilde{e}_k|^2 \le \frac{1}{\sigma_{e_k}^2}\left(\frac{\left|\widehat{h}_k\right|^2}{2^{\frac{L_k}{B}\frac{1}{T_k^{\max} - \frac{L_k C_k}{w_k F_s^{\max}}}} - 1} - \frac{\theta_k B N_0}{p_k}\right)\right\} \\
&= 1 - \exp\left(\frac{1}{\sigma_{e_k}^2}\left(\frac{\theta_k B N_0}{p_k} - \frac{\left|\widehat{h}_k\right|^2}{2^{\frac{L_k}{B}\frac{1}{T_k^{\max} - \frac{L_k C_k}{w_k F_s^{\max}}}} - 1}\right)\right). \quad (55)
\end{aligned}$$

The second equality above is due to $\widetilde{e}_k = (1/\sigma_{e_k})e_k$. The last equality holds since $\widetilde{e}_k \sim \mathcal{CN}(0, 1)$ and thus $|\widetilde{e}_k|^2$ follows an exponential distribution with mean one. So, in this case, (14) becomes

$$\frac{1}{\sigma_{e_k}^2}\left(\frac{\theta_k B N_0}{p_k} - \frac{\left|\widehat{h}_k\right|^2}{2^{\frac{L_k}{\theta_k B\left(T_k^{\max} - \frac{L_k C_k}{w_k F_s^{\max}}\right)}} - 1}\right) \le \ln \xi_k^{\max}. \tag{56}$$

Combining the above two cases, we can obtain the equivalent constraint in (15).

Appendix B
Proof of Lemma 2

Proof: Problem (44) can be further decomposed into three quadratic subproblems

$$\min_{x_k} \; a_k x_k^2 + b_{k,1} x_k \quad \text{s.t. (18b)} \tag{57}$$

$$\min_{p_k} \; a_k p_k^2 + b_{k,2} p_k \quad \text{s.t. (17c)} \tag{58}$$

$$\min_{\theta_k} \; a_k \theta_k^2 + b_{k,3} \theta_k \quad \text{s.t. (17d)} \tag{59}$$

$$\min_{w_k} \; a_k w_k^2 + b_{k,4} w_k \quad \text{s.t. (17f).} \tag{60}$$

Because (57)–(60) are quadratic problems with 1-D variable, the optimal solutions only appear at three points, i.e., two endpoints of a feasible set and the stationary point of the objective function. So we can express their optimal solutions as in (45)–(48), respectively. ■

References

[1] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[2] J. Pan and J. McElhannon, "Future edge cloud and edge computing for Internet of Things applications," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 439–449, Feb. 2018.

[3] J. Wang, D. Feng, S. Zhang, J. Tang, and T. Q. S. Quek, "Computation offloading for mobile edge computing enabled vehicular networks," *IEEE Access*, vol. 7, pp. 62624–62632, 2019.

[4] B. Cao, L. Zhang, Y. Li, D. Feng, and W. Cao, "Intelligent offloading in multi-access edge computing: A state-of-the-art review and framework," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 56–62, Mar. 2019.

[5] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 4, pp. 974–983, Apr. 2015.

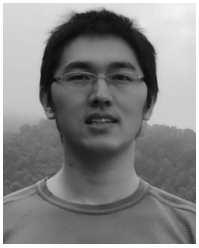
[6] "Mobile-edge computing-introductory technical," ETSI, Sophia Antipolis, France, White Paper, Sep. 2014.

[7] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.

[8] K. Kumar, J. Liu, Y. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Netw. Appl.*, vol. 18, no. 1, pp. 129–140, Feb. 2013.
[9] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
[10] E. Cuervo *et al.*, "MAUI: Making smartphones last longer with code offload," in *Proc. IEEE Int. Conf. Mobile Syst. Appl. Services (MobiSys)*, San Francisco, CA, USA, Jun. 2010, pp. 49–62.
[11] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "CloneCloud: Elastic execution between mobile device and cloud," in *Proc. 6th Conf. Comput. Syst. (EuroSys)*, Salzburg, Austria, Apr. 2011, pp. 301–314.
[12] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Orlando, FL, USA, Mar. 2012, pp. 945–953.
[13] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "CONCERT: A cloud-based architecture for next-generation cellular systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 14–22, Dec. 2014.
[14] 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.
[15] 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.
[16] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. S. Quek, "Offloading in mobile edge computing: Task allocation and computational frequency scaling," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3571–3584, Aug. 2017.
[17] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.
[18] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1757–1771, May 2016.
[19] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.
[20] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.
[21] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang, and R. P. Liu, "Energy-efficient admission of delay-sensitive tasks for mobile edge computing," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2603–2616, Jun. 2018.
[22] C. Wang, F. R. Yu, C. Liang, Q. Chen, and L. Tang, "Joint computation offloading and interference management in wireless cellular networks with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7432–7445, Aug. 2017.
[23] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.
[24] Q. Pham, T. Leanh, N. H. Tran, and C. S. Hong, "Decentralized computation offloading and resource allocation in heterogeneous networks with mobile edge computing," 2018. [Online]. Available: arXiv:1803.00683.
[25] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint computation offloading and user association in multi-task mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12313–12325, Dec. 2018.
[26] J. Du, L. Zhao, J. Feng, and X. Chu, "Computation offloading and resource allocation in mixed fog/cloud computing systems with min–max fairness guarantee," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1594–1608, Apr. 2018.
[27] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.
[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. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994–6009, Sep. 2017.
[30] Z. Wei, B. Zhao, J. Su, and X. Lu, "Dynamic edge computation offloading for Internet of Things with energy harvesting: A learning method," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4436–4447, Jun. 2019.
[31] M. Min, L. Xiao, Y. Chen, P. Cheng, D. Wu, and W. Zhuang, "Learning-based computation offloading for IoT devices with energy harvesting," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1930–1941, Feb. 2019.
[32] T. T. Nguyen, L. Le, and Q. Le-Trung, "Computation offloading in MIMO based mobile edge computing systems under perfect and imperfect CSI estimation," *IEEE Trans. Services Comput.*, early access, Jan. 14, 2019, doi: 10.1109/TSC.2019.2892428.
[33] K. Wang, T. Chang, W. Ma, and C. Chi, "A semidefinite relaxation based conservative approach to robust transmit beamforming with probabilistic SINR constraints," in *Proc. 18th Eur. Signal Process. Conf.*, Aalborg, Denmark, Aug. 2010, pp. 407–411.
[34] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.
[35] T. Van Chien, E. Bjornson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6384–6399, Sep. 2016.
[36] C. Pan, H. Ren, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Weighted sum-rate maximization for the ultra-dense user-centric TDD C-RAN downlink relying on imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1182–1198, Feb. 2019.
[37] S. Melendez and M. P. McGarry, "Computation offloading decisions for reducing completion time," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2017, pp. 160–164.
[38] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.
[39] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM J. Optim.*, vol. 12, no. 2, pp. 479–502, 2002.
[40] S. Kim, R. Pasupathy, and S. G. Henderson, "A guide to sample average approximation," in *Handbook of Simulation Optimization*. New York, NY, USA: Springer, 2015, pp. 207–243.
[41] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.
[42] M. Mahdavi, T. Yang, and R. Jin, "Online stochastic optimization with multiple objectives," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2013.
[43] A. Liu, V. K. N. Lau, and B. Kananian, "Stochastic successive convex approximation for non-convex constrained stochastic optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4189–4203, Aug. 2019.
[44] A. Liu, V. K. N. Lau, and M. Zhao, "Online successive convex approximation for two-stage stochastic nonconvex optimization," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5941–5955, Nov. 2018.
[45] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento, "A parallel decomposition method for nonconvex stochastic multi-agent optimization problems," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2949–2964, Jun. 2016.
[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[47] Q. Wu, M. Tao, D. W. K. Ng, W. Chen, and R. Schober, "Energy-efficient resource allocation for wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2312–2327, Mar. 2016.

Jun Wang received the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2017.

He was an Assistant Professor with China Three Gorges University, Yichang, China, from 2017 to 2018. Since May 2018, he has been a Postdoctoral Research Fellow with Shenzhen University, Shenzhen, China. From August 2019 to August 2020, he was a Visiting Scholar with the Department of ECE, University of Waterloo, Waterloo, ON, Canada. His current research interests include 5G, mobile-edge computing, and Internet of Things.



Daquan Feng (Member, IEEE) received the Ph.D. degree in information engineering from the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China, in 2015.

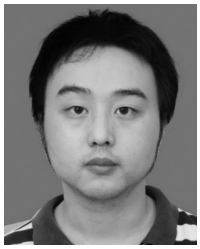
He was a Research Staff with State Radio Monitoring Center, Beijing, China, and then a Postdoctoral Research Fellow with the Singapore University of Technology and Design, Singapore. He was a visiting student with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, from 2011 to 2014. He is currently an Assistant Professor with the Shenzhen Key Laboratory of Digital Creative Technology, Guangdong Province Engineering Laboratory for Digital Creative Technology, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include URLLC communications, MEC, and massive IoT networks.

Dr. Feng is an Associate Editor of IEEE COMMUNICATIONS LETTERS.

Shengli Zhang (Senior Member, IEEE) received the B.Eng. degree in electronic engineering and the M.Eng. degree in communication and information engineering from the University of Science and Technology of China, Hefei, China, in 2002 and 2005, respectively, and the Ph.D. degree from the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong, in 2008.

Then, he joined the Communication Engineering Department, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. From March 2014 to March 2015, he was a Visiting Associate Professor with Stanford University, Stanford, CA, USA. He is the pioneer of physical-layer network coding. He has published more than 30 IEEE top journal papers and ACM top conference papers, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON COMMUNICATIONS, and ACM Mobicom. His research interests include physical-layer network coding, cooperative wireless networks, and blockchain.

Dr. Zhang severed as an Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE WIRELESS COMMUNICATIONS LETTERS, and *IET Communications*.



An Liu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Peking University, Beijing, China, in 2011 and 2004, respectively.

From 2008 to 2010, he was a Visiting Scholar with the Department of ECEE, University of Colorado at Boulder, Boulder, CO, USA. He has been a Postdoctoral Research Fellow from 2011 to 2013, a Visiting Assistant Professor in 2014, and a Research Assistant Professor with the Department of ECE, Hong Kong University of Science and Technology, Hong Kong, from 2015 to 2017. He is currently a Distinguished Research Fellow with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His research interests include wireless communications, stochastic optimization, compressive sensing, and machine/deep learning for communications.

Dr. Liu is serving as an Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE WIRELESS COMMUNICATIONS LETTERS.

Xiang-Gen Xia (Fellow, IEEE) received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1983, the M.S. degree in mathematics from Nankai University, Tianjin, China, in 1986, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1992.

He was a Senior/Research Staff Member with Hughes Research Laboratories, Malibu, CA, USA, from 1995 to 1996. In September 1996, he joined the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA, where he is the Charles Black Evans Professor. He has authored the book *Modulated Coding for Intersymbol Interference Channels* (New York, NY, USA: Marcel Dekker, 2000). His current research interests include space-time coding, MIMO and OFDM systems, digital signal processing, and SAR and ISAR imaging.

Dr. Xia received the National Science Foundation Faculty Early Career Development (CAREER) Program Award in 1997, the Office of Naval Research Young Investigator Award in 1998, and the Outstanding Overseas Young Investigator Award from the National Nature Science Foundation of China in 2001, the 2019 Information Theory Outstanding Overseas Chinese Scientist Award, and the Information Theory Society of Chinese Institute of Electronics. He has served as an Associate Editor for numerous international journals, including IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON MOBILE COMPUTING, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He is the Technical Program Chair of the Signal Processing Symposium Globecom 2007 in Washington, DC, USA, and the General Co-Chair of ICASSP 2005 in Philadelphia.