Optimized Multiuser Computation Offloading with Multi-antenna NOMA

Feng Wang, Jie Xu, and Zhiguo Ding

1School of Information Engineering, Guangdong University of Technology, Guangzhou, China
2School of Computing and Communications, Lancaster University, Lancaster, UK
Email: fengwang13@gdut.edu.cn, jieyu@gdut.edu.cn, z.ding@lancaster.ac.uk

Abstract—Mobile edge computing (MEC) has been regarded as a promising technique to enhance the computation capabilities of wireless devices, by enabling them to offload computation-intensive tasks to base stations (BSs) at the network edge. This paper studies a new multi-user MEC system with multi-antenna non-orthogonal multiple access (NOMA)-based computation offloading. In this system, multiple users simultaneously offload their computation tasks to one multi-antenna BS over the same time/frequency resources for remote execution, and the BS uses successive interference cancellation (SIC) for information decoding. We consider the partial offloading case, such that each user can partition the computation task into two parts for local computing and offloading, respectively. Under this setup, we minimize the weighted sum of the energy consumption at all users subject to their computation latency constraints. The decision variables include the task partition, local central processing unit (CPU) frequencies, and offloading power and rates at the users, and the SIC decoding order at the BS. We present an efficient algorithm to obtain the globally optimal solution to this problem by applying the Lagrange dual method. Numerical results show that the proposed NOMA-based partial offloading design can significantly improve the energy efficiency of the multiuser MEC system, as compared to benchmark schemes with orthogonal multiple access (OMA)-based partial offloading, and with only local computing or full offloading.

I. INTRODUCTION

Future Internet of things (IoT) needs to support emerging ultra-low-latency applications such as augmented reality, autonomous driving, and tele-surgery [1]. Towards this end, the next-generation wireless networks need to accommodate billions of wireless devices with real-time communication and communication capabilities. This is a challenge task, since the IoT devices are normally with low power and small size, while the computation tasks are generally intensive and latency-critical. Conventionally, mobile cloud computing (MCC) is an efficient technique to provide IoT devices with ample computation resources at (remote) centralized cloud servers. However, due to the long propagation distance between the IoT devices and the cloud, MCC may not be able to meet the critical latency requirements for emerging IoT applications. To overcome the limitation of MCC, mobile edge computing (MEC) has been recently proposed as an alternative solution to reduce the computation latency for these IoT devices. In MEC systems, distributed MEC servers are dedicatedly deployed at the network edge (such as cellular base stations (BSs)) to provide cloud-like computing, such that these IoT devices can offload their computation-intensive tasks to the BSs for remote execution [2]–[4].

To fully reap the aforementioned benefits in MEC, some technical challenges require to be addressed, especially for radio/computation resource allocation in multiuser computation offloading. Specifically, in order to simultaneously support the remote computation of multiple users, the wireless communication resources (e.g., the time and frequency resources) at the BSs and the computation resources at the MEC servers should be efficiently shared among multiple end users. To cope with this issue, various joint communication and computation resource allocation approaches have been proposed in the literature to improve the performance of computation offloading in multiuser MEC systems (see e.g., [5]–[10]). For example, the authors in [5]–[7] studied the binary offloading case when the computation tasks at each user is not partitionable but can only be offloaded as a whole, in which frequency-division multiple access (FDMA) [5], [6] and code division multiple access (CDMA) [7] are employed for multiuser computation offloading. On the other hand, the works [8]–[10] studied the partial offloading case when the computation task at each user is partitionable to be different parts for the user’s local computing and the MEC server’s remote computing, respectively, in which time-division multiple access (TDMA) or orthogonal frequency-division multiple access (OFDMA) based multiuser computation offloading are considered. Despite the recent research progress, these existing works considered generally suboptimal multiple access techniques for offloading, such as orthogonal schemes (e.g., TDMA and FDMA) in [5], [6], [8]–[10] and non-orthogonal CDMA schemes by treating interference as noise in [7]. These approaches, however, cannot fully explore the capacity of the multiple access channel from the multiple users to the BS, and thus may fundamentally limit the performance of multiuser MEC systems.

Recently, non-orthogonal multiple access (NOMA) has been recognized as one of the key techniques in the fifth-generation (5G) cellular networks [11]–[13]. Unlike conventional orthogonal multiple access (OMA) techniques such as FDMA and TDMA, NOMA enables multiple users to communicate with the BS concurrently at the same time/frequency/code resources. By implementing sophisticated multiuser detection techniques such as successive interference cancellation (SIC) at the receivers, the NOMA system is expected to achieve a much higher spectral efficiency than the OMA counterpart [14]–[16]. Actually, for an uplink NOMA system or equivalently a multiple access channel from users to BS, it has been well established that the capacity region is achievable when the users employ Gaussian signaling with optimally controlled coding rate, and the BS receiver adopts the minimum mean
square error (MMSE)-SIC decoding with a properly designed decoding order for different users (see, e.g., [14], [15]). Motivated by the benefit of NOMA over OMA, it is expected that NOMA can be exploited as an efficient method to further improve the performance of multiuser computation offloading for MEC.

In this paper, we consider a multi-antenna NOMA-based MEC system consisting of one multi-antenna BS (integrated with an MEC server) and multiple single-antenna users. Each user has certain computation tasks that need to be successfully executed within a finite-duration block. With NOMA, these users can simultaneously offload their computation input bits to the BS over the same time/frequency resources. We focus on the partial offloading case for the purpose of initial investigation. Under this setup, we pursue an energy-efficient design to minimize the weighted sum of the energy consumption at all users while ensuring the successful execution of all the tasks within this block, by jointly optimizing each user’s task partition, local central processing unit (CPU) frequencies, and offloading power and rate, as well as the BS’s decoding order for MMSE-SIC. To our best knowledge, this is the first attempt to utilize the emerging NOMA technique for multiuser computation offloading to improve the performance of MEC systems. However, the weighted sum-energy minimization problem is generally difficult to solve, which is due to the fact that this problem does not admit explicit functions for the users’ offloading rates due to the initially unknown decoding order at the BS. To tackle this challenge, we present an efficient algorithm to obtain the globally optimal solution to this problem by applying the Lagrange dual method. Numerical results show that the proposed NOMA-based partial offloading design significantly improves the energy efficiency of the multiuser MEC system, as compared to benchmark schemes with OMA-based partial offloading, and with only local computing only or full offloading.

Notations: \( \mathbb{C}^{x \times y}, \mathbb{R}^{x \times y}, \) and \( \mathbb{R}_+^{x \times y} \) denote the sets of all complex-valued, real-valued, and non-negative real-valued \( x \times y \) matrices, respectively; \( I \) denotes an identity matrix with appropriate dimensions; \( E[\cdot] \) denotes the statistical expectation. \( \|x\| \) and \( x^t \) denote the Euclidean norm and the transpose of a vector \( x \), respectively; \( |x| \) denotes the absolute value of a scalar \( x \). \( |S| \) denotes the cardinality of a set \( S \). The distribution of a circularly symmetric complex Gaussian (CSCG) random vector \( x \) with mean \( \mu \) and covariance \( \Sigma \) is denoted by \( x \sim \mathcal{CN}(\mu, \Sigma) \), where \( \sim \) stands for “distributed as”.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider an MEC system consisting of one single BS equipped with \( N \) antennas and a set \( K = \{1, \ldots, K\} \) of users each with a single antenna. The BS is integrated with an MEC server to help remotely execute the offloaded computational tasks from the \( K \) users. The computation offloading is implemented over a frequency band with bandwidth \( B \). We consider a frequency non-selective quasi-static channel model, in which the wireless channels remain unchanged during a given block of our interest with finite duration \( T \). Within this block, each user \( k \in K \) has a computation task with \( L_k \) input bits in total that should be executed before the end of this block. We consider the partial offloading case, in which the task at user \( k \) can be arbitrarily partitioned into two parts for local computing and offloading, with the numbers of input bits being denoted by \( \ell_k \) and \( (L_k - \ell_k) \), respectively, where \( 0 \leq \ell_k \leq L_k \). It is assumed that the BS perfectly knows the computation information of all the \( K \) users and the channel state information (CSI) from/to the \( K \) users. In accordance with such information, the BS can coordinate the computation offloading and the local computing for these users.

A. NOMA-based Multiuser Partial Offloading

The multiuser computation offloading is implemented as shown in Fig. 1, where the duration-\( T \) block of interest is divided into three slots, for the users’ task offloading in the uplink, the remote task execution at the BS, and the users’ computation results downloading, respectively. In the first slot with duration \( \Delta^x \), each user \( k \in K \) offloads partial of its input bits to the BS based on NOMA. After collecting these bits from the \( K \) users, in the second slot with duration \( \Delta^{exe} \), the MEC server at the BS remotely executes the offloaded tasks on behalf of these users by developing multiple virtual machines (VMs) each for one user [3]. In the third slot with duration \( \Delta^{rx} \), the BS sends the computation results back to these \( K \) users. In order for each user \( k \in K \) to obtain the computation results before the end of this block, we have
\[
\Delta^x + \Delta^{exe} + \Delta^{rx} = T, \tag{1}
\]
where the strict equality is set in order to minimize the related communication and computation energy consumption. As the MEC server generally has sufficient computation resource, it can adopt a high CPU frequency to minimize the execution time for the offloaded tasks. As a result, we consider the execution time at the MEC server to be negligible, i.e., \( \Delta^{exe} \approx 0 \). Furthermore, as the computed results are of small size and the BS is with high transmit power in general, we also consider the time for these users to download the computed results is negligible, i.e., \( \Delta^{rx} \approx 0 \). Therefore, we re-express (1) as
\[
\Delta^x = T. \tag{2}
\]

Now, we focus on the multiuser computation offloading in the first slot. The \( K \) users employ uplink NOMA to offload their respective computation input bits to the BS.
simultaneously. Let \( x_k \in C \) denote the unit-power task-bearing signal for offloading by user \( k \) with \( \mathbb{E}|x_k|^2 = 1 \), and \( p_k > 0 \) the associated transmit power. The received signal \( y \in \mathbb{C}^{N \times 1} \) by the BS is then expressed as [13]

\[
y = \sum_{k=1}^{K} \sqrt{p_k} h_k x_k + z,
\]

where \( h_k \in \mathbb{C}^{N \times 1} \) is the uplink channel vector from user \( k \) to the BS, and \( z \in \mathbb{C}^{N \times 1} \) denotes the additive white Gaussian noise at the BS, which is normalized to be of unit power with \( z \sim \mathcal{CN}(0, I) \).

At the receiver side, the BS employs the MMSE-SIC to decode information from the \( K \) users. Let the permutation \( \pi \) over \( K \) denote the successive decoding order at the BS: the BS receiver first decodes the information \( x_{\pi(1)} \) transmitted by user \( \pi(1) \), then decodes \( x_{\pi(2)} \) by cancelling the interference from \( x_{\pi(1)} \), followed by \( x_{\pi(3)} \), \( x_{\pi(4)} \), and so on, until \( x_{\pi(K)} \). By employing capacity-achieving Gaussian signaling at these users (e.g., setting \( x_k \)'s to be CSGC random variables), the achievable rate (in bits/sec/Hz) at user \( \pi(k) \) under given decoding order \( \pi \) is [14]

\[ r_{\pi(1)} = \log_2 \left( 1 + \frac{\sum_{i=1}^{k-1} p_{\pi(i)} \|h_{\pi(i)}\|^2}{\sum_{i=1}^{k-1} p_{\pi(i)} \|h_{\pi(i)}\|^2} \right), \quad k \in \mathcal{K}. \]

By allowing the BS to properly design the decoding order and employ time-sharing among different decoding orders, the capacity region for the \( K \) users is achievable and it corresponds to the following polymatroid [14]:

\[ \mathcal{C}(p) = \left\{ r \in \mathbb{R}_+^{K \times 1} : \sum_{k \in J} r_k \leq \log_2 \left( 1 + \sum_{k \in J} p_k \|h_k\|^2 \right), \forall J \subseteq \mathcal{K} \}, \]

where \( r_k \) denotes the achievable rate of user \( k \in \mathcal{K} \), \( r \triangleq [r_1, \ldots, r_K] \), and \( p \triangleq [p_1, \ldots, p_K] \).

Over the block with duration \( T \), the maximum number of bits that can be offloaded from user \( k \) to the BS is given by \( r_k BT \) (with bandwidth \( B \)), which should be no smaller than \( (L_k - \ell_k) \). Therefore, we have

\[ r_k \geq \frac{L_k - \ell_k}{B}, \quad \forall k \in \mathcal{K}. \]

Furthermore, we consider the transmit energy consumption as the sole energy budget at user \( k \in \mathcal{K} \) for computation offloading, which is expressed as

\[ E_{TX}^k = p_k T. \]

\[ (P1) : \min_{\ell, p, r} \sum_{k=1}^{K} \alpha_k \left( \frac{\zeta_k C_k i_k^3}{T^2} + p_k T \right) \]

s.t. (6) and \( r \in \mathcal{C}(p) \)

\[ p_k \geq 0, \quad \forall k \in \mathcal{K} \]

\[ 0 \leq \ell_k \leq L_k, \quad \forall k \in \mathcal{K}. \]
to the fact that this problem does not admit explicit functions for \( r_k \)'s due to the initially unknown decoding order at the BS. How to infer the optimal decoding order from the optimal solution to (P1) is important for practical implementation, but it is generally a challenging task. On the other hand, based on the definition of \( C(p) \) in (5), (11b) corresponds to a total number of \( (2^K - 1) \) inequality constraints, which increases exponentially with respect to \( K \). Therefore, it is practically infeasible to directly consider these constraints in (P1).

III. OPTIMAL SOLUTION TO PROBLEM (P1)

To tackle the difficulty, in this section we leverage the Lagrangian dual method to solve problem (P1) to obtain the optimal task partition \( \ell \), transmit powers \( p \) and rates \( r \) for offloading, as well as the associated decoding order at the BS receiver.

As (P1) is convex and satisfies the Slater’s condition, strong duality holds between (P1) and its dual problem. Let \( \lambda_k \geq 0 \) denote the Lagrangian dual variable associated with the \( k \)th constraint in (6), and define \( \lambda \triangleq [\lambda_1, \ldots, \lambda_K]^\top \). The partial Lagrangian of (P1) is given by

\[
L(\ell, p, r, \lambda) = \sum_{k=1}^{K} \alpha_k \left( \frac{\zeta_k C_k^3 \ell_k^2}{T^2} + p_k T \right) + \sum_{k=1}^{K} \lambda_k \left( L_k - \frac{\ell_k}{BT} - r_k \right). \tag{12}
\]

The Lagrange dual function is then defined as

\[
g(\lambda) = \min_{\ell, p, r} L(\ell, p, r, \lambda) \tag{13a}
\]

\[
s.t. \quad r \in C(p), \quad p_k \geq 0, \quad \forall k \in K \tag{13b}
\]

\[
0 \leq \ell_k \leq L_k, \quad \forall k \in K. \tag{13c}
\]

Correspondingly, the Lagrange dual problem of (P1) is

\[
(D1): \max_{\lambda} g(\lambda) \tag{14a}
\]

\[
s.t. \quad \lambda_k \geq 0, \quad \forall k \in K. \tag{14b}
\]

In the following, we solve problem (P1) by first solving problem (13) to obtain \( g(\lambda) \) under given \( \lambda \), and then solving (D1) to find the optimal \( \lambda \). Denote by \( (\ell^{\text{opt}}, p^{\text{opt}}, r^{\text{opt}}) \) and \( \lambda^{\text{opt}} \) the optimal solutions to (P1) and (D1), respectively.

A. Evaluating \( g(\lambda) \) by Solving Problem (13)

First, we obtain the dual function \( g(\lambda) \) under given \( \lambda \) by solving problem (13). The Lagrangian in (12) can be rearranged as

\[
L(\ell, p, r, \lambda) = \sum_{k=1}^{K} T \alpha_k p_k - \sum_{k=1}^{K} \lambda_k r_k + \sum_{k=1}^{K} \left( \frac{\alpha_k \zeta_k C_k^3 \ell_k^2}{T^2} - \frac{\lambda_k}{BT} \ell_k \right) + \sum_{k=1}^{K} \lambda_k L_k / BT. \tag{15}
\]

Therefore, problem (13) can be decomposed into the \((K + 1)\) subproblems as follows by dropping the constant term \( \sum_{k=1}^{K} \lambda_k L_k / (BT) \), where (16) corresponds to \( K \) subproblems each for one user \( k \in K \) as follows.

\[
\min_{0 \leq \ell_k \leq L_k} \frac{\alpha_k \zeta_k C_k^3}{T^2} \ell_k^2 + \frac{\lambda_k}{BT} \ell_k, \quad \forall k \in K. \tag{16}
\]

\[
\min_{r, p} \sum_{k=1}^{K} T \alpha_k p_k - \sum_{k=1}^{K} \lambda_k r_k \tag{17a}
\]

\[
s.t. \quad r \in C(p), \quad p_k \geq 0, \quad \forall k \in K. \tag{17b}
\]

For convenience of presentation, let \( \ell^*_k \) and \( (p^{*}, r^{*}) \) denote the optimal solutions to the \( k \)th subproblem in (16) and the subproblem in (17), respectively.

For each subproblem \( k \in K \) in (16), \( \ell^*_k \) can be obtained based on the Karush-Kuhn-Tucker (KKT) conditions [17].

**Lemma 3.1:** The optimal solution \( \ell^*_k \) to (16) is given by

\[
\ell^*_k = \min \left( \sqrt{\frac{T \lambda_k}{3B \alpha_k \zeta_k C_k}} L_k \right), \quad \forall k \in K. \tag{18}
\]

**Proof:** See Appendix A. \[\blacksquare\]

In order to solve subproblem (17), we borrow the following lemma that is verified in [15].

**Lemma 3.2:** For any given \( \lambda \) and \( p \), the optimal solution to

\[
\max_{r} \sum_{k=1}^{K} \lambda_k r_k \quad s.t. \quad r \in C(p) \tag{19}
\]

is obtained by one vertex \( r^{(\pi)} \triangleq [r_1^{(\pi)}, \ldots, r_K^{(\pi)}] \) of the polymatroid \( C(p) \), where \( r_k^{(\pi)} \) is given in (4) and the permutation \( \pi \) is determined such that \( \lambda_{\pi(1)} \geq \cdots \geq \lambda_{\pi(K)} \geq 0 \).

Based on Lemma 3.2 and by substituting (4), the subproblem (17) reduces to optimizing over the transmit powers \( p \) as follows.

\[
\max_{p} \sum_{k=1}^{K} \left[ - T \alpha_k p_k + (\lambda_{\pi(k)} - \lambda_{\pi(k+1)}) \right] \times \log_2 \left( 1 + \sum_{i=1}^{k} p_{\pi(i)} \| h_{\pi(i)} \|^2 \right) \tag{20a}
\]

\[
s.t. \quad p_k \geq 0, \quad \forall k \in K. \tag{20b}
\]

where \( \lambda_{\pi(k+1)} = 0 \) is defined for notational convenience. Note that problem (20) is convex and the optimal \( p^{*} \) can thus be efficiently obtained by standard convex solvers, e.g., CVX [18]. By substituting \( p^{*} \) into (4), the offloading rate \( r^{*} \) can be obtained.

**Remark 3.1:** It is worth noting that the transmit power \( p^{*} \) for offloading is unique due to the strict convexity of problem (20) (also see [16]). However, the offloading rate \( r^{*} \) is generally non-unique; this is due to the fact that if there exist any two users \( i \) and \( j \) such that \( \lambda_i = \lambda_j = \lambda_{ij} = \lambda_{ij} \), then the two decoding orders (i.e., decoding the message of user \( i \) first followed by that of user \( j \), and the reverse order) are both optimal for problem (17). Therefore, suppose that \( J_1, \ldots, J_M \subseteq K \) denote \( M \) disjoint subsets such that \( \lambda_{i} \)'s are equal, \( \forall k \in J_m \), for each \( 1 \leq m \leq M \) and \(|J_m| \geq 2\); then there generally exist a total of \( \prod_{m=1}^{M} |J_m| \) optimal decoding
orders for problem (20). Here, one can choose any one of them only for the purpose of evaluating the dual function \( g(\lambda) \).

By combining the optimal solutions of \( \ell_k^\ell \)'s for the subproblems in (16) and \((p^\ell, r^\ell)\) for subproblem (17), the dual function \( g(\lambda) \) in (13) is finally obtained.

**Algorithm 1 for Solving Problem (P1)**

1. **Initialization:** Given an ellipsoid \( E^{(0)} \subseteq \mathbb{R}^{K+1} \), centered at \( \lambda^{(0)} \) and containing the optimal dual solution \( \lambda^{\text{opt}} \).
2. Set \( \ell = 1 \).
3. **Repeat:**
   - Obtain \( r^\ell \) based on (18); set the permutation \( \pi \) such that \( \lambda_{\pi(1)} \geq \cdots \geq \lambda_{\pi(K)} \geq 0 \), then obtain \( p^\ell \) by solving problem (20), and get \( r^\ell \) from (4).
   - Update the ellipsoid \( E^{(\ell+1)} \) based on \( E^{(\ell)} \) and the subgradient of \( g(\lambda) \) given by (21). Set \( \lambda^{(\ell+1)} \) as the center of ellipsoid \( E^{(\ell+1)} \).
   - Set \( \ell = \ell + 1 \).
4. **Until** the stopping criteria for the ellipsoid method is met.
5. Set \( \lambda^{\text{opt}} \leftarrow \lambda^{(\ell)} \).
6. **Output:** Obtain \( \ell^{\text{opt}} \) based on (18) by replacing \( \lambda \) with \( \lambda^{\text{opt}} \), obtain \( p^{\text{opt}} \) by solving problem (20) with \( \lambda^{\text{opt}} \), and construct the optimal offloading rate \( r^{\text{opt}} \) and the associated decoding order based on Proposition 3.1.

**B. Finding the Optimal \( \lambda^{\text{opt}} \) to Maximize \( g(\lambda) \) in (D1)**

With \( g(\lambda) \) obtained, we solve the dual problem (D1) to obtain the optimal \( \lambda^{\text{opt}} \) to maximize \( g(\lambda) \). Note that \( g(\lambda) \) is convex but may not be differentiable in general. As a result, (D1) can be solved by subgradient based methods such as the ellipsoid method [19], by using the fact that the subgradient of the objective function \( g(\lambda) \) is

\[
\biggl[ r_1^m - L_1 - \ell_k^m \frac{B}{T}, \ldots, r_K^m - L_K - \ell_K^m \frac{B}{T} \biggr].
\]

**C. Constructing Primal Optimal Solution and Decoding Order for (P1)**

Based on the dual optimal solution \( \lambda^{\text{opt}} \) to (D1), we need to obtain the primal optimal solution \((\ell^{\text{opt}}, p^{\text{opt}}, r^{\text{opt}})\) to (P1), as well as the optimal decoding order at the BS receiver. Note that under any given \( \lambda \), the optimal solution of \( \ell^* \) and \( p^* \) to problem (13) is unique. Therefore, by replacing \( \lambda^{\text{opt}} \) with \( \lambda^* \) in Lemma 3.1 and solving problem (20) with \( \lambda^{\text{opt}} \), one can obtain the primal optimal \( \ell^{\text{opt}} \) and \( p^{\text{opt}} \) to (P1).

It remains to determine the primal optimal offloading rate \( r^{\text{opt}} \) and the associated optimal decoding order at the BS for the MMSE-SIC. First, consider the case when \( \lambda^{\text{opt}}'s \) are different from each other. Denote \( \pi^{\text{opt}} \) as the permutation such that \( \lambda^{\text{opt}}_{\pi^{\text{opt}}(1)} > \cdots > \lambda^{\text{opt}}_{\pi^{\text{opt}}(K)} \geq 0 \), which then corresponds to the optimal decoding order at the BS receiver for (P1). In this case, the primal optimal offloading rate to (P1) is obtained as \( r^{\text{opt}} = r^{\pi^{\text{opt}}} \) by (4).

Next, we consider the case when there exist some \( \lambda^{\text{opt}}'s \) that are equal to each other. In this case, it is shown in Remark 3.1 that the offloading rate \( r^* \) and the associated decoding order to problem (17) (and thus (13)) are generally not unique, and thus may not be primal optimal to (P1). Therefore, we need an additional step to construct the primal optimal offloading rate \( r^{\text{opt}} \) to (P1) via proper time-sharing among different decoding orders as follows.

In particular, let \( J_1, \ldots, J_M \subseteq K \) denote \( M \) disjoint subsets such that \( \lambda^{\text{opt}}_{J_i} \)'s are equal, \( \forall j \in J_i \), for each \( 1 \leq m \leq M \) and \(|J_m| \geq 2\). Define the set \( I \triangleq \{1, \ldots, K\} \setminus \bigcup_{m=1}^M |J_m| \}. As a result, under the optimal dual solution \( \lambda^{\text{opt}} \), problem (17) (or equivalently (13)) admits a number of \(|Z| \) optimal decoding orders, denoted by \( \pi^{(1)}, \ldots, \pi^{(|Z|)}, \) and \(|Z| \) associated optimal offloading rates by (4), denoted by \( r^{(1)}, \ldots, r^{(|Z|)} \). Here, each rate-tuple \( r^{i} \) corresponds to one vertex of the polymatroid \( C(p^{\text{opt}}) \); however, the primal optimal rate-tuple \( r^{\text{opt}} \) may lie on a surface of this polymatroid in order to satisfy (6) in problem (P1). Hence, it is necessary to employ time-sharing among the \(|Z| \) number of \( r^{(i)}'s \). Towards this end, we partition the duration-\( T \) block into a total number of \(|Z| \) time slots. In each slot \( i \in \{1, \ldots, |Z|\} \) with duration \( t^{(i)} \), the \( K \) users transmit with rates \( r^{(i)} \) and the BS decodes with order \( \pi^{(i)} \).

In order to find the optimal time-sharing strategy to solve problem (P1) while satisfying (6), we obtain the optimal duration \( t^{(i)}'s \) by solving the following feasibility problem:

Find \( t^{(1)}, \ldots, t^{(|Z|)} \)
\[
\begin{align*}
\text{s.t.} \quad \sum_{i \in Z} r^{(i)}t^{(i)} & \geq L_k - \ell_k \frac{B}{T}, \quad \forall k \in K \\
\sum_{i \in Z} t^{(i)} & \leq T \\
0 \leq t^{(i)} & \leq T, \quad \forall i \in Z.
\end{align*}
\]

Note that (22) is a linear program (LP) and can thus be efficiently solved via CVX. Based on the above analysis, we have the optimal primal solution of the offloading rate \( r^{\text{opt}} \) to (P1) in the following proposition.

**Proposition 3.1:** When there exist some \( \lambda^{\text{opt}}'s \) that are equal to each other, the optimal offloading rate \( r^{\text{opt}} \) to (P1) is obtained by time-sharing among the \(|Z| \) optimal offloading rates, i.e., \( r^{(1)}, \ldots, r^{(|Z|)} \), by partitioning the block into \(|Z| \) slots, with \( t^{(i)} \) obtained by (22) denoting the duration of slot \( i \in \{1, \ldots, |Z|\} \). Then, in each slot \( i \), the \( K \) users transmit with offloading rate \( r^{(i)} \) and the BS receiver decodes with the corresponding order \( \pi^{(i)} \).

In summary, we present Algorithm 1 to obtain the optimal solution \((\ell^{\text{opt}}, p^{\text{opt}}, r^{\text{opt}})\) to the computation latency-constrained weighted sum-energy consumption minimization problem (P1).

**IV. Numerical Results**

In this section, we provide numerical results to validate the performance of the proposed NOMA-based partial offloading scheme for the multiuser MEC system, as compared with the following three benchmark schemes.

1) **Full offloading only:** All the \( K \) users execute their computation tasks by offloading the computation input bits to the BS via an NOMA protocol. This scheme corresponds to solving (P1) by setting \( \ell_k = 0, \forall k \in K \).
Energy consumption at user 2, \( E_2 \) (J)

Average sum energy at users (J)

\[
\begin{align*}
E_1 & = 0.4, 0.8, 1.2, 1.6, 2, 2.2 \\
E_2 & = 0.4, 0.8, 1.2, 1.6, 2, 2.2 \\
\end{align*}
\]

∀ \( \text{input bits at different users are set to be identical, i.e., } \zeta_k \) the computation at each user, we set \( \text{density (PSD) at the BS receiver as } \delta_k \) the computation latency constraints. Note that in the case with \( \alpha_1 = \alpha_2 \) under our setup, the OMA-based partial offloading generally achieves lower energy consumption at both users than the other two benchmark schemes. This validates the effectiveness of our proposed NOMA-based partial offloading in minimizing the users’ energy consumption while satisfying the computation latency constraints. Note that in the case with \( \alpha_1 = \alpha_2 \) under our setup, the OMA-based partial offloading is observed to achieve the same energy-saving performance as the NOMA-based counterpart.

Next, we consider a more general scenario when the BS is equipped with \( N = 4 \) antennas to serve a total of \( K = 4 \) users. We consider the Rayleigh fading channel model, and express the channel vector between each user \( k \in K \) and the BS as \( h_k = \sqrt{G_0 \left( \frac{d_k}{d_0} \right)^{-\theta}} \tilde{h}_k \), where \( G_0 = -40 \) dB corresponds to the path loss at a reference distance of \( d_0 = 1 \) m, \( d_k \) denotes the distance between user \( k \) and the BS, \( \theta = 3.7 \) denotes the path loss exponent, and \( \tilde{h}_k \) is a CSCG random vector with \( \tilde{h}_k \sim CN(0, I) \). We consider that the four users are located with different distances to the BS, and particularly set \( d_1 = 200 \) m, \( d_2 = 300 \) m, \( d_3 = 400 \) m, and \( d_4 = 500 \) m. The rate weights for different users are set to be identical as \( \alpha_k = 1 \), \( \forall k \in K \). Then our objective corresponds to minimizing the sum energy consumption at all the \( K \) users. The results are obtained by averaging over 500 randomized channel realizations.

Fig. 3 shows the average sum energy consumption at these users versus the number of computation input bits \( L \) at each user.

2) Local computing only: All the \( K \) users execute their computation tasks via local computing only. This scheme corresponds to solving (P1) by setting \( \ell_k = L_k, \forall k \in K \). The resultant weighted sum-energy consumption at the users can be expressed in closed-form as \( \sum_{k=1}^{K} \alpha_k \zeta_k C_k^2 L_k^2 / T^2 \).

3) OMA-based partial offloading [8]: The \( K \) users adopt a TDMA protocol for computation offloading, where the duration-\( T \) block is partitioned into \( K \) slots. In the \( k \)th time slot, user \( k \) offloads part of its computation input bits to the BS. We jointly optimize the time slot allocation among the users, together with the transmit power and the number of computation input bits for each user’s offloading, as well as the local CPU frequency per user, in order to minimize the weighted sum-energy consumption at the \( K \) users.

In the simulation, we set the system bandwidth for computation offloading as \( B = 1 \) MHz and the noise power spectrum density (PSD) at the BS receiver as \( N_0 = -174 \) dBm/Hz. For the computation at each user, we set \( C_k = 10^5 \) cycles/bit, \( \zeta_k = 10^{-28}, \forall k \in K \). In addition, the numbers of computation input bits at different users are set to be identical, i.e., \( L = L_k, \forall k \in K \).

First, we consider the basic scenario with a single-antenna BS (with \( N = 1 \)) serving \( K = 2 \) users. The number of computation input bits at each user is set as \( L = 6 \times 10^5 \), the block duration is \( T = 0.1 \) sec, and the power loss of the signal propagation from each user to the BS is \( |h_1|^2 = |h_2|^2 = -140 \) dB, which approximately corresponds to a distance of 500 meters (m). Fig. 2 shows the energy consumption tradeoff between the two users, which is obtained by exhausting the non-negative user weights \( \alpha_1 \) and \( \alpha_2 \) with \( \alpha_1 + \alpha_2 = 1 \). Note that the two users’ energy consumptions by the full-offloading-only scheme are much larger than those by the local-computing-only scheme, and thus cannot be shown in this figure. It is observed that the NOMA-based partial offloading generally achieves lower energy consumption at both users than the other two benchmark schemes. This validates the effectiveness of our proposed NOMA-based partial offloading in minimizing the users’ energy consumption while satisfying the computation latency constraints. Note that in the case with \( \alpha_1 = \alpha_2 \) under our setup, the OMA-based partial offloading is observed to achieve the same energy-saving performance as the NOMA-based counterpart.

Fig. 4. The average sum energy consumption at the users versus the duration \( T \) of the time block.
verage sum energy consumption than the local-computing-only and the full-offloading-only schemes. This shows the benefit of partial offloading by exploiting both the local computation resources at the users and the remote MEC resources at the BS. It is also observed that the NOMA-based partial offloading outperforms the OMA-based one, while the performance gain becomes more significant when \( L \) becomes large.

Fig. 4 shows the average sum energy consumption at these users versus the time block duration \( T \), where \( L = 2 \times 10^5 \). In general, we have similar observations as in Fig. 3. Particularly, the performance gain of the NOMA-based partial offloading over the OMA one is observed to become more considerable when \( T \) becomes small. By combining this with Fig. 3, it is evident that NOMA-based partial offloading becomes increasingly important in energy saving when the offloading resources become stringent.

V. CONCLUSION

This paper proposed to improve the performance of multiuser MEC system by exploring the full potential of multiuser computation offloading via multi-antenna NOMA. We developed an efficient design framework to minimize the weighted sum of the energy consumption at all users subject to their computation latency constraints, by jointly optimizing the users’ CPU frequencies for local computing, their transmit powers and rates for computation offloading, as well as the BS’s decoding order for SIC. Leveraging the Lagrange dual method, we obtained the globally optimal solution to this problem. Numerical results demonstrate the benefit of the proposed NOMA-based joint communication and computation design, as compared to benchmark schemes when all the users perform local computing only, full offloading only, or employ OMA-based computation offloading.

APPENDICES

A. Proof of Lemma 3.1

Given \( \lambda \), we solve subproblem (16) for each user \( k \in \mathcal{K} \). The objective of (16) is a convex function with respect to \( \ell_k \) and the constraint \( 0 \leq \ell_k \leq L_k \) is linear. Therefore, (16) is a convex problem and satisfies the Slater’s condition. The Lagrangian of the kth subproblem in (16) is given by

\[
\mathcal{L}_k = \frac{\alpha_k \zeta_k C_k^3}{T^2} \ell_k^2 - \lambda_k \ell_k - \beta_k \ell_k - \beta_k^* (L_k - \ell_k),
\]

where \( \beta_k \geq 0 \) and \( \beta_k^* \geq 0 \) are the Lagrangian multipliers associated with \( \ell_k \geq 0 \) and \( \ell_k \leq L_k \), respectively. Denote by \( (\beta_k^*, \beta_k^*) \) the optimal dual solution. Based on the KKT conditions [17], it follows that

\[
\begin{align*}
L_k &\geq \ell_k^* \geq 0, \\
\ell_k^* &\geq 0, \\
\beta_k^* &\geq 0, \\
\beta_k^* \ell_k^* &= 0, \\
3\alpha_k \zeta_k C_k^3 \ell_k^2 - \lambda_k - \beta_k^* + \beta_k^* &= 0,
\end{align*}
\]

where (24a) denotes the primal and dual feasibility conditions, (24b) collects the complementary slackness conditions, and the LHS term of (24c) is the first-order derivative of \( \mathcal{L}_k \) with respect to \( \ell_k \). Based on (24b), it follows that \( \beta_k^* = \beta_k^* = 0 \) for \( 0 < \ell_k^* < L_k \). Further from (24a) and (24c), we have

\[
\ell_k^* = \left( \frac{L_k}{3 \alpha_k \zeta_k C_k^3} \right) \frac{1}{\sqrt{\beta_k}},
\]

where \( [a, b] \) denotes the projection of \( x \) onto box \( [a, b] \). Since \( \sqrt{\frac{T \lambda_k}{3 \alpha_k \zeta_k C_k^3}} \geq 0 \), it follows that

\[
\ell_k^* = \min \left( \sqrt{\frac{T \lambda_k}{3 \alpha_k \zeta_k C_k^3}}, L_k \right), \quad \forall k \in \mathcal{K}.
\]

This lemma is thus verified.

REFERENCES

[1] M. Chiang and T. Zhang, “Fog and IoT: An overview of research opportunities,” IEEE Internet Thing J., vol. 3, no. 6, pp. 854–864, Jun. 2016.
[2] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, “Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks,” IEEE Signal Process. Mag., vol. 31, no. 6, pp. 45–55, Nov. 2014.
[3] P. Mach and Z. Becvar, “Mobile edge computing: A survey on architecture and computation offloading,” 2017. [Online]. Available: https://arxiv.org/abs/1702.05309
[4] Y. Mao, C. You, J. Zhuang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” 2017. [Online]. Available: https://arxiv.org/abs/1701.01090
[5] M.-H. Chen, M. Dong, and B. Liang, “Joint offloading decision and resource allocation for mobile cloud with computing access point,” in Proc. IEEE ICASSP, Shanghai, China, Mar. 2016, pp. 3516–3520.
[6] 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,” to appear in IEEE Trans. Wireless Commun., 2017. [Online]. Available: https://arxiv.org/abs/1702.00892
[7] 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.
[8] C. You, K. Huang, K. Chae, and B. Kim, “Energy-efficient resource allocation for mobile-edge computation offloading,” IEEE Trans. Wireless Commun., vol. 16, no. 3, pp. 1397–1411, Mar. 2017.
[9] F. Wang, J. Xu, X. Wang, and S. Cui, “Joint offloading and computing optimization in wireless powered mobile-edge computing system,” in Proc. IEEE ICC, Paris, France, May 2017, pp. 1–6.
[10] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, “Joint computation and communication cooperation for mobile edge computing,” 2017. [Online]. Available: https://arxiv.org/abs/1704.06777
[11] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, “Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends,” IEEE Commun. Mag., vol. 53, no. 9, pp. 74–81, Sep. 2015.
[12] Z. Ding, Y. Liu, J Choi, Q. Sun, M. Elkashlan, C.-L. I, and H. V. Poor, “Application of non-orthogonal multiple access in LTE and 5G networks,” IEEE Commun. Mag., vol. 55, no. 2, pp. 185–191, Feb. 2017.
[13] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, “A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends,” to appear in IEEE J. Sel. Areas Commun., 2017. [Online]. Available: https://arxiv.org/abs/1706.05347
[14] D. Tse and P. Viswanath, Fundamentals of Wireless Communication. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[15] D. Tse and S. Hanly, “Multi-access fading channels-Part I: Polymatroid structure, optimal resource allocation and throughput capacities,” IEEE Trans. Inf. Theory, vol. 22, no. 7, pp. 2796–2815, Nov. 1998.
[16] M. Mohseni, R. Zhang, and J. M. Cioffi, “Optimized transmission for fading multiple-access and broadcast channels with multiple antennas,” IEEE J. Sel. Areas Commun., vol. 24, no. 8, pp. 1627–1639, Aug. 2006.
[17] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[18] M. Grant, S. Boyd, and Y. Ye, “CVX: Matlab software for disciplined convex programming,” 2009. [Online]. Available: http://cvxr.com/cvx/
[19] S. Boyd, “Ellipsoid method,” Stanford University, California, USA. [Online]. Available: http://stanford.edu/class/ee364b/lectures/ellipsoid_method_notes.pdf