Latency Minimization in Intelligent Reflecting Surface Assisted D2D Offloading Systems

Yanzhen Liu, Qiyu Hu, Yunlong Cai, Senior Member, IEEE, and Markku Juntti, Fellow, IEEE

Abstract—In this letter, we investigate an intelligent reflecting surface (IRS) aided device-to-device (D2D) offloading system, where an IRS is employed to assist in computation offloading from a group of users with intensive tasks to another group of idle users. To minimize the system latency while cutting down the heavy overhead in exchange of channel state information (CSI), we study the joint design of beamforming and resource allocation on mixed timescales. Specifically, the high-dimensional passive beamforming vector at the IRS is updated in a frame-based manner based on the channel statistics, where each frame consists of a number of time slots, while the offloading ratio and user matching strategy are optimized relied on the low-dimensional real-time effective channel coefficients in each time slot. A novel mixed-integer stochastic successive convex approximation (MISSCA) algorithm is proposed to tackle the challenging problem. The convergence property and the computational complexity of the proposed algorithm are then examined. Simulation results show that our proposed algorithm significantly outperforms the conventional benchmarks.

Index Terms—Intelligent reflecting surface, D2D, computation offloading, latency minimization.

I. INTRODUCTION

Device-to-device (D2D) offloading is an effective technique to reduce the computation latency of intensive-task applications by assigning part of the tasks to nearby users via D2D links [1]–[3]. However, the potential of D2D computation offloading has not been fully exploited since the offloading link between the users is far from perfect [2]. Specifically, the D2D users that are far from each other typically suffer from a low offloading success rate due to the limited transmit power. Moreover, the communication links used for offloading tasks are very likely to be blocked by the obstructions, especially in the indoor environment. Therefore, it is necessary to improve the performance of D2D offloading systems from a communication perspective.

Fortunately, the propagation-induced impairments can be mitigated by the intelligent reflecting surface (IRS), which has been envisioned as an innovative hardware-efficient technology for the beyond fifth-generation (B5G) wireless system [4], [5]. IRS can modify the signal propagation by dynamically adjusting its reflection coefficients such that the desired and interfering signals can be enhanced and suppressed, respectively. As a result, by smartly coordinating these reflecting elements, IRS is able to create a favorable signal propagation environment to improve the wireless communication coverage, throughput, and energy efficiency substantially [6]–[10]. The authors of [6] developed a block coordinate descent (BCD) technique to minimize the latency for the IRS-aided mobile edge computing (MEC) system. In [7], resource allocation for IRS-assisted wireless-powered MEC has been further investigated. Moreover, the authors of [8]–[10] developed efficient algorithms via exploiting the channel statistics to reduce the system overhead in IRS-aided systems.

To the best of our knowledge, although the IRS-assisted computation offloading to the edge server has been investigated in [6], [7], the IRS-aided D2D offloading system has not been well studied. In this work, we investigate an IRS-aided D2D offloading system, where an IRS is employed to assist in computation offloading from a group of users with intensive tasks to another group of idle users. To minimize the system latency while cutting down the heavy overhead in exchange of channel state information (CSI), we study the joint design of beamforming and resource allocation on mixed timescales. Specifically, the high-dimensional passive beamforming vector at the IRS is updated in a frame-based manner based on the channel statistics, where each frame consists of a number of time slots, while the offloading ratio and user matching strategy are optimized relied on the low-dimensional real-time effective channel coefficients in each time slot. Note that the conventional stochastic successive convex approximation (SSCA) algorithm which only deals with smooth objective functions and continuous variables [8], [11] cannot effectively tackle the challenging problem with discrete variables. Therefore, in this work a novel mixed-integer stochastic successive convex approximation (MISSCA) algorithm is proposed, which is a significant extension of the method in [8], [11]. The convergence property and the computational complexity of the proposed algorithm are also examined. Simulation results show that our proposed algorithm significantly outperforms the conventional benchmarks.

Notations: Scalars, vectors, and matrices are denoted by lower case, boldface lower case, and boldface upper case letter, respectively. For a matrix $A$, $A^T$, $\text{conj}(A)$, and $A^H$ denote its transpose, conjugate, and conjugate transpose, respectively. The imaginary unit is denoted by $j$ and $|\cdot|$ denotes the absolute value of a complex scalar. $\odot$ is the Hadamard product and $\angle$ computes the phase of a complex scalar or a complex vector element-wise.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider an IRS-assisted D2D offloading system consisting of $I$ users which have intensive tasks to be processed, indexed by set $\mathcal{I} \triangleq \{1, 2, \ldots, I\}$, $J$ ($J \geq I$) idle users which are able to provide computing services, indexed by $\mathcal{J} \triangleq \{1 + 1, 1 + 2, \ldots, I + J\}$, and an IRS equipped with a controller and $M$ reflecting elements, as illustrated in Fig. 1. We assume partial offloading model [6], thus the users in $\mathcal{I}$ can divide their tasks into two parts. One portion is computed by the local CPU and the other portion is offloaded to a scheduled user in $\mathcal{J}$ and processed by the CPU of the helper.

A. Computation Model

We use a tuple $\{L_i, C_i\}$ to characterize the task of user $i$ in set $\mathcal{I}$, where $L_i$ (in bits) denotes the size of the task and $C_i$ (in CPU cycles/bit) denotes the CPU cycles required to compute
1-bit data of the task at user $i$. Moreover, we let $f_i$ and $f_j$ (in CPU cycles/s) denote the computing resource of user $i$ in set $\mathcal{I}$ and user $j$ in set $\mathcal{J}$, respectively. As mentioned above, we adopt the partial offloading strategy, i.e., user $i$ in set $\mathcal{I}$ offloads $\rho_{i,j}L_i$ bits to a matched D2D user $j$ in set $\mathcal{J}$, where $0 \leq \rho_{i,j} \leq 1$ denotes the offloading ratio, and the remaining $(1-\rho_{i,j})L_i$ bits are computed locally. Hence, the delay of the local computing can be expressed as
\begin{equation}
T_{L_{i,j}}^L = \frac{(1-\rho_{i,j})C_iL_i}{f_i}.
\end{equation}
By neglecting the feedback delay for the length of computation results is generally short [6], [7], the D2D offloading can be expressed as the summation of the transmission delay and computation delay:
\begin{equation}
T_{i,j}^D = \frac{\rho_{i,j}L_i}{r_{i,j}} + \frac{\rho_{i,j}C_iL_i}{f_j},
\end{equation}
where $r_{i,j}$ is the communication rate between user $i$ and user $j$. Then, we express the total delay for the task processing of user $i$ as
\begin{equation}
T_{i,j} = \max\{T_{i,j}^L, T_{i,j}^D\}.
\end{equation}
Without loss of generality, we assume that the users in $\mathcal{I}$ only can offload tasks to one user in set $\mathcal{J}$. Moreover, we define a binary decision variable $u_{i,j}$ to indicate the status of the D2D link established between the two groups of users, i.e., $u_{i,j} = 1$ when user $i$ establishes a D2D link with user $j$ and $u_{i,j} = 0$ otherwise.

### B. Communication Model

For the sake of exposition, we assume that the users are all equipped with a single antenna. For a given D2D pair of user $i$ in $\mathcal{I}$ and user $j$ in $\mathcal{J}$, we let $h_{i,j} \in \mathbb{C}$ denote the channel coefficient between user $i$ and user $j$, $g_i \in \mathbb{C}^{M \times 1}$ denote the channel vector between user $i$ and the IRS, and $g_j \in \mathbb{C}^{M \times 1}$ denote the channel vector between user $j$ and the IRS. Moreover, we let $s_i \sim \mathcal{CN}(0,1)$ denote the transmit symbol of user $i$. Then, the received signal at user $j$ is given by
\begin{equation}
y_j = \sum_{i=1}^{I} \sqrt{p_i}(h_{i,j} + g_j^H \Phi g_i)s_i + n_j,
\end{equation}
where $p_i$ denotes the transmit power of user $i$ and $n_j \sim \mathcal{CN}(0, \sigma_j^2)$ denotes the additive white Gaussian noise with zero mean and variance $\sigma_j^2$. $\Phi \in \mathbb{C}^{M \times M}$ denotes the passive beamforming matrix at the IRS, which is a diagonal matrix due to no signal processing over its passive reflecting elements. We define the passive beamforming vector $\phi \in \mathbb{C}^{M \times 1}$ as
\begin{equation}
\text{diag}\{\phi\}. \quad \text{Then, the maximum achievable transmission rate between user $i$ and user $j$ can be obtained as}
\end{equation}
\begin{equation}
r_{i,j} = \frac{B}{1 + \sum_{i' \neq i} p_{i'}[h_{i',j} + g_j^H \phi]^2 + \sigma_j^2}.
\end{equation}
A. Short-term Offloading Ratio and User Matching Strategy

With fixed long-term IRS passive beamforming vector \( \phi^t \) in frame \( t \), for given real-time effective channel coefficients \( \{ h_{i,j}^t, \forall i,j \} \) in time slot \( k \), the short-term optimization problem of the offloading ratio and user matching strategy design yields
\[
\min_{\rho_{i,j} \leq 1} \sum_{i,j=1}^{I+J} u_{i,j} w_{i,j} T_{i,j} \quad \text{s.t.} \quad (6b),(6c). \tag{7a}
\]
Note that \( \rho_{i,j} \) is not coupled with \( u_{i,j} \) in the constraints. Hence we can first optimize the offloading ratio to minimize \( T_{i,j} \), which provides the following \( I \times J \) parallel subproblems
\[
\min_{\rho_{i,j} \leq 1} \max \{ T_{i,j}^L, T_{i,j}^D \} \quad \forall i,j. \tag{8}
\]
Since both \( T_{i,j}^L \) and \( T_{i,j}^D \) are linear functions of \( \rho_{i,j} \), it is readily seen that the optimal \( \rho_{i,j}^* \) should satisfy \( T_{i,j}^* = T_{i,j}^D \), and we obtain
\[
\rho_{i,j}^* = \frac{C_{1,f} r_{i,j} + f_i + f_j}{C_{1,f} (f_i + f_j) r_{i,j} + f_i + f_j}. \tag{9}
\]
By substituting (9) into (3), we obtain
\[
T_{i,j}^* = \frac{C_{1,L_i} f_i + C_{1,L_j} f_j^2}{f_i + f_j} + \frac{C_{1,L_i} f_i^2}{f_i + f_j}. \tag{10}
\]

B. Long-term Passive Beamforming Design

The long-term optimization problem of the passive beamforming vector \( \phi \) is given by
\[
(P2) \min_{\theta} f(\theta, x^*) = \mathbb{E}_H \{ (g(\theta, x^*) \}
\]
where \( \theta \triangleq \mathcal{H} \phi, x^* \triangleq \{ \rho_{i,j}^*, u_{i,j}^* \} \) denoting the collection of the optimal short-term variables and
\[
g(\theta, x^*) \triangleq \sum_{i=1}^J w_i T_{i,m_i}^*(\theta), \tag{12}
\]
where we define \( \pi_i \triangleq \sum_{j=1}^{I+J} u_{i,j}^* \) for denoting the user in \( \mathcal{J} \) which is matched with user \( i \).

\( P2 \) is difficult to handle due to the highly non-convex objective function. Hence we seek to approximate (11) by using a quadratic surrogate function [11]. Specifically, at the end of each frame, the channel samples \( \{ (g_{i,m_i}^t, h_{i,m_i}^t), \forall i \} \) are obtained and the surrogate objective function is updated based on the CSI samples and the short-term variables \( x^* \) in time slot \( (t+1)A_s \) as
\[
\hat{f}^t(\theta) = (f^t)^T(\theta - \theta^t) + \varpi \| \theta - \theta^t \|^2, \tag{13}
\]
where \( \theta^t \) is the current value of \( \theta \), \( \varpi > 0 \) is a constant, and \( f^t \) denotes the approximation of the partial derivatives \( \frac{\partial f}{\partial \theta} \), which is updated based on the following expression
\[
f^t = (1 - \varphi(t)) f^{t-1} + \varphi(t) \frac{\partial g}{\partial \theta}(\theta^t, x^*), \tag{14}
\]
where \( \{ \varphi(t) \} \) is a sequence to be properly chosen and the expression of \( \frac{\partial g}{\partial \theta} \) is omitted here due to space limitation. In practice, its value can be obtained via the “PyTorch” platform.

Algorithm 1 Proposed two-timescale MISSCA algorithm

1: Initialize the long-term variable \( \theta^0 \). Set \( f^{-1} = 0, f^{-1} = 0, t = 0, \) and \( k = 0 \). Choose proper sequences \( \{ \varphi(t) \}, \{ \gamma^t \} \) and set a proper value for \( \varpi \).
2: repeat
3: Obtain the effective CSI \( \{ \tilde{h}_{i,j}^t, \forall i,j \} \) for time slot \( k \).
4: Compute \( \rho_{i,j}^t, \forall i,j \) and \( T_{i,j}^L, T_{i,j}^D \) based on (9) and (10), respectively. Obtain the optimal matching variable \( u_{i,j}^t \) via the KM algorithm.
5: Update \( k = k + 1 \).
6: until the frame ends, i.e., \( k = (t+1)A_s \).
7: Obtain the CSI samples \( \{ (g_{i,m_i}^t, h_{i,m_i}^t), \forall i \} \). Compute the surrogate function (13) based on (14).
8: Obtain the optimal solution via (17) and update \( \theta^t \) based on (18).
9: Update \( t = t + 1 \) and return to step 2.

Subsequently, we aim to solve the approximated problem at time frame \( t \), which is given by
\[
\min_{\theta} \hat{f}^t(\theta). \tag{16}
\]
This is a convex quadratic problem and the solution can be readily derived as
\[
\hat{\theta}^t = \theta^t - \frac{f^t}{2\varpi}. \tag{17}
\]
Then, the long-term variable is updated as
\[
\theta^{t+1} = (1 - \gamma^t) \theta^t + \gamma^t \hat{\theta}^t, \tag{18}
\]
where similarly \( \{ \gamma^t \} \) denotes a sequence of parameters and the convergence can be guaranteed if we choose \( \varphi(t) \) and \( \gamma^t \) by following the conditions \( \lim_{t \to \infty} \gamma^t = 0, \sum_{t=0}^\infty \gamma^t = \infty, \sum_{t=0}^\infty (\varphi(t))^2 < \infty, \lim_{t \to \infty} \gamma^t = 0, \sum_{t=0}^\infty \gamma^t = \infty, \sum_{t=0}^\infty (\gamma^t)^2 < \infty \) and \( \lim_{t \to \infty} \frac{\gamma^t}{\varpi} = 0 \). The convergence proof is given in Appendix A.

C. Practical Implementation

The proposed MISSCA algorithm is summarized in Algorithm 1 and the computational complexity is given by \( O(I^2 J + \frac{1}{\lambda} IM) \). In practice, this algorithm can be implemented on a user in set \( \mathcal{J} \), which is referred to as master user. Specifically, for each time slot, each user \( j \) in set \( \mathcal{J} \) estimates the effective channel coefficients \( \{ h_{i,j}, \forall i \} \) and feeds them to the master user which performs the proposed algorithm to obtain the short-timescale variables, i.e., the offloading ratios and user matching variables. The offloading ratios are fed back to the users in set \( \mathcal{I} \) through signaling channels, and the matched users in set \( \mathcal{J} \) are notified to provide services. At the end of each frame, the full channel samples \( \{ (g_{i,m_i}, h_{i,m_i}), \forall i \} \) are estimated at these scheduled users in set \( \mathcal{J} \) using the IRS related channel estimation algorithms [13] and then sent to the master user. Finally, the long-timescale IRS passive beamforming vector is updated based on the proposed algorithm. As a result, we can obtain that the required CSI overhead of the proposed two-timescale algorithm in a frame is \( A_s(IJ + IM) \), while that of the single-timescale algorithm is \( A_s(IJ + IM) \), which is significantly reduced.

IV. Simulation Results

In this section, we present simulation results to verify the effectiveness of our proposed two-timescale algorithm. A top view of the simulation setup is given in Fig. 3. The users in \( \mathcal{I} \) and \( \mathcal{J} \) are randomly located in two circles with radius of \( R_1 = \)
10 m and $R_2 = 10$ m, respectively. The coordinates of the centers for these two circles are $(-5 \text{ m}, 0 \text{ m})$ and $(5 \text{ m}, 0 \text{ m})$, respectively, and the IRS is located at $(0 \text{ m}, y_1 = 0 \text{ m})$. The heights of the users are all set to 1 m and the height of the IRS is 3 m. Unless otherwise specified, we consider $T = 4$, $J = 6$, $M = 32$, $A_s = 100$, $A_f = 100$, $w_i = 1$, $\forall i, f_i = 1 \times 10^{10}$ CPU cycles/s, $\forall i, p_i = 24 $ dBm, $\forall i, C_i = 1 \times 10^3$ CPU cycles/bit [14], $\forall i, B = 15$ MHz, and the noise spectral density as $-174$ dBm/Hz. Furthermore, the length of the computation tasks $L_i$ is uniformly distributed from 0.1 Mbits to 0.5 Mbits and the computation resource of user $j$ follows the uniform distribution within $f_j \in [1 \times 10^{10}, 5 \times 10^{10}]$ CPU cycles/s. We adopt the Rician channel with a Rician factor $\beta = 20$ and the path loss is modeled as $P_{LS} = C_0 (\frac{d_{link}}{D_0})^{-\alpha}$, where $C_0$ is the path loss at the reference distance $D_0 = 1$ m and is set to $C_0 = -30$ dB, $d_{link}$ is the link distance, and $\alpha$ is the path loss exponent where we set it for the link between the users as $\alpha_u = 3.8$ and the link between the user and IRS as $\alpha_{ul} = 2.2$ [6], respectively.

In this work, we consider the following algorithms:

- **Proposed TTS**: We employ Algorithm 1 for designing the offloading ratio, the matching strategy, and the IRS passive beamforming vector.
- **STS**: We employ a single-timescale scheme which optimizes the offloading ratio and the matching strategy, as well as the passive beamforming vector, based on the real-time high-dimensional full CSI in each time slot.
- **Max-to-max TTS**: We employ a heuristic user matching scheme which repeatedly matches the user in set $I$ that has the maximum task size $L_i^{\text{max}}$ to the user in set $J$ that has the strongest computation capacity $f_j^{\text{max}}$ and removes the matched users from sets $I$ and $J$. The offloading ratio and the passive beamforming vector are optimized based on the corresponding steps in Algorithm 1.
- **Random IRS**: We employ a scheme where the reflection coefficients of the IRS are randomly generated. Then, the offloading ratio is designed based on (9) and the matching strategy is optimized based on the KM algorithm.
- **No IRS**: We do not employ the IRS in the system. The offloading ratio and the matching strategy are designed based on the random IRS method.

Fig. 3 shows the convergence performance of Algorithm 1. As we can see, the system delay converges quickly within 20 iterations. Fig. 4 compares the STS scheme and the proposed TTS scheme in terms of required CSI overhead in a frame, where we assume that the number of quantization bits for each element of the CSI vector equals 8. We can conclude that our proposed TTS algorithm can significantly reduce the CSI overhead compared to the STS scheme. Hence the proposed algorithm is much more suitable for practical design.

Fig. 6 shows the weighted sum system delay of different schemes versus the number of reflecting elements of the IRS. As we can see, the proposed TTS algorithm significantly outperforms the schemes of no IRS, random IRS, and max-to-max TTS. In addition, we observe a small gap between our proposed TTS design and the STS scheme, especially when the number of reflecting elements is small. This is because the line-of-sight (LoS) component generally dominates the IRS-related links, thus rendering valid designs based on the channel statistics. Moreover, it is observed that the delay of the random IRS scheme increases with $M$, since a large number of randomly generated reflecting coefficients will lead to strong interference. On the contrary, the latency of our proposed TTS scheme decreases quickly with $M$, for the proposed algorithm can efficiently mitigate the interference and enhance the transmission rate by properly designing the reflecting coefficients.

Fig. 7 presents the system delay of different algorithms when the location of the IRS is moving along the $y$ axis. It is readily seen that when the IRS is far away from the users, the delay of our proposed algorithm approaches that of no IRS. When the IRS is located at the central location, the best latency performance is achieved. However, for the random IRS scheme, it is better to place the IRS about 10 m or 15 m far from the center, since this distance achieves a good balance between enhancing the desired signal and increasing the interference.

Fig. 8 indicates the latency performance of various schemes versus the number of available users $J$. We can see that our proposed TTS algorithm almost achieves the performance of the STS scheme and significantly outperforms the others. We also observe that the delay of the proposed TTS scheme decreases rapidly when $J$ increases, while the latency of the heuristic max-to-max strategy is not sensitive to $J$. This is because the proposed TTS scheme jointly allocates the communication and computation resources and it can find the optimal matching solution, while the heuristic max-to-max scheme only allocates the computation resources. It is possible that the matched user has a bad communication link and the system latency will not be significantly reduced.
V. CONCLUSION

In this letter, we investigated an IRS-aided D2D offloading system. To minimize the system latency while cutting down the heavy overhead in exchange of CSI, we studied the joint design of beamforming and resource allocation on mixed timescales. A novel MISSCA algorithm has been proposed to tackle the challenging problem. The convergence property and the computational complexity have also been examined. Simulation results show that our proposed algorithm significantly outperforms the conventional benchmarks.

APPENDIX A

CONVERGENCE PROOF OF ALGORITHM 1

Let $x^\ast(\theta, H)$ denote the optimal short-term variables under input $\theta$ and $H$. Then, the proof relies on the following lemma.

**Lemma 1.** We have

$$
\|x^\ast(\theta_1, H) - x^\ast(\theta_2, H)\| \leq B_x \sqrt{\|\theta_1 - \theta_2\|^2}, \text{ w.p.1,} \quad (19)
$$

$$
\|f(\theta_1, x^\ast(\theta_1, H)) - f(\theta_2, x^\ast(\theta_2, H))\| \leq B_f \sqrt{\|\theta_1 - \theta_2\|^2}, \quad (20)
$$

for any $\theta_1, \theta_2$ and some constant $B_x > 0, B_f > 0$, where w.p.1 is the abbreviation of with probability one.

**Proof:** There are totally $2^{|J|}$ kinds of situations for the user matching strategy and we define $x^{m_1}(\theta, H) \triangleq \{p^\ast_1, x_1^\ast, \ldots, x_{m_1}^\ast\}$ to characterize one of the matching strategy. Then, we have

$$
g(\theta, x^{m_1}(\theta, H)) \neq g(\theta, x^{m_2}(\theta, H)), \forall m_1 \neq m_2, \text{ w.p.1.} \quad (21)
$$

By considering that the probability density function of the joint distribution of $g(\theta, x^{m_1}(\theta, H)) \times g(\theta, x^{m_2}(\theta, H))$ is bounded and the integral along the line $g(\theta, x^{m_1}(\theta, H)) = g(\theta, x^{m_2}(\theta, H))$ equals zeros (because the measure of this line is 0), (21) can be easily verified. We then prove that

$$
\lim_{\Delta\theta \to 0} \|x^\ast(\theta, H) - x^\ast(\theta + \Delta\theta, H)\| = 0, \text{ w.p.1.} \quad (22)
$$

Based on (21), we have

$$
g(\theta, x^\ast(\theta, H)) < g(\theta, x^{m}(\theta, H)), \text{ w.p.1,} \forall x^{m}(\theta, H) \neq x^\ast(\theta, H). \quad (23)
$$

Moreover, since $g(\theta, x^{m}(\theta, H)), \forall m$ are Lipschitz continuous functions with respect to $\theta$. We have

$$
\lim_{\Delta\theta \to 0} g(\theta + \Delta\theta, x^\ast(\theta + \Delta\theta, H)) < \lim_{\Delta\theta \to 0} g(\theta + \Delta\theta, x^\ast(\theta + \Delta\theta, H)), \text{ w.p.1,} \forall x^\ast(\theta, H) \neq x^\ast(\theta, H). \quad (24)
$$

As a result, we have

$$
\lim_{\Delta\theta \to 0} x^\ast(\theta + \Delta\theta, H) = x^\ast(\theta, H), \text{ w.p.1,} \quad (25)
$$

which proves (19). Moreover, since $f(\theta, x^\ast(\theta, H))$ can be written as $E_H \{\min_m g(\theta, x^{m}(\theta, H))\}$ and $g(\theta, x^{m}(\theta, H)), \forall m$ are Lipschitz continuous, (20) holds immediately. Then, based on the Lemma 1 in [15], it can be verified that the proposed algorithm converges to a stationary point of $P$, w.p.1.

REFERENCES

[1] L. Pu, X. Chen, J. Xu, and X. Fu, “D2D fogging: An energy-efficient and incentive-aware task offloading framework via network-assisted D2D collaborations,” IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3887–3901, Dec. 2016.

[2] W. Hu and G. Cao, “Quality-aware traffic offloading in wireless networks,” IEEE Trans. Mobile Comput., vol. 16, no. 11, pp. 3182–3195, Nov. 2017.

[3] Y. He, J. Ren, G. Yu, and Y. Cai, “D2D communications meet mobile edge computing for enhanced computation capacity in cellular networks,” IEEE Trans. Wireless Commun., vol. 18, no. 3, pp. 1750–1763, Mar. 2019.

[4] Q. Wu and R. Zhang, “Towards smart and reconﬁgurable environment: Intelligent reﬂecting surface aided wireless network,” IEEE Commun. Mag., vol. 58, no. 1, pp. 106–112, Jun. 2020.

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

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

[7] T. Bai, C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, “Resource allocation for intelligent reﬂecting surface aided wired mobile edge computing in OFDM systems,” IEEE Trans. Wireless Commun., to appear.

[8] M. M. Zhao, Q. Wu, M. J. Zhao, and R. Zhang, “Intelligent reﬂecting surface enhanced wireless networks: Two-timescale beamforming optimization,” IEEE Trans. Wireless Commun., vol. 20, no. 1, pp. 2-17, Jan. 2021.

[9] Z. Peng, T. Li, C. Pan, H. Ren, W. Xu, and M. D. Renzo, “Analysis and optimization for RIS-aided multi-pair communications relying on statistical CSI,” IEEE Trans. Veh. Technol., vol. 70, no. 4, pp. 3897–3901, Apr. 2021.

[10] K. Zhi, C. Pan, H. Ren, and K. Wang, “Statistical CSI-based design for reconﬁgurable intelligent surface-aided massive MIMO systems with direct links,” IEEE Wireless Commun. Lett., vol. 10, no. 5, pp. 1128–1132, May 2021.

[11] 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.

[12] F. Bourgeois and J.-C. Lassalle, “An extension of the Munkres algorithm for the assignment problem to rectangular matrices,” Commun. ACM, vol. 14, no. 12, pp. 802–804, Dec. 1971.

[13] B. Zheng, C. You, and B. Zhang, “Fast channel estimation for IRS-assisted OFDM,” IEEE Wireless Commun. Lett., to appear.

[14] C. Liu, M. Benni, M. Debah, and H. V. Poor, “Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing,” IEEE Trans. Commun., vol. 67, no. 6, pp. 4132–4150, June 2019.

[15] A. Ruszczynski, “Feasible direction methods for stochastic programming problems,” Math. Program., vol. 19, no. 1, pp. 220–229, Dec. 1980.