Joint Mode Selection and Transceiver Design for Device-to-Device Communications Underlaying Multi-User MIMO Cellular Networks

Jingran Lin, Qingjiang Shi, Qiang Li, and Dongmei Zhao

***Abstract*—Consider a network consisting of one multi-antenna base station (BS) and multiple pairs of multi-antenna user equipments (UEs). For each UE pair, the communication between transmitter and receiver is established either through BS or via device-to-device (D2D) link. We assume that D2D transmission and cellular transmission are equally prioritized and share the same resources. To improve the network throughput, we maximize the sum rate by jointly optimizing the transmission mode of each UE pair and the associated transceivers. Due to the NP-hardness of this problem, we first perform some efficient approximation to it and then design an iterative algorithm, which is guaranteed to converge to a stationary solution by solving a series of weighted minimum mean square error (WMMSE) problems. The proposed algorithm has two distinguishing features. First, it only solves the WMMSE problem inexactly in each iteration, which thereby has a simplified algorithm structure and accelerated convergence behavior than the classical one. Second, we further fit the WMMSE problem into the alternating direction method of multipliers (ADMM) framework, making it amenable to parallel and distributed computation. Finally, the approximated problem is solved efficiently and distributively, with simple closed-form solutions in each step.**

***Index Terms*—Device-to-device (D2D) communication, mode selection, transceiver design, weighted minimum mean square error (WMMSE), alternating direction method of multipliers (ADMM).**

I. Introduction

Device-to-device (D2D) communication, which allows two user equipments (UEs) to exchange information directly without the intervention of a base station (BS), has been widely accepted as an important key technology in the fifth-generation (5G) mobile networks [1], [2]. By exploiting the proximity gain, reuse gain, hop gain, and pairing gain, D2D can remarkably improve the network performance in terms of spectral and energy efficiencies [3]–[5]. However, to achieve these potential benefits, D2D communications should be appropriately managed. Many important yet challenging issues, such as D2D discovery, D2D synchronization, transmission mode selection, wireless resource allocation, power control, interference mitigation, etc., need to be carefully addressed [6], [7]. In this paper, we focus on the joint transmission mode selection and interference mitigation problem for the D2D communication underlaying cellular networks. Although such a problem has been intensively investigated, most current studies impose some constraints onto the setting of D2D communications, e.g., the UE status [4], [8], the UE/antenna number [8], [9], the resource sharing strategy [7], the priority of D2D transmission [4], etc. As a consequence, the D2D performance gains are not fully utilized. Naturally, to maximally explore the D2D potentials, a more general and flexible formulation of the joint mode selection and interference mitigation problem should be considered.

A. State of the Art

So far, various schemes have been proposed for joint mode selection and interference mitigation in D2D communications. Basically, the literature can be categorized according to the network size and the way of algorithm execution (centralized or distributed).

1) Network Size: Most early studies formulated this problem in single-antenna networks which consist of only one cellular UE and one pair of D2D UEs [8]–[10]. They usually selected the transmission mode (e.g., cellular mode or underlay/overlay D2D mode) for the D2D UE pair by exhaustive search, and then optimized the power allocation to achieve the best performance in throughput and/or power efficiency. Recently, to accommodate the explosive demand for wireless data, the network size (e.g., the UE number and/or the antenna number) has been increasing sharply and inevitably. The approach of exhaustive search is thereby computationally prohibitive, and some advanced techniques, such as game theory [11], graph theory [12], and nonlinear optimization [13], have been introduced to help the mode selection and interference mitigation. Specifically, as far as the multi-antenna D2D communication is concerned, beamforming is widely utilized to manage the interference. For instance, many studies applied zero-forcing (ZF) beamforming to avoid the interference between cellular UEs and D2D UEs [14]–[16]. A more general beamforming scheme was addressed in [17], where the transceivers were designed to maximize the entire D2D and cellular transmission rates, and the sequential quadratic programming method was

J. Lin and Q. Li are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: jingranlin@uestc.edu.cn, and lq@uestc.edu.cn). Q. Shi is with the School of Software Engineering, Tongji University, Shanghai 200092, China (e-mail: qing.j.shi@gmail.com). D. Zhao is with the ZTE Corporation, Shenzhen 518063, China (e-mail: zdmzzp@sina.com).

The work of J. Lin was supported in part by the natural Science Foundation of China (NSFC) under Grant 61671120, and in part by the Sichuan Science and Technology Program under Grant 2018JY0147. The work of Q. Shi was supported in part by NSFC under Grant 61671411, in part by National Key Research and Development Project under grant 2017YFE0119300, and in part by the Fundamental Research Funds for the Central Universities under grant 22120180113. The work of Q. Li was supported in part by NSFC under Grant 61531009.

Part of this work has been reported in IEEE ICASSP 2016 [1].

adopted to solve this problem. This approach outperforms the ZF approach, but does not involve the issue of mode selection. The authors in [18] considered the joint mode selection and transceiver design for rate maximization in multi-input multi-output (MIMO) D2D networks, and then solved the problem by applying the approach of successive convex approximation (SCA). In this work, however, each UE pair may perform cellular and D2D transmissions simultaneously, thus rendering heavy operational burdens. Obviously, a more practical mode selection strategy should be exclusive, i.e., the communication within each UE pair is established either through base station (BS) or via D2D link.

2) Algorithm Execution: In addition to the centralized algorithms, many distributed algorithms have been proposed since it may not be easy to solve the high-dimensional D2D communication problems as the network size increases. For instance, several heuristic algorithms were developed to distributively solve the joint mode selection and power control problem [19], [20]. In [21], some energy-splitting variables were introduced such that the mode selection and the resource allocation could be decoupled and optimized independently. Another popular approach worth mentioning is game theory, which has been successfully applied to tackle many distributed selfish optimization problems about joint mode selection and resource allocation in D2D communications [22], [23].

In summary, despite the differences in problem formulation and algorithm design, most current studies divided the UEs into cellular UEs and D2D UEs in advance, and fixed their statuses during the entire communication process. Moreover, they usually took cellular UEs as the primary users, and thus their quality of service (QoS) requests were delivered with priority. In this circumstance, one important task of D2D management was to limit the interference from D2D UEs to cellular UEs. To guarantee this, cellular UEs were set to communicate with BS constantly, while only D2D UEs were involved in mode selection. In addition, to better protect cellular transmissions from D2D interference in underlay D2D communications, D2D UEs usually reused the cellular resources in uplink only, while keeping idle in downlink [7], [24], [25].

Obviously, this has significantly affected the flexibility of D2D management. Thus, the performance gains of D2D are not fully utilized. Actually, there are still continuing debates in the Third Generation Partnership Project (3GPP) on the mode selection between cellular and D2D communications and its priority, although a lot of progress has been made in the standardization of D2D discovery, channel models, deployment scenarios, evaluation methodologies, and so on [24]–[26]. In this paper, we relax these constraints and formulate the D2D communication problem from a more general perspective. A key point is that we equally prioritize the cellular transmission and the D2D transmission. Let us explain the reasonability of this setting. The idea that the D2D UEs exchange data by reusing the cellular resources is quite similar to the concept of the secondary user (SU) introduced in cognitive ratio (CR) networks [5]. This is one possible reason why a lot of studies have prioritized cellular UEs over D2D UEs, and imposed some constraints on D2D communications. However, there are some essential differences between D2D and CR. In particular, the D2D connection utilizes licensed spectrum bands and is supervised by a central entity (e.g., the cellular BS), whereas in CR the SU is not controlled by the primary user (PU) networks. In brief, the involvement of the cellular network in the control plane is the key difference between D2D and CR [2], [5], [25]. Since D2D communications are fully controlled by the cellular network, to maximally utilize the performance gains of D2D, it may be more reasonable to assume that cellular UEs and D2D UEs have the same priority. A similar idea has been mentioned in [4], [7], [19], but no details have been provided on how to fulfill it.

B. Contributions and Organization

In this paper, we consider the D2D communication underlaying multi-user MIMO networks. By jointly optimizing the transmission mode and the associated transceiver for each UE pair, we aim to maximize the network throughput. Departing from most current studies, we formulate this problem from a more general perspective. Specifically, we prioritize the cellular transmission and D2D transmission equally. They share the same resources in both uplink and downlink. The status of each UE pair (cellular or D2D) is not specified in advance, and all the UE pairs are involved in mode selection. The transmission mode of each UE pair can freely switch between cellular and D2D, depending on its contribution to the total throughput and its interference to other UE pairs. This setting endows more flexibilities to D2D management, but yields a challenging NP-hard problem [27]. As a compromise, we pursue some efficient approximate solutions with manageable complexity.

Specifically, by performing the weighted minimum mean square error (WMMSE) reformulation [28], [29], we develop an algorithm which handles this problem by iteratively solving a series of WMMSE problems. In particular, the WMMSE problem in each iteration is only solved inexactly to simplify the algorithm structure and accelerate the convergence. For this reason, the algorithm is referred to as the iterative algorithm based on inexact WMMSE (IA-IWMMSE). We show that IA-IWMMSE is guaranteed to converge to a stationary solution. Next, to facilitate the algorithm execution, we further fit the WMMSE problem into the framework of alternating direction method of multipliers (ADMM) [30], so that IA-IWMMSE can be executed in a distributed manner. Finally, the problem can be solved distributively and efficiently, with a simple closed-form solution in each step.

The rest of this paper is organized as follows. The system model and problem formulation are given in Section II. In Section III, we propose IA-IWMMSE to solve the problem and show its convergence. In Section IV, we next show how to distributively solve the WMMSE problem in IA-WMMSE, with the help of ADMM. Simulation results are provided in Section V. Section VI concludes this paper.

Notations: Boldface capital and little letters denote matrices and vectors, respectively. Italic letters denote scalars. For a given matrix $\mathbf{X}$, we denote its transpose, Hermitian, and inverse by $\mathbf{X}^T$, $\mathbf{X}^\dagger$, and $\mathbf{X}^{-1}$, respectively. Similarly, we denote the transpose and Hermitian of vector $\mathbf{x}$ by $\mathbf{x}^T$ and $\mathbf{x}^\dagger$. We use $\|\cdot\|_p$ to denote the l_p-norm, $p = 0, 1, 2$. Little subscripts "m"

TABLE I
SUMMARY OF THE VARIABLES USED IN THIS PAPER

Uplink (UL)	$\mathbf{G}_m^{\mathrm{U}}$	Cellular UL channel between TX_UE$_m$ and BS
	$\mathbf{F}_{m,n}^{\mathrm{U}}$	D2D UL channel between TX_UE$_n$ and RX_UE$_m$
	$\mathbf{v}_m^{\mathrm{U}}$	UL TX beamformer of TX_UE$_m$
	$\mathbf{u}_{d,m}^{\mathrm{U}}$	D2D UL RX beamformer of RX_UE$_m$ for TX_UE$_m$
	$\mathbf{u}_{c,m}^{\mathrm{U}}$	Cellular UL RX beamformer of BS for TX_UE$_m$
Downlink (DL)	$\mathbf{G}_m^{\mathrm{D}}$	Cellular DL channel between BS and RX_UE$_m$
	$\mathbf{F}_{m,n}^{\mathrm{D}}$	D2D DL channel between TX_UE$_n$ and RX_UE$_m$
	$\mathbf{v}_{d,m}^{\mathrm{D}}$	D2D DL TX beamformer of TX_UE$_m$ for RX_UE$_m$
	$\mathbf{v}_{c,m}^{\mathrm{D}}$	Cellular DL TX beamformer of BS for RX_UE$_m$
	$\mathbf{u}_{d,m}^{\mathrm{D}}$	D2D DL RX beamformer of RX_UE$_m$ for TX_UE$_m$
	$\mathbf{u}_{c,m}^{\mathrm{D}}$	Cellular DL RX beamformer of RX_UE$_m$ for BS

and "n" denote the UE index; little subscripts "c" and "d" denote cellular mode and D2D mode; capital superscripts "U" and "D" denote uplink and downlink. For instance, $\mathbf{x}_{d,m}^{\mathrm{U}}$ is some vector variable associated with UE m in the uplink D2D transmission, while $\mathbf{x}_{c,n}^{\mathrm{D}}$ is some vector variable associated with UE n in the downlink cellular transmission.

II. System Model and Problem Formulation

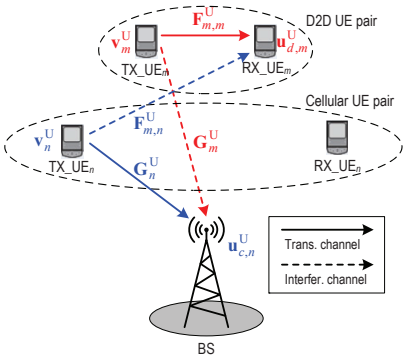
A. System Model

As shown in Fig. 1, we consider a network that consists of one BS and a set $\mathcal{M} = \{1, 2, \cdots, M\}$ of UE pairs, where each UE pair includes a transmitter UE (TX_UE) and a receiver UE (RX_UE). The BS is equipped with N_b antennas, while each UE with N_u antennas. By employing the frequency division duplexing (FDD) strategy, each UE pair can communicate either in cellular mode or in D2D mode. In cellular mode, the BS adopts the decode-and-forward strategy to transmit data from one TX_UE to its intended RX_UE through orthogonal uplink and downlink frequency bands. In D2D mode, one TX_UE sends data to its intended RX_UE directly by reusing the cellular resources in both uplink and downlink. We assume the interfering broadcast channel (IBC) model [28], i.e., each transmitter generates interference to all the other receivers.

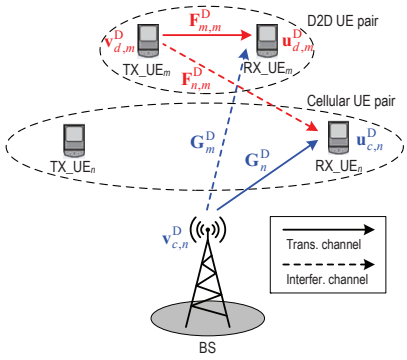
The channel and transceiver variables used in uplink and downlink are summarized in Table I. Without loss of generality, let us assume in Fig.1 that the mth UE pair (the transmitter and receiver are denoted as TX_UE$_m$ and RX_UE$_m$) selects the D2D mode, while the nth UE pair (the transmitter and receiver are denoted as TX_UE$_n$ and RX_UE$_n$) selects the cellular mode. Therefore, the system depicted in Fig. 1 works as follows.

The uplink transmissions are shown in Fig. 1(a). In the uplink band, TX_UE$_m$ sends data to RX_UE$_m$ directly through the uplink D2D channel $\mathbf{F}_{m,m}^{\mathrm{U}} \in \mathbb{C}^{N_u \times N_u}$, using the TX beamformer $\mathbf{v}_m^{\mathrm{U}} \in \mathbb{C}^{N_u \times 1}$, while TX_UE$_n$ sends data to BS via the uplink cellular channel $\mathbf{G}_n^{\mathrm{U}} \in \mathbb{C}^{N_b \times N_u}$, using the TX beamformer $\mathbf{v}_n^{\mathrm{U}} \in \mathbb{C}^{N_u \times 1}$. Then, RX_UE$_m$ decodes the data from TX_UE$_m$, using the RX beamformer $\mathbf{u}_{d,m}^{\mathrm{U}} \in \mathbb{C}^{N_u \times 1}$, while BS decodes the data from TX_UE$_n$, using the RX beamformer $\mathbf{u}_{c,n}^{\mathrm{U}} \in \mathbb{C}^{N_b \times 1}$.

The downlink transmissions are shown in Fig. 1(b). In the downlink band, TX_UE$_m$ sends data to RX_UE$_m$ directly through the downlink D2D channel $\mathbf{F}_{m,m}^{\mathrm{D}} \in \mathbb{C}^{N_u \times N_u}$, using the TX beamformer $\mathbf{v}_{d,m}^{\mathrm{D}} \in \mathbb{C}^{N_u \times 1}$, while BS forwards the data from TX_UE$_n$ (decoded in uplink) to RX_UE$_n$ via the downlink cellular channel $\mathbf{G}_n^{\mathrm{D}} \in \mathbb{C}^{N_u \times N_b}$, using the TX beamformer $\mathbf{v}_{c,n}^{\mathrm{D}} \in \mathbb{C}^{N_b \times 1}$. Then, RX_UE$_m$ decodes the data from TX_UE$_m$, using the RX beamformer $\mathbf{u}_{d,m}^{\mathrm{D}} \in \mathbb{C}^{N_u \times 1}$, while RX_UE$_n$ decodes the data from BS, using the RX beamformer $\mathbf{u}_{c,n}^{\mathrm{D}} \in \mathbb{C}^{N_u \times 1}$.



(a) Transmissions in uplink band



(b) Transmissions in downlink band

Fig. 1. D2D communications underlaying multi-user MIMO cellular networks. The solid lines denote the intended transmit channels, and the dashed lines denote the interfering channels.

Compared with many current schemes, our setting endows more flexibilities to the D2D management, thereby making it possible to fully utilize the network resources and maximize the system throughput. On the other hand, this flexible setting also introduces extra interferences, and then complicates the network management. To achieve the potential benefits of this flexible setting, we should appropriately select the transmission mode for each UE pair and carefully design the associated transceivers to mitigate the interference.

B. Problem Statement

We first define the D2D set $\mathcal{D}$ and the cellular set $\mathcal{D}^{\perp}$, with $\mathcal{D} \cap \mathcal{D}^{\perp} = \emptyset$ and $\mathcal{D} \cup \mathcal{D}^{\perp} = \mathcal{M}$. If UE pair m works in D2D mode, we have $m \in \mathcal{D}$ and $\mathbf{u}_{c,m}^{\mathrm{U}} = \mathbf{0}, \mathbf{v}_{c,m}^{\mathrm{D}} = \mathbf{0}, \mathbf{u}_{c,m}^{\mathrm{D}} = \mathbf{0}$; otherwise, we have $m \in \mathcal{D}^{\perp}$ and $\mathbf{u}_{d,m}^{\mathrm{U}} = \mathbf{0}$, $\mathbf{v}_{d,m}^{\mathrm{D}} = \mathbf{0}$, $\mathbf{u}_{d,m}^{\mathrm{D}} = \mathbf{0}$. Next, we formulate the signal-to-interference-plus-noise-ratio (SINR) and rate terms of each UE pair. As shown in Fig. 1, the uplink D2D transmission between TX_UE$_m$ and

RX_UE$_m$ via channel $\mathbf{F}^{\mathrm{U}}_{m,m}$ is interfered by all the other TX_UE$_n$ via channel $\mathbf{F}^{\mathrm{U}}_{m,n}$, $\forall\ n \in \mathcal{M}$, $n \neq m$, while the downlink D2D transmission between TX_UE$_m$ and RX_UE$_m$ via channel $\mathbf{F}^{\mathrm{D}}_{m,m}$ is interfered by all the other D2D transmitter TX_UE$_n$ via $\mathbf{F}^{\mathrm{D}}_{m,n}$, $\forall\ n \in \mathcal{D}, n \neq m$, and by the BS via channel $\mathbf{G}^{\mathrm{D}}_{m}$. Thus, in D2D mode, the uplink and downlink SINRs of UE pair m are computed as

$$\mathrm{SINR}^{\mathrm{U}}_{d,m} = \frac{\left|(\mathbf{u}^{\mathrm{U}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{U}}_{m,m}\mathbf{v}^{\mathrm{U}}_{m}\right|^2}{\sigma^2\left\|\mathbf{u}^{\mathrm{U}}_{d,m}\right\|_2^2+\sum_{\substack{n\in\mathcal{M}\\ n\neq m}}\left|(\mathbf{u}^{\mathrm{U}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{U}}_{m,n}\mathbf{v}^{\mathrm{U}}_{n}\right|^2}, \tag{1a}$$

$$\mathrm{SINR}^{\mathrm{D}}_{d,m} = \frac{\left|(\mathbf{u}^{\mathrm{D}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{D}}_{m,m}\mathbf{v}^{\mathrm{D}}_{d,m}\right|^2}{\begin{pmatrix}\sigma^2\left\|\mathbf{u}^{\mathrm{D}}_{d,m}\right\|_2^2+\sum_{n\in\mathcal{D}^{\perp}}\left|(\mathbf{u}^{\mathrm{D}}_{d,m})^{\dagger}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}\right|^2\\ +\sum_{\substack{n\in\mathcal{D}\\ n\neq m}}\left|(\mathbf{u}^{\mathrm{D}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}\right|^2\end{pmatrix}}, \tag{1b}$$

where σ^2 is the noise power. Let B^{U} and B^{D} denote the uplink bandwidth and the downlink bandwidth, respectively. Then, the total rate of UE pair m in D2D mode can be expressed as

$$R_{d,m} = B^{\mathrm{U}}\log(1+\mathrm{SINR}^{\mathrm{U}}_{d,m}) + B^{\mathrm{D}}\log(1+\mathrm{SINR}^{\mathrm{D}}_{d,m}). \tag{2}$$

Similarly, in cellular mode, the uplink and downlink SINRs of UE pair m are computed as

$$\mathrm{SINR}^{\mathrm{U}}_{c,m} = \frac{\left|(\mathbf{u}^{\mathrm{U}}_{c,m})^{\dagger}\mathbf{G}^{\mathrm{U}}_{m}\mathbf{v}^{\mathrm{U}}_{m}\right|^2}{\sigma^2\left\|\mathbf{u}^{\mathrm{U}}_{c,m}\right\|_2^2+\sum_{\substack{n\in\mathcal{M}\\ n\neq m}}\left|(\mathbf{u}^{\mathrm{U}}_{c,m})^{\dagger}\mathbf{G}^{\mathrm{U}}_{n}\mathbf{v}^{\mathrm{U}}_{n}\right|^2}, \tag{3a}$$

$$\mathrm{SINR}^{\mathrm{D}}_{c,m} = \frac{\left|(\mathbf{u}^{\mathrm{D}}_{c,m})^{\dagger}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,m}\right|^2}{\begin{pmatrix}\sigma^2\left\|\mathbf{u}^{\mathrm{D}}_{c,m}\right\|_2^2+\sum_{n\in\mathcal{D}^{\perp}}\left|(\mathbf{u}^{\mathrm{D}}_{c,m})^{\dagger}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}\right|^2\\ +\sum_{\substack{n\in\mathcal{D}\\ n\neq m}}\left|(\mathbf{u}^{\mathrm{D}}_{c,m})^{\dagger}\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}\right|^2\end{pmatrix}}. \tag{3b}$$

The rate of UE pair m in cellular mode is actually determined by the smaller one of the uplink rate and the downlink rate, i.e.,

$$R_{c,m} = \min\{B^{\mathrm{U}}\log(1+\mathrm{SINR}^{\mathrm{U}}_{c,m}), B^{D}\log(1+\mathrm{SINR}^{\mathrm{D}}_{c,m})\}. \tag{4}$$

Therefore, the sum-rate maximization problem based on joint mode selection and transceiver design is formulated as

$$\begin{aligned}(\mathrm{PA}):\ &\max_{\{\mathcal{D},\mathcal{D}^{\perp},\mathbf{V},\mathbf{U}\}}\ \sum_{m\in\mathcal{D}}R_{d,m}+\sum_{m\in\mathcal{D}^{\perp}}R_{c,m}\\ \text{s.t.}\quad &\sum_{m\in\mathcal{M}}\|\mathbf{v}^{\mathrm{D}}_{c,m}\|_2^2\le P_B, &\text{(5a)}\\ &\|\mathbf{v}^{\mathrm{U}}_{m}\|_2^2\le p^{\mathrm{U}}_{m}, \|\mathbf{v}^{\mathrm{D}}_{d,m}\|_2^2\le p^{\mathrm{D}}_{m},\ \forall\, m\in\mathcal{M}, &\text{(5b)}\\ &\mathcal{D}\cap\mathcal{D}^{\perp}=\emptyset,\ \mathcal{D}\cup\mathcal{D}^{\perp}=\mathcal{M}, &\text{(5c)}\end{aligned}$$

where $\mathbf{V}$ and $\mathbf{U}$ denote $\{\mathbf{v}^{\mathrm{U}}_{m},\mathbf{v}^{\mathrm{D}}_{d,m},\mathbf{v}^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$ and $\{\mathbf{u}^{\mathrm{U}}_{d,m}, \mathbf{u}^{\mathrm{U}}_{c,m},\mathbf{u}^{\mathrm{D}}_{d,m},\mathbf{u}^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$, respectively; p^{U}_{m} and p^{D}_{m} are the transmit power budgets of TX_UE$_m$ in uplink and downlink, $\forall m\in\mathcal{M}$; P_B is the BS power budget.

C. Problem Reformulation

In this subsection, we reformulate (PA) to avoid optimizing $\mathcal{D}$ and $\mathcal{D}^{\perp}$ directly. First, we give some more general forms of the downlink SINRs in D2D and cellular modes. Replacing $n\in\mathcal{D}$ and $n\in\mathcal{D}^{\perp}$ in (1b) and (3b) simply by $n\in\mathcal{M}$, we get

$$\overline{\mathrm{SINR}}^{\mathrm{D}}_{d,m} = \frac{\left|(\mathbf{u}^{\mathrm{D}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{D}}_{m,m}\mathbf{v}^{\mathrm{D}}_{d,m}\right|^2}{\begin{pmatrix}\sigma^2\left\|\mathbf{u}^{\mathrm{D}}_{d,m}\right\|_2^2+\sum_{n\in\mathcal{M}}\left|(\mathbf{u}^{\mathrm{D}}_{d,m})^{\dagger}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}\right|^2\\ +\sum_{\substack{n\in\mathcal{M}\\ n\neq m}}\left|(\mathbf{u}^{\mathrm{D}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}\right|^2\end{pmatrix}}, \tag{6a}$$

$$\overline{\mathrm{SINR}}^{\mathrm{D}}_{c,m} = \frac{\left|(\mathbf{u}^{\mathrm{D}}_{c,m})^{\dagger}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,m}\right|^2}{\begin{pmatrix}\sigma^2\left\|\mathbf{u}^{\mathrm{D}}_{c,m}\right\|_2^2+\sum_{\substack{n\in\mathcal{M}\\ n\neq m}}\left|(\mathbf{u}^{\mathrm{D}}_{c,m})^{\dagger}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}\right|^2\\ +\sum_{n\in\mathcal{M}}\left|(\mathbf{u}^{\mathrm{D}}_{c,m})^{\dagger}\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}\right|^2\end{pmatrix}}. \tag{6b}$$

The rates of UE pair m in D2D and cellular modes can then be computed as

$$\bar{R}_{d,m} = B^{\mathrm{U}}\log(1+\mathrm{SINR}^{\mathrm{U}}_{d,m}) + B^{\mathrm{D}}\log(1+\overline{\mathrm{SINR}}^{\mathrm{D}}_{d,m}), \tag{7a}$$

$$\bar{R}_{c,m} = \min\{B^{\mathrm{U}}\log(1+\mathrm{SINR}^{\mathrm{U}}_{c,m}), B^{\mathrm{D}}\log(1+\overline{\mathrm{SINR}}^{\mathrm{D}}_{c,m})\}. \tag{7b}$$

Second, in our original formulation, the mode selection strategy is exclusive; i.e., the communication within a UE pair should be established either through BS or via D2D link. To achieve this, we further introduce the binary vector $\mathbf{d} = [d_1, d_2, \cdots, d_M]$ to indicate the transmission mode of each UE pair. If $d_m = 1$, UE pair m works in D2D mode; if $d_m = 0$, UE pair m works in cellular mode. Then, the total achievable rate of UE pair m is computed as

$$R_m = d_m\bar{R}_{d,m} + (1-d_m)\bar{R}_{c,m}. \tag{8}$$

Proposition 1: (PA) can be equivalently reformulated as the following problem (PB),

$$\begin{aligned}(\mathrm{PB}):\ &\max_{\{\mathbf{V},\mathbf{U},\mathbf{d}\}}\ \sum\nolimits_{m\in\mathcal{M}}R_m\\ \text{s.t.}\quad &\text{(5a) and (5b) satisfied,}\\ &d_m\in\{0,1\},\ \forall m\in\mathcal{M}.\end{aligned} \tag{9}$$

Proof: See Appendix A for the details of proof. □

It has been well known that the sum-rate maximization problem in multi-user MIMO-IBC networks is NP-hard [27]. Compared with the conventional sum-rate maximization problem, (PB) is obviously more challenging. This motivates us to seek some approximate solutions with manageable complexity.

III. Iterative Algorithm Based on Inexact WMMSE

A. Brief Review of WMMSE

The WMMSE approach [28], [29] transforms the sum-rate maximization problem to a weighted mean square error (MSE) minimization problem, and the two problems are equivalent in the sense that they have the same global optimal solutions.

We consider an MIMO-IBC network consisting of M pairs of UEs, where each UE pair contains one TX_UE and one RX_UE. Denote $\mathbf{H}_{n,m}$ as the channel between TX_UE$_m$ and RX_UE$_n$. Let $\mathbf{v}_m$ and $\mathbf{u}_m$ denote the TX beamformer of TX_UE$_m$ and the RX beamformer of RX_UE$_m$, respectively. We can express the SINR and rate of RX_UE$_m$ as

$$\begin{cases}\mathrm{SINR}_m = \frac{|\mathbf{u}_m^{\dagger}\mathbf{H}_{m,m}\mathbf{v}_m|^2}{\sigma^2\|\mathbf{u}_m\|_2^2+\sum_{n\neq m}|\mathbf{u}_m^{\dagger}\mathbf{H}_{m,n}\mathbf{v}_n|^2},\\ R_m = \log(1+\mathrm{SINR}_m)\end{cases} \tag{10}$$

Utilizing the well-known relation between SINR and MSE, we have $\max_{\{\mathbf{u}_m\}}(1+\mathrm{SINR}_m)=\max_{\{\mathbf{u}_m\}} e_m^{-1}$, where e_m is the MSE at RX_UE$_m$, which is given by

$$e_m=|1-\mathbf{u}_m^{\dagger}\mathbf{H}_{m,m}\mathbf{v}_m|^2+\sum_{n\neq m}|\mathbf{u}_m^{\dagger}\mathbf{H}_{m,n}\mathbf{v}_n|^2+\sigma^2\|\mathbf{u}_m\|_2^2. \tag{11}$$

Note that $\log(e_m)=\min_{\{w_m>0\}} w_m e_m-\log(w_m)+1$, then the following sum-rate maximization problem (PC) and the weighted sum-MSE minimization problem (PD) are equivalent

$$\begin{aligned}(\mathrm{PC}):\ &\max_{\{\mathbf{v}_m,\mathbf{u}_m\}}\ \sum_m R_m\\ &\text{s.t.}\quad \|\mathbf{v}_m\|_2^2\le P_m,\ \forall\ m,\end{aligned}$$

$$\begin{aligned}(\mathrm{PD}):\ &\min_{\{\mathbf{v}_m,\mathbf{u}_m,w_m\}}\ \sum_m[w_m e_m-\log(w_m)+1]\\ &\text{s.t.}\qquad \|\mathbf{v}_m\|_2^2\le P_m,\ \forall\ m,\end{aligned}$$

where w_m is the weight of e_m, and P_m is the transmit power budget of TX_UE$_m$.

The weighted sum-MSE minimization problem (PD) can be solved by the block coordinate descent (BCD) technique which updates $\{\mathbf{v}_m\}_{m=1}^M$ $\{\mathbf{u}_m\}_{m=1}^M$ and $\{w_m\}_{m=1}^M$ alternately, i.e.,

$$\mathbf{v}_m=\left(\sum\nolimits_{n=1}^{M} w_n\mathbf{H}_{n,m}^{\dagger}\mathbf{u}_n\mathbf{u}_n^{\dagger}\mathbf{H}_{n,m}+\zeta_m\mathbf{I}\right)^{-1}\mathbf{H}_{m,m}^{\dagger}\mathbf{u}_m w_m, \tag{12a}$$

$$\mathbf{u}_m=\left(\sum\nolimits_{n=1}^{M}\mathbf{H}_{m,n}\mathbf{v}_n\mathbf{v}_n^{\dagger}\mathbf{H}_{m,n}^{\dagger}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{H}_{m,m}\mathbf{v}_m, \tag{12b}$$

$$w_m=\left(1-\mathbf{u}_m^{\dagger}\mathbf{H}_{m,m}\mathbf{v}_m\right)^{-1}, \tag{12c}$$

where $\zeta_m\ge 0$ is the Lagrangian multiplier associated with $\|\mathbf{v}_m\|_2^2\le P_m$, and should be chosen carefully such that the Karush-Kuhn-Tucker (KKT) complementarity conditions [31] are satisfied; $\mathbf{u}_m$ is the MMSE receiver such that the MSE, i.e., e_m defined in (11), can be minimized; w_m is updated by the reciprocal of the minimum e_m.

According to *Theorem 3* in [28], the WMMSE algorithm[1] iteratively computes a stationary point of (PC) by repeating the steps in (12).

B. Problem Approximation

In this subsection, we formulate an efficient approximation of (PB) for which the WMMSE approach is applicable.

To this end, we first relax the binary indicator $d_m\in\{0,1\}$ by $0\le d_m\le 1$. After that, we deal with the product terms in $R_m=d_m\bar{R}_{d,m}+(1-d_m)\bar{R}_{c,m}$. Specifically, we introduce some auxiliary variables $\{s_{d,m}^{\mathrm{U}},s_{d,m}^{\mathrm{D}},s_{c,m}\}_{m\in\mathcal{M}}$, and bound them by $d_m B^{\mathrm{U}}\log(1+\mathrm{SINR}_{d,m}^{\mathrm{U}})\ge(s_{d,m}^{\mathrm{U}})^2$, $d_m B^{\mathrm{D}}\log(1+\overline{\mathrm{SINR}}_{d,m}^{\mathrm{D}})\ge(s_{d,m}^{\mathrm{D}})^2$, $(1-d_m)B^{\mathrm{U}}\log(1+\mathrm{SINR}_{c,m}^{\mathrm{U}})\ge s_{c,m}^2$, and $(1-d_m)B^{\mathrm{D}}\log(1+\overline{\mathrm{SINR}}_{c,m}^{\mathrm{D}})\ge s_{c,m}^2$, $m\in\mathcal{M}$. Then, the term of R_m can be replaced by $(s_{d,m}^{\mathrm{U}})^2+(s_{d,m}^{\mathrm{D}})^2+s_{c,m}^2$ in the objective of (PB). Considering that the "quadratic-over-linear" function $h(\omega,\nu)\triangleq\frac{\omega^2}{\nu}$ is convex as $\nu>0$ [31], we can recast $\varrho\nu\ge\omega^2$ as a convex constraint $\varrho\ge\frac{\omega^2}{\nu}$ for $\nu>0$. Following this idea, we get an approximation of (PB) as

$$\begin{aligned}(\mathrm{PE}):\ &\max_{\{\mathbf{V},\mathbf{U},\mathbf{d},\mathbf{s}\}}\ \sum_{m\in\mathcal{M}}(s_{d,m}^{\mathrm{U}})^2+(s_{d,m}^{\mathrm{D}})^2+s_{c,m}^2\\ &\text{s.t.}\quad \text{(5a) and (5b) satisfied,}\\ &\qquad 0\le d_m\le 1,\ \forall m\in\mathcal{M},\qquad (13a)\\ &\qquad B^{\mathrm{U}}\log(1+\mathrm{SINR}_{d,m}^{\mathrm{U}})\ge\tfrac{(s_{d,m}^{\mathrm{U}})^2}{d_m+\epsilon},\ \forall m\in\mathcal{M},\qquad (13b)\\ &\qquad B^{\mathrm{D}}\log(1+\overline{\mathrm{SINR}}_{d,m}^{\mathrm{D}})\ge\tfrac{(s_{d,m}^{\mathrm{D}})^2}{d_m+\epsilon},\ \forall m\in\mathcal{M},\qquad (13c)\\ &\qquad \min\left\{\begin{array}{l}B^{\mathrm{U}}\log(1+\mathrm{SINR}_{c,m}^{\mathrm{U}}),\\ B^{\mathrm{D}}\log(1+\overline{\mathrm{SINR}}_{c,m}^{\mathrm{D}})\end{array}\right\}\ge\tfrac{s_{c,m}^2}{1-d_m+\epsilon},\\ &\qquad\qquad\qquad\qquad\qquad\qquad\forall m\in\mathcal{M},\qquad (13d)\end{aligned}$$

where $\mathbf{s}$ denotes $\{s_{d,m}^{\mathrm{U}},s_{d,m}^{\mathrm{D}},s_{c,m}\}_{m\in\mathcal{M}}$, and ϵ is a small positive number introduced to avoid the numerical problems of zero denominator.

We further apply the WMMSE reformulation to handle the $\log(\cdot)$ terms in (13b) – (13d), thereby generating (PF). From *Theorem 1* in [28] and *Lemma 3* in [29], (PE) and (PF) are equivalent in the sense that they have the same optimal solutions. Specifically, (PF) is expressed as

$$\begin{aligned}(\mathrm{PF}):\ &\min_{\{\mathbf{V},\mathbf{U},\mathbf{W},\mathbf{d},\mathbf{s}\}}\ -\sum_{m\in\mathcal{M}}(s_{d,m}^{\mathrm{U}})^2+(s_{d,m}^{\mathrm{D}})^2+s_{c,m}^2\\ &\text{s.t.}\quad \text{(5a), (5b) and (13a) satisfied,}\\ &\qquad \log(w_{d,m}^{\mathrm{U}})-w_{d,m}^{\mathrm{U}}e_{d,m}^{\mathrm{U}}(\mathbf{V}^{\mathrm{U}},\mathbf{u}_{d,m}^{\mathrm{U}})+1\ge\tfrac{(s_{d,m}^{\mathrm{U}})^2}{B^{\mathrm{U}}(d_m+\epsilon)},\\ &\qquad\qquad\qquad\qquad\forall m\in\mathcal{M},\qquad (14a)\\ &\qquad \log(w_{d,m}^{\mathrm{D}})-w_{d,m}^{\mathrm{D}}e_{d,m}^{\mathrm{D}}(\mathbf{V}^{\mathrm{D}},\mathbf{u}_{d,m}^{\mathrm{D}})+1\ge\tfrac{(s_{d,m}^{\mathrm{D}})^2}{B^{\mathrm{D}}(d_m+\epsilon)},\\ &\qquad\qquad\qquad\qquad\forall m\in\mathcal{M},\qquad (14b)\\ &\qquad \log(w_{c,m}^{\mathrm{U}})-w_{c,m}^{\mathrm{U}}e_{c,m}^{\mathrm{U}}(\mathbf{V}^{\mathrm{U}},\mathbf{u}_{c,m}^{\mathrm{U}})+1\ge\tfrac{s_{c,m}^2}{B^{\mathrm{U}}(1-d_m+\epsilon)},\\ &\qquad\qquad\qquad\qquad\forall m\in\mathcal{M},\qquad (14c)\\ &\qquad \log(w_{c,m}^{\mathrm{D}})-w_{c,m}^{\mathrm{D}}e_{c,m}^{\mathrm{D}}(\mathbf{V}^{\mathrm{D}},\mathbf{u}_{c,m}^{\mathrm{D}})+1\ge\tfrac{s_{c,m}^2}{B^{\mathrm{D}}(1-d_m+\epsilon)},\\ &\qquad\qquad\qquad\qquad\forall m\in\mathcal{M},\qquad (14d)\end{aligned}$$

where $\mathbf{W}$ is the collection of weighting factors $\{w_{d,m}^{\mathrm{U}},w_{c,m}^{\mathrm{U}},w_{d,m}^{\mathrm{D}},w_{c,m}^{\mathrm{D}}\}_{m\in\mathcal{M}}$; $\mathbf{V}^{\mathrm{U}}$ and $\mathbf{V}^{\mathrm{D}}$ denote the uplink TX beamformers $\{\mathbf{v}_m^{\mathrm{U}}\}_{m\in\mathcal{M}}$ and the downlink TX beamformers $\{\mathbf{v}_{d,m}^{\mathrm{D}},\mathbf{v}_{c,m}^{\mathrm{D}}\}_{m\in\mathcal{M}}$; $e_{d,m}^{\mathrm{U}}(\mathbf{V}^{\mathrm{U}},\mathbf{u}_{d,m}^{\mathrm{U}})$ and $e_{d,m}^{\mathrm{D}}(\mathbf{V}^{\mathrm{D}},\mathbf{u}_{d,m}^{\mathrm{D}})$ are the uplink and downlink D2D MSE values of UE pair m; $e_{c,m}^{\mathrm{U}}(\mathbf{V}^{\mathrm{U}},\mathbf{u}_{c,m}^{\mathrm{U}})$ and $e_{c,m}^{\mathrm{D}}(\mathbf{V}^{\mathrm{D}},\mathbf{u}_{c,m}^{\mathrm{D}})$ are the uplink and downlink cellular MSE values of UE pair m. Similar as (11), these

[1]There are key differences between (PB) and the conventional sum-rate maximization problem (PC). First, we consider the underlay D2D communication scenario which involves both D2D uplink/downlink transmissions and cellular uplink/downlink transmissions, while (PC) only considers the cellular downlink transmission. Second, the objective of (PB) adopts a more complicated form involving the product terms of $d_m\bar{R}_{c,m}$ and $d_m\bar{R}_{d,m}$. Therefore, the WMMSE algorithm cannot be directly applied to solve (PB).

MSE values are defined in (15),

$$e^{\mathrm{U}}_{d,m}(\mathbf{V}^{\mathrm{U}},\mathbf{u}^{\mathrm{U}}_{d,m}) = \sigma^2\|\mathbf{u}^{\mathrm{U}}_{d,m}\|_2^2 + |1-(\mathbf{u}^{\mathrm{U}}_{d,m})^\dagger\mathbf{F}^{\mathrm{U}}_{m,m}\mathbf{v}^{\mathrm{U}}_m|^2 + \sum_{\substack{n\in\mathcal{M}\\ n\neq m}} |(\mathbf{u}^{\mathrm{U}}_{d,m})^\dagger\mathbf{F}^{\mathrm{U}}_{m,n}\mathbf{v}^{\mathrm{U}}_n|^2, \tag{15a}$$

$$e^{\mathrm{D}}_{d,m}(\mathbf{V}^{\mathrm{D}},\mathbf{u}^{\mathrm{D}}_{d,m}) = \sigma^2\|\mathbf{u}^{\mathrm{D}}_{d,m}\|_2^2 + |1-(\mathbf{u}^{\mathrm{D}}_{d,m})^\dagger\mathbf{F}^{\mathrm{D}}_{m,m}\mathbf{v}^{\mathrm{D}}_{d,m}|^2 + \sum_{\substack{n\in\mathcal{M}\\ n\neq m}} |(\mathbf{u}^{\mathrm{D}}_{d,m})^\dagger\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}|^2 + \sum_{n\in\mathcal{M}} |(\mathbf{u}^{\mathrm{D}}_{d,m})^\dagger\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}|^2, \tag{15b}$$

$$e^{\mathrm{U}}_{c,m}(\mathbf{V}^{\mathrm{U}},\mathbf{u}^{\mathrm{U}}_{c,m}) = \sigma^2\|\mathbf{u}^{\mathrm{U}}_{c,m}\|_2^2 + |1-(\mathbf{u}^{\mathrm{U}}_{c,m})^\dagger\mathbf{G}^{\mathrm{U}}_{m}\mathbf{v}^{\mathrm{U}}_m|^2 + \sum_{\substack{n\in\mathcal{M}\\ n\neq m}} |(\mathbf{u}^{\mathrm{U}}_{c,m})^\dagger\mathbf{G}^{\mathrm{U}}_{n}\mathbf{v}^{\mathrm{U}}_n|^2, \tag{15c}$$

$$e^{\mathrm{D}}_{c,m}(\mathbf{V}^{\mathrm{D}},\mathbf{u}^{\mathrm{D}}_{c,m}) = \sigma^2\|\mathbf{u}^{\mathrm{D}}_{c,m}\|_2^2 + |1-(\mathbf{u}^{\mathrm{D}}_{c,m})^\dagger\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,m}|^2 + \sum_{n\in\mathcal{M}} |(\mathbf{u}^{\mathrm{D}}_{c,m})^\dagger\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}|^2 + \sum_{\substack{n\in\mathcal{M}\\ n\neq m}} |(\mathbf{u}^{\mathrm{D}}_{c,m})^\dagger\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}|^2. \tag{15d}$$

C. Iterative Algorithm Based on Inexact WMMSE

Due to its non-convex objective, we handle (PF) by iteratively solving its sequential convex approximations (SCA) based on the idea of difference-of-convex (DC) programming [32]. Specifically, in each iteration, we perform the first-order approximation to $[(s^{\mathrm{U}}_{d,m})^2+(s^{\mathrm{D}}_{d,m})^2+s^2_{c,m}]$, and then solve the resultant problem

$$(\mathrm{PG}): \min_{\{\mathbf{V},\mathbf{U},\mathbf{W},\mathbf{d},\mathbf{s}\}} -2\sum_{m\in\mathcal{M}} \hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m} + \hat{s}^{\mathrm{D}}_{d,m}s^{\mathrm{D}}_{d,m} + \hat{s}_{c,m}s_{c,m}$$
$$\text{s.t.} \quad (5a),(5b),(13a) \text{ and } (14a)-(14d) \text{ satisfied},$$

with $\hat{s}^{\mathrm{U}}_{d,m}$, $\hat{s}^{\mathrm{D}}_{d,m}$ and $\hat{s}_{c,m}$ being the iterates of $s^{\mathrm{U}}_{d,m}$, $s^{\mathrm{D}}_{d,m}$ and $s_{c,m}$ in the previous iteration.

Following the WMMSE framework of (12), we divide the variables $\{\mathbf{V},\mathbf{U},\mathbf{W},\mathbf{d},\mathbf{s}\}$ of (PG) into three blocks, i.e., $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$, $\mathbf{U}$, and $\mathbf{W}$, and then apply the BCD method to solve it iteratively. First, fixing $\mathbf{U}$ and $\mathbf{W}$, we solve the following convex problem to update $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$,

$$(\mathrm{PH}): \min_{\{\mathbf{V},\mathbf{d},\mathbf{s}\}} -2\sum_{m\in\mathcal{M}} \hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m} + \hat{s}^{\mathrm{D}}_{d,m}s^{\mathrm{D}}_{d,m} + \hat{s}_{c,m}s_{c,m}$$
$$\text{s.t.} \quad (5a),(5b),(13a), \text{ and } (14a)-(14d) \text{ satisfied}.$$

Problem (PH) is convex can be optimally solved by CVX [33]. Define the mapping function $\Omega(\mathbf{U},\mathbf{W};\hat{\mathbf{s}})$, where $\hat{\mathbf{s}}$ denotes $\{\hat{s}^{\mathrm{U}}_{d,m},\hat{s}^{\mathrm{D}}_{d,m},\hat{s}_{c,m}\}_{m\in\mathcal{M}}$, such that every element in the range of the map is an optimal solution to (PH). Then, the update of $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$ can be simply written as

$$\{\mathbf{V},\mathbf{d},\mathbf{s}\} \in \Omega(\mathbf{U},\mathbf{W};\hat{\mathbf{s}}). \tag{16}$$

Next, fixing $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$ and $\mathbf{W}$, we update $\mathbf{U}$ by solving the following problem,

$$(\mathrm{PI}): \min_{\mathbf{U}} -2\sum_{m\in\mathcal{M}} \hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m} + \hat{s}^{\mathrm{D}}_{d,m}s^{\mathrm{D}}_{d,m} + \hat{s}_{c,m}s_{c,m}$$
$$\text{s.t.} \quad (14a)-(14d) \text{ satisfied}.$$

In this subproblem, $\mathbf{U}$ only appears in the MSE terms in (14), i.e., $\{e^{\mathrm{U}}_{d,m},e^{\mathrm{D}}_{d,m},e^{\mathrm{U}}_{c,m},e^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$. Since we consider a sum-rate maximization problem here, $\mathbf{U}$ is selected to minimize the MSE terms so that the rates of the UE pairs can be maximally improved in this step. In addition, updating $\mathbf{U}$ as the optimal MMSE receiver also maximizes the right-hand side of (14), thereby maximally enlarging the searching space of $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$ in the next iteration. This helps improve the system throughput quickly. Therefore, the optimal solution of $\mathbf{U}$ is given by

$$\mathbf{u}^{\mathrm{U}}_{d,m} = \left[\sum_{n\in\mathcal{M}} \mathbf{F}^{\mathrm{U}}_{m,n}\mathbf{v}^{\mathrm{U}}_n(\mathbf{v}^{\mathrm{U}}_n)^\dagger(\mathbf{F}^{\mathrm{U}}_{m,n})^\dagger + \sigma^2\mathbf{I}\right]^{-1}\mathbf{F}^{\mathrm{U}}_{m,m}\mathbf{v}^{\mathrm{U}}_m \tag{17a}$$

$$\mathbf{u}^{\mathrm{U}}_{c,m} = \left[\sum_{n\in\mathcal{M}} \mathbf{G}^{\mathrm{U}}_{n}\mathbf{v}^{\mathrm{U}}_n(\mathbf{v}^{\mathrm{U}}_n)^\dagger(\mathbf{G}^{\mathrm{U}}_{n})^\dagger + \sigma^2\mathbf{I}\right]^{-1}\mathbf{G}^{\mathrm{U}}_{m}\mathbf{v}^{\mathrm{U}}_m \tag{17b}$$

$$\mathbf{u}^{\mathrm{D}}_{d,m} = \left[\sum_{n\in\mathcal{M}}\begin{bmatrix}\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}(\mathbf{v}^{\mathrm{D}}_{d,n})^\dagger(\mathbf{F}^{\mathrm{D}}_{m,n})^\dagger\\ +\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}(\mathbf{v}^{\mathrm{D}}_{c,n})^\dagger(\mathbf{G}^{\mathrm{D}}_{m})^\dagger\end{bmatrix} + \sigma^2\mathbf{I}\right]^{-1}\mathbf{F}^{\mathrm{D}}_{m,m}\mathbf{v}^{\mathrm{D}}_{d,m} \tag{17c}$$

$$\mathbf{u}^{\mathrm{D}}_{c,m} = \left[\sum_{n\in\mathcal{M}}\begin{bmatrix}\mathbf{F}^{\mathrm{D}}_{m,n}\mathbf{v}^{\mathrm{D}}_{d,n}(\mathbf{v}^{\mathrm{D}}_{d,n})^\dagger(\mathbf{F}^{\mathrm{D}}_{m,n})^\dagger\\ +\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,n}(\mathbf{v}^{\mathrm{D}}_{c,n})^\dagger(\mathbf{G}^{\mathrm{D}}_{m})^\dagger\end{bmatrix} + \sigma^2\mathbf{I}\right]^{-1}\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,m}. \tag{17d}$$

For the sake of brevity, we simplify the formulas of updating $\mathbf{U}$ as

$$\mathbf{U} = \Gamma(\mathbf{V}). \tag{18}$$

Lastly, fixing $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$ and $\mathbf{U}$, we update $\mathbf{W}$ by solving the following problem,

$$(\mathrm{PJ}): \min_{\mathbf{W}} -2\sum_{m\in\mathcal{M}} \hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m} + \hat{s}^{\mathrm{D}}_{d,m}s^{\mathrm{D}}_{d,m} + \hat{s}_{c,m}s_{c,m}$$
$$\text{s.t.} \quad (14a)-(14d) \text{ satisfied}.$$

Similarly, $\mathbf{W}$ is chosen to maximize the right-hand side of (14), which is actually the reciprocal of the minimum MSE, i.e.,

$$w^{\mathrm{U}}_{d,m} = \left[1-(\mathbf{u}^{\mathrm{U}}_{d,m})^\dagger\mathbf{F}^{\mathrm{D}}_{m,m}\mathbf{v}^{\mathrm{U}}_m\right]^{-1} \tag{19a}$$

$$w^{\mathrm{U}}_{c,m} = \left[1-(\mathbf{u}^{\mathrm{U}}_{c,m})^\dagger\mathbf{G}^{\mathrm{U}}_{m}\mathbf{v}^{\mathrm{U}}_m\right]^{-1} \tag{19b}$$

$$w^{\mathrm{D}}_{d,m} = \left[1-(\mathbf{u}^{\mathrm{D}}_{d,m})^\dagger\mathbf{F}^{\mathrm{D}}_{m,m}\mathbf{v}^{\mathrm{D}}_{d,m}\right]^{-1} \tag{19c}$$

$$w^{\mathrm{D}}_{c,m} = \left[1-(\mathbf{u}^{\mathrm{D}}_{c,m})^\dagger\mathbf{G}^{\mathrm{D}}_{m}\mathbf{v}^{\mathrm{D}}_{c,m}\right]^{-1} \tag{19d}$$

where $\mathbf{u}^{\mathrm{U}}_{d,m},\mathbf{u}^{\mathrm{U}}_{c,m},\mathbf{u}^{\mathrm{D}}_{d,m}$ and $\mathbf{u}^{\mathrm{D}}_{c,m}$ are computed as in (17). Again, for the sake of brevity, we simplify the formulas of updating $\mathbf{W}$ as

$$\mathbf{W} = \Upsilon(\mathbf{V},\mathbf{U}). \tag{20}$$

By repeating the steps of (16), (18) and (20), the WMMSE algorithm iteratively solves (PG) with some given $\hat{\mathbf{s}}$. Then, we update $\hat{\mathbf{s}} = \mathbf{s}$, and repeat the above WMMSE loop. Finally, a two-layer (including the outer DC programming layer and the inner WMMSE layer) algorithm was developed to iteratively solve (PF) or the equivalent (PE) with stationary convergence guarantee, which is referred to as the iterative algorithm based on exact WMMSE (IA-EWMMSE) in this paper, since the update of $\hat{\mathbf{s}}$ is based on the exact solution of the inner WMMSE problem (PG). We summarize IA-EWMMSE in Table II

TABLE II
SUMMARY OF IA-EWMMSE

1. Initialize $\{\mathbf{U}, \mathbf{W}, \hat{\mathbf{s}}\}$;
2. **Repeat** *(outer DC programming layer)*
3. **Repeat** *(inner WMMSE layer)*
4. $\{\mathbf{V}, \mathbf{d}, \mathbf{s}\} \in \Omega(\mathbf{U}, \mathbf{W}; \hat{\mathbf{s}})$;
5. $\mathbf{U} = \Gamma(\mathbf{V})$;
6. $\mathbf{W} = \Upsilon(\mathbf{V}, \mathbf{U})$;
7. **Until** some stopping criterion is satisfied;
8. $\hat{\mathbf{s}} = \mathbf{s}$;
9. **Until** some stopping criterion is satisfied;

TABLE III
SUMMARY OF IA-IWMMSE

1. Initialize $\{\mathbf{U}, \mathbf{W}, \hat{\mathbf{s}}\}$;
2. **Repeat**
3. $\{\mathbf{V}, \mathbf{d}, \mathbf{s}\} \in \Omega(\mathbf{U}, \mathbf{W}; \hat{\mathbf{s}})$;
4. $\mathbf{U} = \Gamma(\mathbf{V})$;
5. $\mathbf{W} = \Upsilon(\mathbf{V}, \mathbf{U})$;
6. $\hat{\mathbf{s}} = \mathbf{s}$;
7. **Until** some stopping criterion is satisfied;

(more details can be found in our previous conference paper [1]), where the stopping criterion is satisfied as the difference between the iterates of two adjacent iterations falls below some predefined threshold (same for Tables III and IV).

However, it may be difficult to implement IA-EWMMSE in practice due to the complicated two-layer loops. To simplify the algorithm structure, it is natural to consider designing some algorithm, which updates $\hat{\mathbf{s}}$ based on the inexact solution of (PG) while with convergence guarantee. Motivated by this idea, we propose an algorithm referred to as the iterative algorithm based on inexact WMMSE (IA-IWMMSE). The main steps of IA-IWMMSE are summarized in Table III, where (16), (18), and (20) are executed only once in each iteration to update $\hat{\mathbf{s}}$. Interestingly, although $\hat{\mathbf{s}}$ is updated in an inexact manner, we still have the following convergence result, i.e., Proposition 2, for IA-IWMMSE[2]. As shown later in the simulation results of Fig. 4, in addition to having the simpler single-layer structure, IA-IWMMSE accelerates the convergence, and thereby has lower total complexity than IA-EWMMSE.

Proposition 2: Every accumulation point of the iterates generated by IA-IWMMSE in Table III is a stationary solution of (PE).

Proof: See Appendix B for the details of proof. □

Before closing this section, we summarize our methodology for problem (PA). First, to avoid optimizing the set variables $\mathcal{D}$ and $\mathcal{D}^{\perp}$ directly, we introduce the binary indicators $\{d_m\}_{m\in\mathcal{M}}$, and obtain the equivalent problem (PB). Then, we relax the binary indicators and approximate (PB) by (PE). Since (PE) is still NP-hard, we next perform the WMMSE reformulation to get (PF), which is equivalent to (PE) in the sense that they have the same optimal solutions. However, the non-convex objective of (PF) prevents us from applying the WMMSE algorithm. To tackle this, we iteratively solve its sequential convex approximations, i.e., (PG), by performing the first-order approximation. In each iteration, we further use BCD to divide (PG) into three convex subproblems, i.e., (PH), (PI), and (PJ). Finally, we get a stationary solution of (PF), which is also a stationary solution of (PE). In summary, the original problem (PA) is challenging to solve, and even its approximation (PE) is still NP-hard. As a compromise, we pursue a stationary solution of (PE) with manageable complexity.

[2]After IA-IWMMSE converges, we quantize $\{d_m\}_{m\in\mathcal{M}}$ to binary values to identify the cellular UEs and the D2D UEs. Then, we optimize the transceivers by solving two independent (i.e., uplink and downlink) sum-rate problems with known transmission modes. Since these problems have been intensively studied [28], we omit the details due to the space limitation.

TABLE IV
SUMMARY OF THE INTRODUCED AUXILIARY VARIABLES

Variable Name	Description
$\mathbf{T}_d^{\mathrm{U}}(m)$	$\mathbf{T}_d^{\mathrm{U}}(m) = \mathbf{V}^{\mathrm{U}}$; $\mathbf{T}_d^{\mathrm{U}}(m) = \{\mathbf{t}_{d,n}^{\mathrm{U}}(m)\}_{n\in\mathcal{M}}$, $\mathbf{V}^{\mathrm{U}} = \{\mathbf{v}_n^{\mathrm{U}}\}_{n\in\mathcal{M}}$.
$\mathbf{T}_c^{\mathrm{U}}(m)$	$\mathbf{T}_c^{\mathrm{U}}(m) = \mathbf{V}^{\mathrm{U}}$; $\mathbf{T}_c^{\mathrm{U}}(m) = \{\mathbf{t}_{c,n}^{\mathrm{U}}(m)\}_{n\in\mathcal{M}}$, $\mathbf{V}^{\mathrm{U}} = \{\mathbf{v}_n^{\mathrm{U}}\}_{n\in\mathcal{M}}$.
$\mathbf{T}_d^{\mathrm{D}}(m)$	$\mathbf{T}_d^{\mathrm{D}}(m) = \mathbf{V}^{\mathrm{D}}$; $\mathbf{T}_d^{\mathrm{D}}(m) = \{\mathbf{t}_{d;d,n}^{\mathrm{D}}(m), \mathbf{t}_{d;c,n}^{\mathrm{D}}(m)\}_{n\in\mathcal{M}}$, $\mathbf{V}^{\mathrm{D}} = \{\mathbf{v}_{d,n}^{\mathrm{D}}, \mathbf{v}_{c,n}^{\mathrm{D}}\}_{n\in\mathcal{M}}$.
$\mathbf{T}_c^{\mathrm{D}}(m)$	$\mathbf{T}_c^{\mathrm{D}}(m) = \mathbf{V}^{\mathrm{D}}$; $\mathbf{T}_c^{\mathrm{D}}(m) = \{\mathbf{t}_{c;d,n}^{\mathrm{D}}(m), \mathbf{t}_{c;c,n}^{\mathrm{D}}(m)\}_{n\in\mathcal{M}}$, $\mathbf{V}^{\mathrm{D}} = \{\mathbf{v}_{d,n}^{\mathrm{D}}, \mathbf{v}_{c,n}^{\mathrm{D}}\}_{n\in\mathcal{M}}$.
$x_{d,m}^{\mathrm{U}}, x_{c,m}^{\mathrm{U}}, x_{d,m}^{\mathrm{D}}, x_{c,m}^{\mathrm{D}}$ $y_{d,m}^{\mathrm{U}}, y_{d,m}^{\mathrm{D}}, y_{c,m}$	$B^{\mathrm{U}} x_{d,m}^{\mathrm{U}} = y_{d,m}^{\mathrm{U}}$, $B^{\mathrm{U}} x_{c,m}^{\mathrm{U}} = y_{c,m}$, $B^{\mathrm{D}} x_{d,m}^{\mathrm{D}} = y_{d,m}^{\mathrm{D}}$, $B^{\mathrm{D}} x_{c,m}^{\mathrm{D}} = y_{c,m}$.
$z_{d,m}^{\mathrm{U}}, z_{d,m}^{\mathrm{D}}, z_{c,m}$	$z_{d,m}^{\mathrm{U}} = d_m + \epsilon$, $z_{d,m}^{\mathrm{D}} = d_m + \epsilon$, $z_{c,m} = 1 - d_m + \epsilon$.

IV. DISTRIBUTED ADMM ALGORITHM FOR (PH)

Except for updating $\{\mathbf{V}, \mathbf{d}, \mathbf{s}\}$, IA-IWMMSE can be executed distributively. Unfortunately, updating $\{\mathbf{V}, \mathbf{d}, \mathbf{s}\}$ in the centralized way, e.g., by the interior-point (IP) method [31], requires a per-iteration complexity of $\mathcal{O}(M^3(N_u + N_d)^3)$, which dominates the total computational cost. Obviously, a fully distributed algorithm is preferred in this circumstance, which may compute $\{\mathbf{v}_m^{\mathrm{U}}, \mathbf{v}_{d,m}^{\mathrm{D}}, \mathbf{v}_{c,m}^{\mathrm{D}}\}_{m\in\mathcal{M}}$, $\{d_m\}_{m\in\mathcal{M}}$ and $\{s_{d,m}^{\mathrm{U}}, s_{d,m}^{\mathrm{D}}, s_{c,m}\}_{m\in\mathcal{M}}$ independently. To this end, we further recast (PH) so that it fits into the ADMM framework [30], and then devise a distributed algorithm to solve it, with a simple closed-form solution in each step.

A. ADMM Reformulation of (PH)

To decouple the variables in (PH), we first introduce three groups of auxiliary variables, which are summarized in Table IV.

1) We introduce M copies of the TX beamformers, i.e., $\{\mathbf{T}_d^{\mathrm{U}}(m), \mathbf{T}_c^{\mathrm{U}}(m), \mathbf{T}_d^{\mathrm{D}}(m), \mathbf{T}_c^{\mathrm{D}}(m)\}_{m\in\mathcal{M}}$, where $\mathbf{T}_d^{\mathrm{U}}(m) = \mathbf{T}_c^{\mathrm{U}}(m) = \mathbf{V}^{\mathrm{U}}$, and $\mathbf{T}_d^{\mathrm{D}}(m) = \mathbf{T}_c^{\mathrm{D}}(m) = \mathbf{V}^{\mathrm{D}}$, $\forall m \in \mathcal{M}$. Then, we replace the $\mathbf{V}^{\mathrm{U}}$ in (14a) and (14c) by $\mathbf{T}_d^{\mathrm{U}}(m)$ and $\mathbf{T}_c^{\mathrm{U}}(m)$, respectively, to decouple the $2M$ uplink constraints of $\mathbf{V}^{\mathrm{U}}$. Similarly, we replace the $\mathbf{V}^{\mathrm{D}}$ in (14b) and (14d) by $\mathbf{T}_d^{\mathrm{D}}(m)$ and $\mathbf{T}_c^{\mathrm{D}}(m)$, respectively, to decouple the $2M$ downlink constraints of $\mathbf{V}^{\mathrm{D}}$.

2) We introduce two series of variables: $\{x^{\mathrm{U}}_{d,m}, x^{\mathrm{U}}_{c,m}, x^{\mathrm{D}}_{d,m}, x^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$ and $\{y^{\mathrm{U}}_{d,m},\ y^{\mathrm{D}}_{d,m}, y_{c,m}\}_{m\in\mathcal{M}}$, where $B^{\mathrm{U}}x^{\mathrm{U}}_{d,m} = y^{\mathrm{U}}_{d,m}$, $B^{\mathrm{U}}x^{\mathrm{U}}_{c,m} = y_{c,m}$, $B^{\mathrm{D}}x^{\mathrm{D}}_{d,m} = y^{\mathrm{D}}_{d,m}$, and $B^{\mathrm{D}}x^{\mathrm{D}}_{c,m} = y_{c,m}$, $\forall m \in \mathcal{M}$.

3) We also introduce a series of indicators: $\{z^{\mathrm{U}}_{d,m}, z^{\mathrm{D}}_{d,m}, z_{c,m}\}_{m\in\mathcal{M}}$, where $z^{\mathrm{U}}_{d,m} = d_m+\epsilon$, $z^{\mathrm{D}}_{d,m} = d_m+\epsilon$, and $z_{c,m} = 1-d_m+\epsilon,\ \forall m \in \mathcal{M}$.

Inserting these variables into (PH), we equivalently recast it as

$$\begin{aligned}
(\mathrm{PK}):\ &\min_{\{\mathbf{V},\mathbf{d},\mathbf{s},\mathbf{T},\mathbf{x},\mathbf{y},\mathbf{z}\}} -2\sum_{m\in\mathcal{M}}\begin{pmatrix}\hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m}+\hat{s}^{\mathrm{D}}_{d,m}s^{\mathrm{D}}_{d,m}\\ +\hat{s}_{c,m}s_{c,m}\end{pmatrix}\\
\text{s.t.}\quad &(5a),(5b)\text{ and }(13a)\text{ satisfied},\\
&\log(w^{\mathrm{U}}_{d,m}) - w^{\mathrm{U}}_{d,m}e^{\mathrm{U}}_{d,m}(\mathbf{T}^{\mathrm{U}}_d(m),\mathbf{u}^{\mathrm{U}}_{d,m}) + 1 \geq x^{\mathrm{U}}_{d,m}, \quad \forall m\in\mathcal{M}, &(21a)\\
&\log(w^{\mathrm{D}}_{d,m}) - w^{\mathrm{D}}_{d,m}e^{\mathrm{D}}_{d,m}(\mathbf{T}^{\mathrm{D}}_d(m),\mathbf{u}^{\mathrm{D}}_{d,m}) + 1 \geq x^{\mathrm{D}}_{d,m}, \quad \forall m\in\mathcal{M}, &(21b)\\
&\log(w^{\mathrm{U}}_{c,m}) - w^{\mathrm{U}}_{c,m}e^{\mathrm{U}}_{c,m}(\mathbf{T}^{\mathrm{U}}_c(m),\mathbf{u}^{\mathrm{U}}_{c,m}) + 1 \geq x^{\mathrm{U}}_{c,m}, \quad \forall m\in\mathcal{M}, &(21c)\\
&\log(w^{\mathrm{D}}_{c,m}) - w^{\mathrm{D}}_{c,m}e^{\mathrm{D}}_{c,m}(\mathbf{T}^{\mathrm{D}}_c(m),\mathbf{u}^{\mathrm{D}}_{c,m}) + 1 \geq x^{\mathrm{D}}_{c,m}, \quad \forall m\in\mathcal{M}, &(21d)\\
&y^{\mathrm{U}}_{d,m} \geq \tfrac{(s^{\mathrm{U}}_{d,m})^2}{z^{\mathrm{U}}_{d,m}},\ y^{\mathrm{D}}_{d,m} \geq \tfrac{(s^{\mathrm{D}}_{d,m})^2}{z^{\mathrm{D}}_{d,m}},\ y_{c,m} \geq \tfrac{s^2_{c,m}}{z_{c,m}}, \quad \forall m\in\mathcal{M}, &(21e)\\
&z^{\mathrm{U}}_{d,m} \geq \epsilon,\ z^{\mathrm{D}}_{d,m} \geq \epsilon,\ z_{c,m} \geq \epsilon, \quad \forall m\in\mathcal{M}, &(21f)\\
&B^{\mathrm{U}}x^{\mathrm{U}}_{d,m} = y^{\mathrm{U}}_{d,m},\ B^{\mathrm{D}}x^{\mathrm{D}}_{d,m} = y^{\mathrm{D}}_{d,m},\ \forall m\in\mathcal{M}, &(21g)\\
&B^{\mathrm{U}}x^{\mathrm{U}}_{c,m} = y_{c,m},\ B^{\mathrm{D}}x^{\mathrm{D}}_{c,m} = y_{c,m},\ \forall m\in\mathcal{M}, &(21h)\\
&z^{\mathrm{U}}_{d,m} = d_m+\epsilon,\ z^{\mathrm{D}}_{d,m} = d_m+\epsilon,\ z_{c,m} = 1-d_m+\epsilon, \quad \forall m\in\mathcal{M}, &(21i)\\
&\mathbf{t}^{\mathrm{U}}_{d,n}(m) = \mathbf{v}^{\mathrm{U}}_n,\ \mathbf{t}^{\mathrm{U}}_{c,n}(m) = \mathbf{v}^{\mathrm{U}}_n, \quad \forall m,n\in\mathcal{M}, &(21j)\\
&\mathbf{t}^{\mathrm{D}}_{d;d,n}(m) = \mathbf{t}^{\mathrm{D}}_{c;d,n}(m) = \mathbf{v}^{\mathrm{D}}_{d,n}, \quad \forall m,n\in\mathcal{M}, &(21k)\\
&\mathbf{t}^{\mathrm{D}}_{c;d,n}(m) = \mathbf{t}^{\mathrm{D}}_{c;c,n}(m) = \mathbf{v}^{\mathrm{D}}_{c,n}, \quad \forall m,n\in\mathcal{M}. &(21l)
\end{aligned}$$

The partial *augmented Lagrangian function* [30] of (PK) is defined as (22) at the top of next page, where $c>0$ is the penalty parameter; $\mathbf{\Phi}, \boldsymbol{\psi}$, and $\boldsymbol{\theta}$ are the associated Lagrangian multipliers. Dividing the variables $\{\mathbf{V},\mathbf{d},\mathbf{s},\mathbf{T},\mathbf{x},\mathbf{y},\mathbf{z}\}$ into two blocks of $\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\}$ and $\{\mathbf{T},\mathbf{d},\mathbf{x}\}$, we perform the two-block ADMM framework shown in Table V to iteratively solve (PK), with global convergence guarantee [30].

Remark 1: In ADMM framework, (PK) is divided into two simple subproblems of $\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\}$ and $\{\mathbf{T},\mathbf{d},\mathbf{x}\}$. Moreover, due to their separable structures, these two subproblems can be further separated into smaller problems. More interestingly, all these smaller problems have closed-form solutions. Finally, (PK), or the equivalent (PH), can be solved distributively and efficiently. We will elaborate more on this in the following subsections.

B. Update $\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\}$

The problem of $\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\}$ is totally separable among $\mathbf{v}^{\mathrm{U}}_m$, $\mathbf{v}^{\mathrm{D}}_{d,m}$, $\{s^{\mathrm{U}}_{d,m}, y^{\mathrm{U}}_{d,m}, z^{\mathrm{U}}_{d,m}\}$, $\{s^{\mathrm{D}}_{d,m}, y^{\mathrm{D}}_{d,m},\ z^{\mathrm{D}}_{d,m}\}$, $\{s_{c,m}, y_{c,m}, z_{c,m}\}$ for $m\in\mathcal{M}$, and $\{\mathbf{v}^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$. Therefore, we can update the $(5M+1)$ subproblems independently and in parallel.

TABLE V
TWO-BLOCK ADMM FRAMEWORK FOR (PK)

1.	**Repeat**
2.	Update $\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\}$ with other variables fixed: $\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\} = \operatorname{argmin}_{\{\mathbf{V},\mathbf{s},\mathbf{y},\mathbf{z}\}} \mathcal{L}_c(\cdot)$ s.t. (5a), (5b), (21e) and (21f),
3.	Update $\{\mathbf{T},\mathbf{d},\mathbf{x}\}$ with other variables fixed: $\{\mathbf{T},\mathbf{d},\mathbf{x}\} = \operatorname{argmin}_{\{\mathbf{T},\mathbf{d},\mathbf{x}\}} \mathcal{L}_c(\cdot)$ s.t. (13a) and (21a) – (21d),
4.	Update Lagrangian variables $\{\mathbf{\Phi},\boldsymbol{\psi},\boldsymbol{\theta}\}$ [30],
5.	**Until** some stopping criterion is satisfied

1) Update $\mathbf{v}^{\mathrm{U}}_m$, $\mathbf{v}^{\mathrm{D}}_{d,m}$, and $\{\mathbf{v}^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$: The problem of $\mathbf{v}^{\mathrm{U}}_m$ is expressed as

$$\begin{aligned}
(\mathrm{PL}):\ \min_{\mathbf{v}^{\mathrm{U}}_m}\ &\frac{c}{2}\sum_{n\in\mathcal{M}}\begin{pmatrix}\|\mathbf{v}^{\mathrm{U}}_m - \mathbf{t}^{\mathrm{U}}_{d,m}(n) - \frac{1}{c}\boldsymbol{\phi}^{\mathrm{U}}_{d,m}(n)\|^2_2\\ +\|\mathbf{v}^{\mathrm{U}}_m - \mathbf{t}^{\mathrm{U}}_{c,m}(n) - \frac{1}{c}\boldsymbol{\phi}^{\mathrm{U}}_{c,m}(n)\|^2_2\end{pmatrix}\\
\text{s.t.}\ &\|\mathbf{v}^{\mathrm{U}}_m\|^2_2 \leq p^{\mathrm{U}}_m, &(23)
\end{aligned}$$

which is solved as

$$\begin{cases}
\mathbf{v}^{\mathrm{U}}_m = \frac{\sum_{n\in\mathcal{M}}\left[\boldsymbol{\phi}^{\mathrm{U}}_{d,m}(n)+\boldsymbol{\phi}^{\mathrm{U}}_{c,m}(n)+c\left(\mathbf{t}^{\mathrm{U}}_{d,m}(n)+\mathbf{t}^{\mathrm{U}}_{c,m}(n)\right)\right]}{2(cM+\alpha^{\mathrm{U}}_m)}\\
\alpha^{\mathrm{U}}_m = \left[\frac{\left\|\sum_{n\in\mathcal{M}}\left[\boldsymbol{\phi}^{\mathrm{U}}_{d,m}(n)+\boldsymbol{\phi}^{\mathrm{U}}_{c,m}(n)+c\left(\mathbf{t}^{\mathrm{U}}_{d,m}(n)+\mathbf{t}^{\mathrm{U}}_{c,m}(n)\right)\right]\right\|_2}{2\sqrt{p^{\mathrm{U}}_m}} - cM\right]^+
\end{cases} \tag{24}$$

where α^{U}_m is the Lagrangian multiplier for $\|\mathbf{v}^{\mathrm{U}}_m\|^2_2 \leq p^{\mathrm{U}}_m$, and $[\cdot]^+ = \max\{0,\ \cdot\}$.

The problems of $\mathbf{v}^{\mathrm{D}}_{d,m}$ and $\{\mathbf{v}^{\mathrm{D}}_{c,m}\}_{m\in\mathcal{M}}$ can be similarly solved. We omit the details for brevity.

2) Update $\{s^{\mathrm{U}}_{d,m}, y^{\mathrm{U}}_{d,m}, z^{\mathrm{U}}_{d,m}\}$, $\{s^{\mathrm{D}}_{d,m}, y^{\mathrm{D}}_{d,m}, z^{\mathrm{D}}_{d,m}\}$, and $\{s_{c,m}, y_{c,m}, z_{c,m}\}$: The problem related to $\{s^{\mathrm{U}}_{d,m}, y^{\mathrm{U}}_{d,m}, z^{\mathrm{U}}_{d,m}\}$ is expressed as

$$\begin{aligned}
(\mathrm{PM}):\ \min_{\{s^{\mathrm{U}}_{d,m}, y^{\mathrm{U}}_{d,m}, z^{\mathrm{U}}_{d,m}\}}\ &\begin{bmatrix}\frac{c}{2}(y^{\mathrm{U}}_{d,m} - B^{\mathrm{U}}x^{\mathrm{U}}_{d,m} - \frac{1}{c}\psi^{\mathrm{U}}_{d,m})^2\\ +\frac{c}{2}(z^{\mathrm{U}}_{d,m} - d_m - \epsilon + \frac{1}{c}\theta^{\mathrm{U}}_{d,m})^2\\ -2\hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m}\end{bmatrix}\\
\text{s.t.}\quad &y^{\mathrm{U}}_{d,m} \geq \tfrac{(s^{\mathrm{U}}_{d,m})^2}{z^{\mathrm{U}}_{d,m}},\ z^{\mathrm{U}}_{d,m} \geq \epsilon. &(25)
\end{aligned}$$

Exploring the first-order optimality conditions, we obtain

$$\begin{cases}
s^{\mathrm{U}}_{d,m} = \frac{\hat{s}^{\mathrm{U}}_{d,m}z^{\mathrm{U}}_{d,m}}{\beta^{\mathrm{U}}_{d,m}},\\
y^{\mathrm{U}}_{d,m} = \frac{cB^{\mathrm{U}}x^{\mathrm{U}}_{d,m}+\psi^{\mathrm{U}}_{d,m}+\beta^{\mathrm{U}}_{d,m}}{c},\\
z^{\mathrm{U}}_{d,m} = \max\left\{\left[d_m+\epsilon+\frac{(\hat{s}^{\mathrm{U}}_{d,m})^2}{c\beta^{\mathrm{U}}_{d,m}} - \frac{\theta^{\mathrm{U}}_{d,m}}{c}\right],\ \epsilon\right\},
\end{cases} \tag{26}$$

where $\beta^{\mathrm{U}}_{d,m} \geq 0$ is the Lagrangian multiplier of $y^{\mathrm{U}}_{d,m} \geq \frac{(s^{\mathrm{U}}_{d,m})^2}{z^{\mathrm{U}}_{d,m}}$, and should be chosen properly such that the Karush-Kuhn-Tucker (KKT) complementarity conditions [31] are satisfied. Specifically, we define $\beta^*_{d,m} \triangleq \frac{(\hat{s}^{\mathrm{U}}_{d,m})^2}{\theta^{\mathrm{U}}_{d,m}-cd_m}$. If $\beta^*_{d,m} \geq 0$ and $y^{\mathrm{U}}_{d,m} \leq \frac{(s^{\mathrm{U}}_{d,m})^2}{z^{\mathrm{U}}_{d,m}}|_{\beta^{\mathrm{U}}_{d,m}=\beta^*_{d,m}}$, we solve the cubic equation

$$\frac{cB^{\mathrm{U}}x^{\mathrm{U}}_{d,m}+\psi^{\mathrm{U}}_{d,m}+\beta^{\mathrm{U}}_{d,m}}{c} = \frac{(\hat{s}^{\mathrm{U}}_{d,m})^2\epsilon}{(\beta^{\mathrm{U}}_{d,m})^2} \tag{27}$$

$$
\begin{aligned}
\mathcal{L}_c\left(\mathbf{V},\mathbf{d},\mathbf{s},\mathbf{T},\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{\Phi},\boldsymbol{\psi},\boldsymbol{\theta}\right) = & -2\sum_{m\in\mathcal{M}}\left(\hat{s}^{\mathrm{U}}_{d,m}s^{\mathrm{U}}_{d,m}+\hat{s}^{\mathrm{D}}_{d,m}s^{\mathrm{D}}_{d,m}+\hat{s}_{c,m}s_{c,m}\right)\\
&+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{M}}\left\{\begin{array}{r}\mathrm{Re}\left[(\boldsymbol{\phi}^{\mathrm{U}}_{d,n}(m))^{\dagger}(\mathbf{t}^{\mathrm{U}}_{d,n}(m)-\mathbf{v}^{\mathrm{U}}_{n})+(\boldsymbol{\phi}^{\mathrm{U}}_{c,n}(m))^{\dagger}(\mathbf{t}^{\mathrm{U}}_{c,n}(m)-\mathbf{v}^{\mathrm{U}}_{n})\right]\\ +\frac{c}{2}\left[\|\mathbf{t}^{\mathrm{U}}_{d,n}(m)-\mathbf{v}^{\mathrm{U}}_{n}\|_2^2+\|\mathbf{t}^{\mathrm{U}}_{c,n}(m)-\mathbf{v}^{\mathrm{U}}_{n}\|_2^2\right]\end{array}\right\}\\
&+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{M}}\left\{\begin{array}{r}\mathrm{Re}\left[(\boldsymbol{\phi}^{\mathrm{D}}_{d;d,n}(m))^{\dagger}(\mathbf{t}^{\mathrm{D}}_{d;d,n}(m)-\mathbf{v}^{\mathrm{D}}_{d,n})+(\boldsymbol{\phi}^{\mathrm{D}}_{d;c,n}(m))^{\dagger}(\mathbf{t}^{\mathrm{D}}_{d;c,n}(m)-\mathbf{v}^{\mathrm{D}}_{c,n})\right]\\ +\frac{c}{2}\left[\|\mathbf{t}^{\mathrm{D}}_{d;d,n}(m)-\mathbf{v}^{\mathrm{D}}_{d,n}\|_2^2+\|\mathbf{t}^{\mathrm{D}}_{d;c,n}(m)-\mathbf{v}^{\mathrm{D}}_{c,n}\|_2^2\right]\end{array}\right\}\\
&+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{M}}\left\{\begin{array}{r}\mathrm{Re}\left[(\boldsymbol{\phi}^{\mathrm{D}}_{c;d,n}(m))^{\dagger}(\mathbf{t}^{\mathrm{D}}_{c;d,n}(m)-\mathbf{v}^{\mathrm{D}}_{d,n})+(\boldsymbol{\phi}^{\mathrm{D}}_{c;c,n}(m))^{\dagger}(\mathbf{t}^{\mathrm{D}}_{c;c,n}(m)-\mathbf{v}^{\mathrm{D}}_{c,n})\right]\\ +\frac{c}{2}\left[\|\mathbf{t}^{\mathrm{D}}_{c;d,n}(m)-\mathbf{v}^{\mathrm{D}}_{d,n}\|_2^2+\|\mathbf{t}^{\mathrm{D}}_{c;c,n}(m)-\mathbf{v}^{\mathrm{D}}_{c,n}\|_2^2\right]\end{array}\right\}\\
&+\sum_{m\in\mathcal{M}}\left\{\psi^{\mathrm{U}}_{d,m}\left(B^{\mathrm{U}}x^{\mathrm{U}}_{d,m}-y^{\mathrm{U}}_{d,m}\right)+\psi^{\mathrm{D}}_{d,m}\left(B^{\mathrm{D}}x^{\mathrm{D}}_{d,m}-y^{\mathrm{D}}_{d,m}\right)+\frac{c}{2}\left[\left(B^{\mathrm{U}}x^{\mathrm{U}}_{d,m}-y^{\mathrm{U}}_{d,m}\right)^2+\left(B^{\mathrm{D}}x^{\mathrm{D}}_{d,m}-y^{\mathrm{D}}_{d,m}\right)^2\right]\right\}\\
&+\sum_{m\in\mathcal{M}}\left\{\psi^{\mathrm{U}}_{c,m}\left(B^{\mathrm{U}}x^{\mathrm{U}}_{c,m}-y_{c,m}\right)+\psi^{\mathrm{D}}_{c,m}\left(B^{\mathrm{D}}x^{\mathrm{D}}_{c,m}-y_{c,m}\right)+\frac{c}{2}\left[\left(B^{\mathrm{U}}x^{\mathrm{U}}_{c,m}-y_{c,m}\right)^2+\left(B^{\mathrm{D}}x^{\mathrm{D}}_{c,m}-y_{c,m}\right)^2\right]\right\}\\
&+\sum_{m\in\mathcal{M}}\left\{\begin{array}{r}\theta^{\mathrm{U}}_{d,m}(z^{\mathrm{U}}_{d,m}-d_m-\epsilon)+\theta^{\mathrm{D}}_{d,m}(z^{\mathrm{D}}_{d,m}-d_m-\epsilon)+\theta_{c,m}(z_{c,m}-1+d_m-\epsilon)\\ +\frac{c}{2}\left[(z^{\mathrm{U}}_{d,m}-d_m-\epsilon)^2+(z^{\mathrm{D}}_{d,m}-d_m-\epsilon)^2+(z_{c,m}-1+d_m-\epsilon)^2\right]\end{array}\right\}
\end{aligned}
\tag{22}
$$

to find the optimal $\beta^{\mathrm{U}}_{d,m}$; otherwise, we solve the following quartic equation

$$
\frac{cB^{\mathrm{U}}x^{\mathrm{U}}_{d,m}+\psi^{\mathrm{U}}_{d,m}+\beta^{\mathrm{U}}_{d,m}}{c}=\frac{(\hat{s}^{\mathrm{U}}_{d,m})^2}{(\beta^{\mathrm{U}}_{d,m})^2}\cdot\left[d_m+\epsilon+\frac{(\hat{s}^{\mathrm{U}}_{d,m})^2}{c\beta^{\mathrm{U}}_{d,m}}-\frac{\theta^{\mathrm{U}}_{d,m}}{c}\right] \tag{28}
$$

to get the optimal $\beta^{\mathrm{U}}_{d,m}$. Since both cubic and quartic equations can be solved analytically, the closed-form solution to $\beta^{\mathrm{U}}_{d,m}$ is achievable.

The problems of $\{s^{\mathrm{D}}_{d,m},y^{\mathrm{D}}_{d,m},z^{\mathrm{D}}_{d,m}\}$ and $\{s_{c,m},y_{c,m},z_{c,m}\}$ can be solved similarly. We omit the details for brevity.

C. Update $\{\mathbf{T},\mathbf{d},\mathbf{x}\}$

It can be easily observed that updating $\{\mathbf{T},\mathbf{d},\mathbf{x}\}$ is completely separable among $\{\mathbf{T}^{\mathrm{U}}_d(m),x^{\mathrm{U}}_{d,m}\}$, $\{\mathbf{T}^{\mathrm{D}}_d(m),\ x^{\mathrm{D}}_{d,m}\}$, $\{\mathbf{T}^{\mathrm{U}}_c(m),x^{\mathrm{U}}_{c,m}\}$, $\{\mathbf{T}^{\mathrm{D}}_c(m),x^{\mathrm{D}}_{c,m}\}$ and $\{d_m\}$, $\forall m\in\mathcal{M}$. Then, we update the $5M$ subproblems independently and in parallel.

1) Update $\{\mathbf{T}^{\mathrm{U}}_d(m),x^{\mathrm{U}}_{d,m}\},\{\mathbf{T}^{\mathrm{D}}_d(m),x^{\mathrm{D}}_{d,m}\},\{\mathbf{T}^{\mathrm{U}}_c(m),x^{\mathrm{U}}_{c,m}\}$, and $\{\mathbf{T}^{\mathrm{D}}_c(m),x^{\mathrm{D}}_{c,m}\}$: The problem related to $\{\mathbf{T}^{\mathrm{U}}_d(m),x^{\mathrm{U}}_{d,m}\}$ is expressed as

$$
\begin{aligned}
(\mathrm{PN}):\ &\min_{\{\mathbf{T}^{\mathrm{U}}_d(m),x^{\mathrm{U}}_{d,m}\}}\ \frac{c}{2}\left[\begin{array}{r}\sum_{n\in\mathcal{M}}\|\mathbf{t}^{\mathrm{U}}_{d,n}(m)-\mathbf{v}^{\mathrm{U}}_{n}+\frac{1}{c}\boldsymbol{\phi}^{\mathrm{U}}_{d,n}(m)\|_2^2\\ +\left(B^{\mathrm{U}}x^{\mathrm{U}}_{d,m}-y^{\mathrm{U}}_{d,m}-\frac{1}{c}\psi^{\mathrm{U}}_{d,m}\right)^2\end{array}\right]\\
&\mathrm{s.t}\quad \log(w^{\mathrm{U}}_{d,m})-w^{\mathrm{U}}_{d,m}e^{\mathrm{U}}_{d,m}(\mathbf{T}^{\mathrm{U}}_d(m),\mathbf{u}^{\mathrm{U}}_{d,m})+1\geq x^{\mathrm{U}}_{d,m}.
\end{aligned}
\tag{29}
$$

It is solved as

$$
\left\{\begin{array}{l}
\mathbf{t}^{\mathrm{U}}_{d,m}(m)=[2\delta^{\mathrm{U}}_{d,m}w^{\mathrm{U}}_{d,m}(\mathbf{F}^{\mathrm{U}}_{m,m})^{\dagger}\mathbf{u}^{\mathrm{U}}_{d,m}(\mathbf{u}^{\mathrm{U}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{U}}_{m,m}+c\mathbf{I}]^{-1}\\
\qquad\qquad\times[c\mathbf{v}^{\mathrm{U}}_{m}-\boldsymbol{\phi}^{\mathrm{U}}_{d,m}(m)+2\delta^{\mathrm{U}}_{d,m}w^{\mathrm{U}}_{d,m}(\mathbf{F}^{\mathrm{U}}_{m,m})^{\dagger}\mathbf{u}^{\mathrm{U}}_{d,m}]\\
\mathbf{t}^{\mathrm{U}}_{d,n}(m)=[2\delta^{\mathrm{U}}_{d,m}w^{\mathrm{U}}_{d,m}(\mathbf{F}^{\mathrm{U}}_{m,n})^{\dagger}\mathbf{u}^{\mathrm{U}}_{d,m}(\mathbf{u}^{\mathrm{U}}_{d,m})^{\dagger}\mathbf{F}^{\mathrm{U}}_{m,n}+c\mathbf{I}]^{-1}\\
\qquad\qquad\times[c\mathbf{v}^{\mathrm{U}}_{n}-\boldsymbol{\phi}^{\mathrm{U}}_{d,n}(m)],\ \forall n\neq m\\
x^{\mathrm{U}}_{d,m}=\frac{cB^{\mathrm{U}}y^{\mathrm{U}}_{d,m}+B^{\mathrm{U}}\psi^{\mathrm{U}}_{d,m}-\delta^{\mathrm{U}}_{d,m}}{c(B^{\mathrm{U}})^2}
\end{array}\right.
\tag{30}
$$

where $\delta^{\mathrm{U}}_{d,m}\geq 0$ is the associated Lagrangian multiplier and should be properly chosen such that the KKT complementarity conditions are satisfied. That is, if (29) holds for $\delta^{\mathrm{U}}_{d,m}=0$, we have $\delta^{\mathrm{U}}_{d,m}=0$; otherwise, we choose some $\delta^{\mathrm{U}}_{d,m}>0$ such that (29) holds for equality, which can be easily done by bisection search.

Similarly, the problems of $\{\mathbf{T}^{\mathrm{D}}_d(m),x^{\mathrm{D}}_{d,m}\}$, $\{\mathbf{T}^{\mathrm{U}}_c(m),x^{\mathrm{U}}_{c,m}\}$ and $\{\mathbf{T}^{\mathrm{D}}_c(m),x^{\mathrm{D}}_{c,m}\}$ can be solved.

2) Update d_m: The problem of d_m is

$$
\begin{aligned}
(\mathrm{PO}):\ &\min_{d_m}\left[\begin{array}{l}(\theta_{c,m}-\theta^{\mathrm{U}}_{d,m}-\theta^{\mathrm{D}}_{d,m})d_m+\frac{c}{2}(d_m+\epsilon-z^{\mathrm{U}}_{d,m})^2\\ +\frac{c}{2}[(d_m+\epsilon-z^{\mathrm{D}}_{d,m})^2+(d_m-1-\epsilon+z_{c,m})^2]\end{array}\right]\\
&\mathrm{s.t.}\ \ 0\leq d_m\leq 1,
\end{aligned}
\tag{31}
$$

which is solved as

$$
d_m=\left[\frac{c\left(z^{\mathrm{U}}_{d,m}+z^{\mathrm{D}}_{d,m}-z_{c,m}+1-\epsilon\right)-(\theta_{c,m}-\theta^{\mathrm{U}}_{d,m}-\theta^{\mathrm{D}}_{d,m})}{3c}\right]_0^1 \tag{32}
$$

where $[\,\cdot\,]_0^1$ denotes the projection onto the range of $[0,1]$.

Since (PH) can be solved distributively, we obtain the fully distributed IA-IWMMSE by updating $\{\mathbf{V},\mathbf{d},\mathbf{s}\}$ based on the ADMM algorithm in Table V.

Remark 2: The complexity comparison of the ADMM algorithm and the IP algorithm for (PH) is listed in Table VI. With the problem dimension being $M(N_u+N_b)$, the per-iteration complexity of the IP algorithm is $\mathcal{O}(M^3(N_u+N_b)^3)$. By contrast, the complexity of the ADMM algorithm is dominated by the update of $\{\mathbf{T},\mathbf{d},\mathbf{x}\}$. As shown in (30), the per-iteration complexity is $\mathcal{O}(\max\{M^2N_u^2,M^2N_uN_b\})$ if we use the rank-one update rule to compute the matrix inverse. Obviously, the ADMM algorithm is more efficient.

V. Numerical Examples

Consider a network consisting of one multi-antenna BS and $M=10$ pairs of multi-antenna UEs. The BS is located at the

TABLE VI
COMPLEXITY COMPARISON OF ADMM ALGORITHM AND IP ALGORITHM

	Per-Iteration Complexity
IP Algorithm	$\mathcal{O}(M^3(N_u+N_b)^3)$
ADMM Algorithm	$\mathcal{O}(\max\{M^2N_u^2, M^2N_uN_b\})$

center of a hexagonal cell with the side length being $d = 1$ km, and the UEs are randomly deployed in the cell. We assume the Rayleigh channel with zero mean and variance $L(200/r)^3$, with r being the distance between transmitter and receiver, and L being the shadowing effect satisfying $10\log_{10}(L) \sim \mathcal{N}(0, 64)$. We assume that all TX_UEs have same transmit power budgets in uplink and downlink, i.e., $p_1^{\mathrm{U}} = p_2^{\mathrm{U}} = \cdots = p_M^{\mathrm{U}}$, $p_1^{\mathrm{D}} = p_2^{\mathrm{D}} = \cdots = p_M^{\mathrm{D}}$. The background noise spectral density is -174 dBm/Hz, and the noise power depends on the bandwidth values, e.g., B^{U} and B^{D}. The parameter ϵ is set as 10^{-6} in the following simulations.

We first show the convergence behaviour of the ADMM algorithm for (PH). To this end, we solve (PH) by CVX and then solve the equivalent (PK) by ADMM under the same network configurations. The typical convergence traces (starting from the same initial point) are shown in Fig. 2. It is clear that the ADMM algorithm can converge to the CVX solution, while the convergence behaviour is sensitive to the value of c. In our simulation, the ADMM algorithm typically converges in about 200 iterations with $c = 10$.

Next, we validate the efficiency advantage of the ADMM algorithm against CVX (the IP algorithm) for (PH). In Fig. 3, we compare the CPU running times of the two approaches at different UE pair numbers, which are obtained by running the simulations on a computer with i5-3320M CPU and 8 GB RAM. The results show that the ADMM algorithm has higher efficiency regardless of the UE pair number M, owing to the simple closed-form solutions in each step. We should mention here that the results in Fig. 3 are not consistent with those in Table VI. This can be explained as follows. In Table VI we actually compare the per-iteration complexities of the two algorithms, while in Fig. 3 we show the total CPU times they need to solve (PH). The mismatch between them is mainly due to the fact that the two algorithms require different numbers of iterations to achieve the optimal solution of (PH).

In Fig. 4, we compare the convergence behaviours of IA-IWMMSE and IA-EWMMSE. In IA-EWMMSE, we update $\hat{\mathbf{s}}$ only after (PG) has been exactly solved (i.e., at the red circles in Fig. 4). Notice that updating $\hat{\mathbf{s}}$ yields apparent performance improvement, especially in the first few DC programming iterations. Generally, IA-EWMMSE converges after 250 WMMSE iterations. By contrast, we update $\hat{\mathbf{s}}$ in each WMMSE iteration in IA-IWMMSE, and thereby the most of performance improvement can be achieved in the first few iterations. As a consequence, IA-IWMMSE is more efficient and converges faster (typically in 100 WMMSE iterations). It can also be observed in Fig. 4 that the performance of IA-IWMMSE is better than that of IA-EWMMSE. This is due to the fact that the two algorithms may converge to different stationary solutions. Since (PF) is non-convex, it is difficult

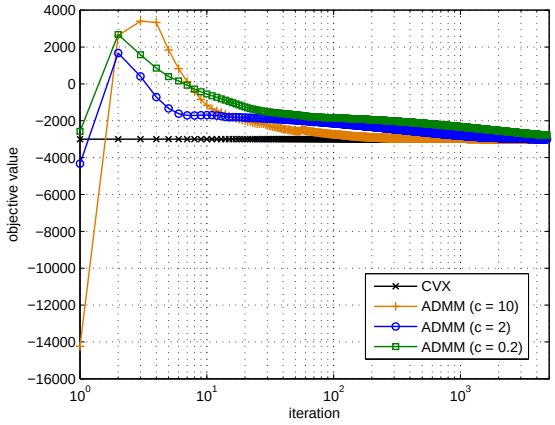


Fig. 2. Typical convergence traces of the ADMM algorithm for (PH), with $M = 10$, $N_b = 3$, and $N_u = 2$.

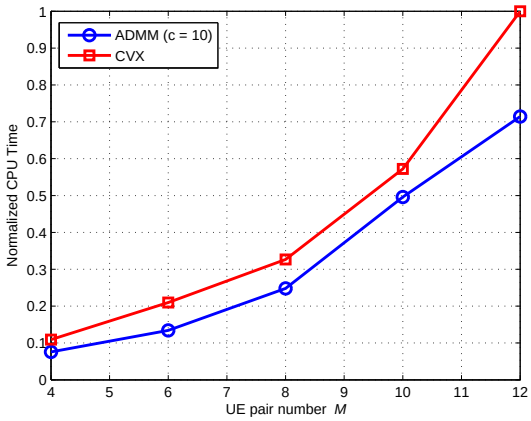


Fig. 3. CPU time comparison at different UE pair numbers M, with $N_b = 3$ and $N_u = 2$.

to achieve the global optimum. Both IA-EWMMSE and IA-IWWMSE are only guaranteed to converge to a stationary solution (local optimum) of (PF). The two algorithms have the same computational steps, while the order of execution is different (see Tables II and III). In consequence, they have different convergence traces and finally reach different local optimums. However, since the objective function of (PF) depends on the random D2D and cellular channels, there is no theoretical guarantee that one local optimum will be better than the other. This is confirmed in Fig. 5, where we record the sum rates and iteration numbers for convergence of the two algorithms with 50 different channels. Interestingly, although the two algorithms may behave differently in each simulation trial, their average performances are close to each other. The advantage of IA-IWMMSE against IA-EWMMSE mainly lies in the simplified algorithm structure (see Table III) and the accelerated convergence behaviour (see Figs. 4 and 5).

After that, we compare IA-IWMMSE with the following approaches to show its advantages in sum-rate maximization. (1) IA-EWMMSE. (2) All-cellular approach; i.e., all UE pairs communicate through BS. (3) All-D2D approach; i.e., all

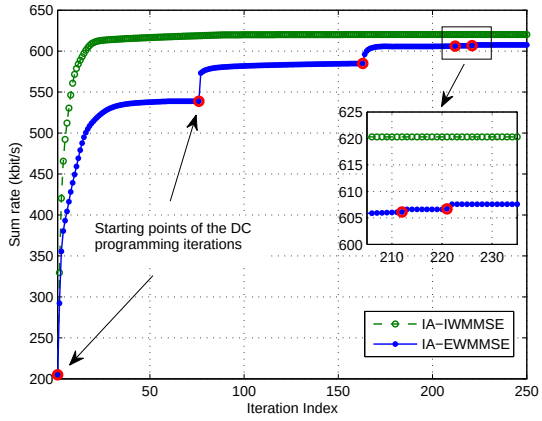


Fig. 4. Typical convergence traces of IA-IWMMSE and IA-EWMMSE, with $M = 10$, $N_b = 3$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, and $p_m^{\mathrm{D}} = p_m^{\mathrm{U}} = 0$ dBm, $\forall\, m$.

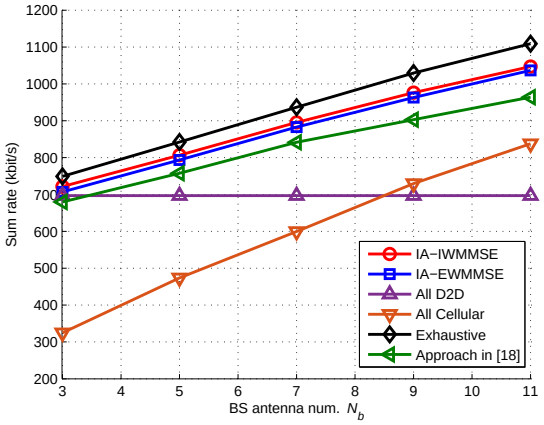


Fig. 6. Sum-rate comparison at different BS antenna numbers N_b, with $M = 10$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{U}} = 0$ dBm, $p_m^{\mathrm{D}} = 0$ dBm, $\forall\, m$.

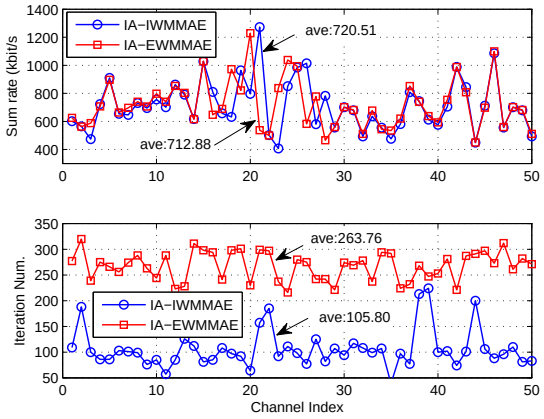


Fig. 5. Comparison of IA-IWMMSE and IA-EWMMSE for 50 different channels with $M = 10$, $N_b = 3$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{D}} = p_m^{\mathrm{U}} = 0$ dBm, $\forall\, m$.

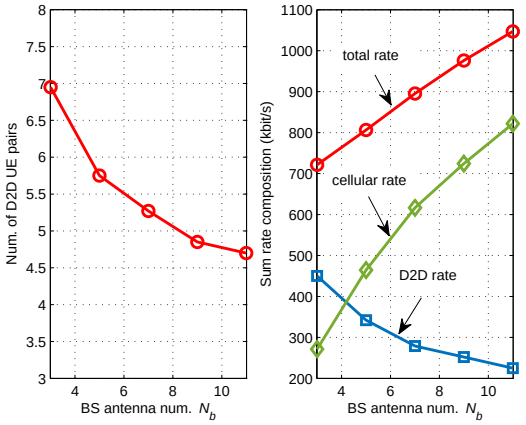


Fig. 7. Some details of IA-IWMMSE at different BS antenna numbers N_b, with $M = 10$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{U}} = 0$ dBm, $p_m^{\mathrm{D}} = 0$ dBm, $\forall\, m$.

UE pairs communicate via the D2D link. (4) Exhaustive search approach, which compares the throughput results of all possible transmission mode settings and select the best one. This approach is computationally intolerable in practice, yet its results can be used to show the gap between our approximate solution and the optimum. (5) The approach in [18], which considers the joint mode selection and transceiver design for sum rate maximization in MIMO D2D networks, and solves it by utilizing the method of SCA. Besides the solution approach, our work differs from [18] mainly in that we employ an exclusive (i.e., either cellular or D2D) mode selection strategy for each UE pair, while [18] allows simultaneous cellular and D2D transmission (i.e., it does not use d_m when computing the achievable rate of the mth UE pair, see R_m in (8)) to further improve the network throughput, which, of course, raises heavy operational burden and signaling cost in practice. To avoid unfair comparison, as simultaneous cellular and D2D transmission occurs for some UE pair in [18], we force it to select the mode contributing more to the throughput.

In Fig. 6, we compare the sum rates of different approaches at distinct BS antenna numbers N_b. Among these approaches, the all-D2D approach works independently of the BS antenna number, while the other approaches achieve higher throughput as N_b increases, since more space diversity can be exploited. The exhaustive search approach surely achieves the highest sum rate by exploring all the mode selection possibilities. By performing the system-level optimization on transmission mode selection and transceiver design, IA-IWMMSE and IA-EWMMSE achieve the sum rates close to that of the exhaustive search approach and outperform the other approaches. The performance gap between IA-IWMMSE and the exhaustive search approach is mainly due to the relaxation of d_m and the first-order approximation in solving (PB). Although the approach in [18] also jointly optimizes the transmission mode and the associated transceivers for each UE pair, it is not designed for the exclusive mode selection strategy. Consequently, when required to select only one kind of transmission for each UE pair, it suffers from performance degradation because the mode selection may not be optimal. When all the UE pairs communicate through BS, the network tends to be congested.

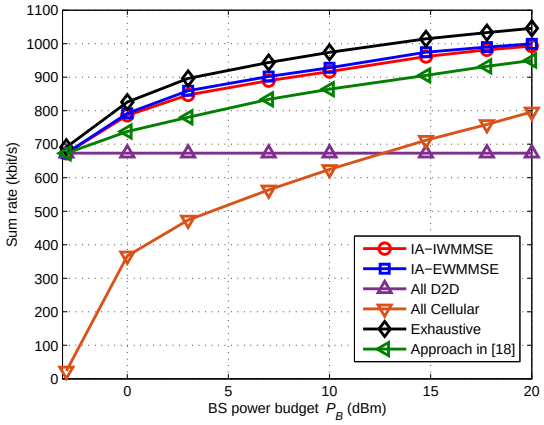


Fig. 8. Sum-rate comparison at different BS power budgets P_B, with $M = 10$, $N_b = 10$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $p_m^{\mathrm{U}} = -3$ dBm, $p_m^{\mathrm{D}} = -3$ dBm, $\forall\, m$.

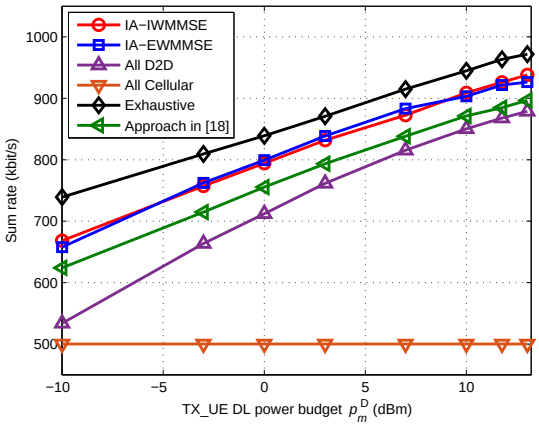


Fig. 10. Sum-rate comparison at different UE downlink power budgets p_m^{D}, with $M = 10$, $N_b = 5$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{U}} = 0$ dBm, $\forall\, m$.

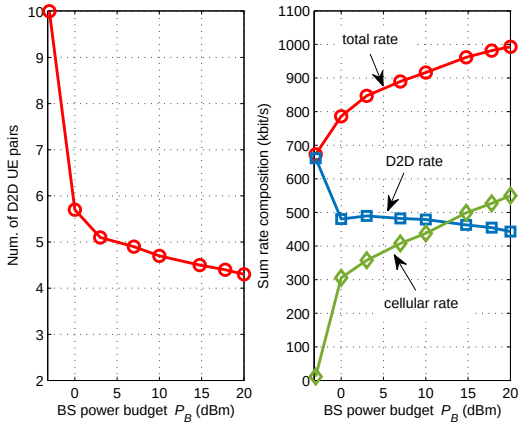


Fig. 9. Some details of IA-IWMMSE at different BS power budgets P_B, with $M = 10$, $N_b = 10$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $p_m^{\mathrm{U}} = -3$ dBm, $p_m^{\mathrm{D}} = -3$ dBm, $\forall\, m$.

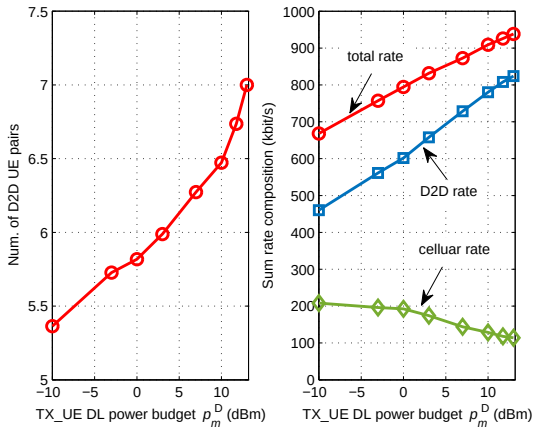


Fig. 11. Some details of IA-IWMMSE at different UE DL power budgets p_m^{D}, with $M = 10$, $N_b = 5$, $N_u = 2$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{U}} = 0$ dBm, $\forall\, m$

Therefore, the all-cellular approach has the worst performance in the case of small N_b. When N_b increases, the congestion will be alleviated, and hence it gradually outperforms the all-D2D approach. In Fig. 7, we further show some details of IA-IWMMSE. As N_b increases, the extra space diversity in cellular transmission drives more UE pairs to select the cellular mode. Then, the D2D UE pair number decreases and the traffic through BS increases.

We compare the sum rates of these approaches at different BS power budgets P_B in Fig. 8, and show some details of IA-IWMMSE in Fig. 9. In the case of very small BS power budget (e.g., $P_B = -3$ dBm), IA-IWMMSE suggests that all the UE pairs should select the D2D mode. Therefore, IA-IWMMSE and the all-D2D approach have the same performance. IA-EWMMSE and the approach in [18] achieve similar performance at this point, implying that they also select the D2D mode for all UE pairs. The above approaches outperform the all-cellular approach when P_B is small. As P_B increases, the D2D UE pairs are more likely to select the cellular mode for higher rate. As a consequence, the D2D UE pair number decreases and the sum rate increases. The order of achievable sum rate in this simulation is the same as that in Fig. 6 due to the reason aforementioned. We emphasize that IA-IWMMSE and IA-EWMMSE perform close to the exhaustive approach and outperform the other approaches.

We compare the sum rates of these approaches at different UE downlink power budgets p_m^{D} in Fig. 10. In this simulation, the sum rate of the all-cellular approach is independent of p_m^{D}, while the sum rates of other approaches increase with p_m^{D} since they all involve D2D transmission. The descending order of these approaches in terms of sum rate is the exhaustive search approach, IA-IWMMSE (IA-EWMMSE), the approach in [18], the all-D2D approach, and the all-cellular approach. In Fig. 11, we show more details of IA-IWMMSE. In cellular mode, the rate of one UE pair is determined actually by the smaller one of uplink rate and downlink rate, while in D2D mode the rate is the sum of them. Therefore, increasing the UE downlink power budget improves the rates of D2D UE pairs, thereby encouraging UE pairs to select the D2D mode. In consequence, the D2D UE pair number and the sum rate

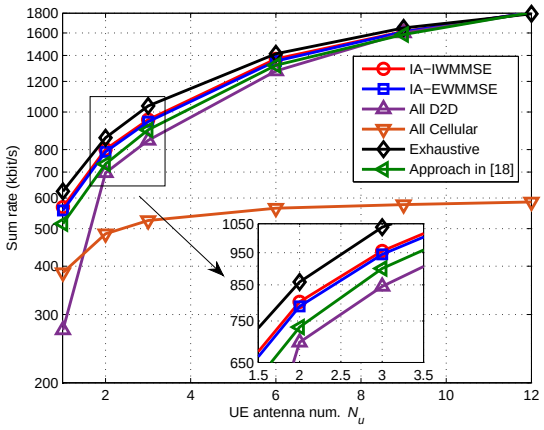


Fig. 12. Sum-rate comparison at different UE antenna number N_u, with $M = 10$, $N_b = 5$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{U}} = 0$ dBm, $p_m^{\mathrm{D}} = 0$ dBm, $\forall\, m$.

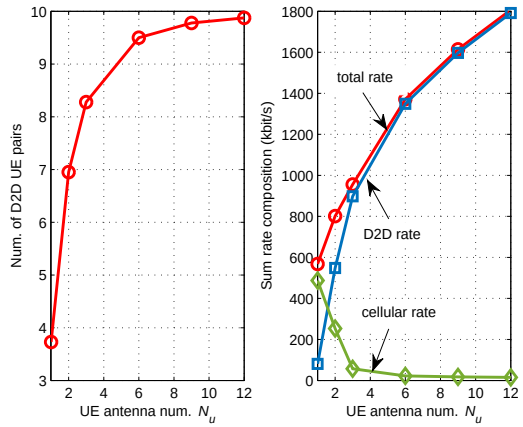


Fig. 13. Some details of IA-IWMMSE at different UE antenna number N_u, with $M = 10$, $N_b = 5$, $B^{\mathrm{U}} = B^{\mathrm{D}} = 15$ kHz, $P_B = 15$ dBm, $p_m^{\mathrm{U}} = 0$ dBm, $p_m^{\mathrm{D}} = 0$ dBm, $\forall\, m$.

increase with p_m^{D}. The proportion of D2D rate in the sum rate increases with p_m^{D} also.

We compare the sum rates of these approaches at different UE antenna numbers N_u in Fig. 12, and show more details of IA-IWMMSE in Fig. 13. As N_u increases, all the approaches in comparison achieve higher sum rates. However, the performance of the all-cellular approach grows far more slowly compared with the other approaches, especially in the case of large N_u. This is reasonable since when N_u is large, the BS antenna number N_b becomes the bottleneck of throughput improvement in cellular transmission, thus limiting the performance of the all-cellular approach. By contrast, the D2D transmission benefits from the extra space diversity introduced by increasing N_u. Then, the other approaches can effectively improve the network throughput by letting more UE pairs to select the D2D mode. Again, among these approaches, IA-IWMMSE achieves the performance close to that of exhaustive search and outperforms the others. It can be observed that both the number of D2D UE pairs and the traffic through D2D link increase with N_u in IA-IWMMSE. Moreover, in the case of $N_u = 12$, the evident advantage of D2D transmission in space diversity drives all the UE pairs to communicate directly, and hence the sum rate curves of all these approaches (except the all-cellular approach) coincide.

VI. Conclusion

In this paper, we consider the joint mode selection and transceiver design problem in underlay D2D multi-user MIMO networks for sum rate maximization. Different from most current studies, we assume that the cellular and D2D transmissions are equally prioritized, and each UE pair can switch between cellular mode and D2D mode freely. We develop a low-complexity algorithm, based on inexact WMMSE, to iteratively compute a stationary solution of the sum-rate maximization problem. To facilitate the algorithm's implementation, we further apply the ADMM framework to decouple the WMMSE problem. Finally, an efficient distributed algorithm is devised, with a simple closed-form update in each step. The efficacy and efficiency of the proposed algorithm have been demonstrated by extensive numerical simulations.

Appendix A
Proof of Proposition 1

Theoretically, (PA) and (PB) can be solved by the method of exhaustive search; i.e., we compare the throughput results of all possible transmission mode settings and select the best one. Therefore, to establish the equivalence between (PA) and (PB), it is sufficient to show that their resultant problems with the same transmission mode setting are equivalent.

The key step in proof is to show that if UE pair m works in D2D mode and UE pair n works in cellular mode, i.e., $d_m = 1$ and $d_n = 0$, then we have $\mathbf{v}_{c,m}^{\mathrm{D}} = \mathbf{0}$ and $\mathbf{v}_{d,n}^{\mathrm{D}} = \mathbf{0}$ at the optimal solution of the resultant problem of (PB). This can be proven by contradiction.

Specifically, assume that UE pair m works in D2D mode, and $\{\mathbf{V}^*, \mathbf{U}^*\}$ is the optimal solution of the resultant problem of (PB) with $\mathbf{v}_{c,m}^{\mathrm{D}} \neq \mathbf{0}$. Then, we can get another feasible solution of the problem by simply letting $\mathbf{v}_{c,m}^{\mathrm{D}} = \mathbf{0}$, which is denoted as $\{\bar{\mathbf{V}}^*, \mathbf{U}^*\}$. From the downlink SINR expressions of (6a) and (6b), we conclude that letting $\mathbf{v}_{c,m}^{\mathrm{D}} = \mathbf{0}$ eliminates the cellular interference from TX_UE$_m$ to all the other downlink transmissions, and then improves $\overline{\mathrm{SINR}}_{d,n}^{\mathrm{D}}, \forall\ n \in \mathcal{M}$, and $\overline{\mathrm{SINR}}_{c,n}^{\mathrm{D}}, \forall\ n \in \mathcal{M}, n \neq m$. According to (7a) and (7b), this produces increased $\bar{R}_{d,n}, \forall\ n \in \mathcal{M}$, for the M D2D transmissions, and non-decreased $\bar{R}_{c,n}, \forall\ n \in \mathcal{M}, n \neq m$, for the $(M-1)$ cellular transmissions. In other words, among the $2M$ rate terms of $\{\bar{R}_{d,n}, \bar{R}_{c,n}\}_{n\in\mathcal{M}}$, only $\bar{R}_{c,m}$ decreases to zero. Recall that UE pair m works in D2D mode, i.e., $d_m = 1$, then from (8) we know that $\bar{R}_{c,m}$ does not count for the throughput. Hence, we claim that $\{\mathbf{V}^*, \mathbf{U}^*\}$ is *not* the optimal solution since $\{\bar{\mathbf{V}}^*, \mathbf{U}^*\}$ yields higher throughput. There comes the contradiction. Similarly, we can show that if UE pair n works in D2D mode, we have $\mathbf{v}_{d,n}^{\mathrm{D}} = \mathbf{0}$ at the optimal solution of the resultant problem of (PB).

In conclusion, for the same transmission mode setting, the resultant problems of (PA) and (PB) are equivalent in the sense that they have the same optimal solutions due to the

same SINR and rate expressions. Based on this, we further claim the equivalence between (PA) and (PB).

Appendix B
Proof of Proposition 2

Using subscript k to denote the iteration index, we express the updating rules in each iteration of IA-IWMMSE as

$$\{\mathbf{V}^{(k+1)}, \mathbf{d}^{(k+1)}, \mathbf{s}^{(k+1)}\} \in \Omega(\mathbf{U}^{(k)}, \mathbf{W}^{(k)}; \hat{\mathbf{s}}^{(k)}), \tag{33a}$$

$$\mathbf{U}^{(k+1)} = \Gamma(\mathbf{V}^{(k+1)}), \tag{33b}$$

$$\mathbf{W}^{(k+1)} = \Upsilon(\mathbf{V}^{(k+1)}, \mathbf{U}^{(k+1)}), \tag{33c}$$

$$\hat{\mathbf{s}}^{(k+1)} = \mathbf{s}^{(k+1)}. \tag{33d}$$

We next define the following two objective functions

$$Q(\mathbf{d}, \mathbf{s}, \mathbf{V}; \mathbf{U}, \mathbf{W}) \triangleq -\sum_{m\in\mathcal{M}} (s_{d,m}^{\mathrm{U}})^2 + (s_{d,m}^{\mathrm{D}})^2 + s_{c,m}^2 \tag{34a}$$

$$\hat{Q}(\mathbf{d}, \mathbf{s}, \mathbf{V}; \mathbf{U}, \mathbf{W}, \hat{\mathbf{s}}) \triangleq \sum_{m\in\mathcal{M}} \begin{bmatrix} (\hat{s}_{d,m}^{\mathrm{U}})^2 + (\hat{s}_{d,m}^{\mathrm{D}})^2 + \hat{s}_{c,m}^2 \\ -2\hat{s}_{d,m}^{\mathrm{U}} s_{d,m}^{\mathrm{U}} - 2\hat{s}_{d,m}^{\mathrm{D}} s_{d,m}^{\mathrm{D}} \\ -2\hat{s}_{c,m} s_{c,m} \end{bmatrix} \tag{34b}$$

where $\hat{Q}(\mathbf{d}, \mathbf{s}, \mathbf{V}; \mathbf{U}, \mathbf{W}, \hat{\mathbf{s}})$ is the first-order approximation of $Q(\mathbf{d}, \mathbf{s}, \mathbf{V}; \mathbf{U}, \mathbf{W})$ at $\mathbf{s} = \hat{\mathbf{s}}$; Q is the objective of (PF), and $\hat{Q}$ is a *general form* of the objective of (PG).

Since the WMMSE iterations produce monotonically decreasing objective values (even if only one WMMSE iteration is executed), and the first-order Taylor expansion is a global overestimate of any concave function, we claim that IA-IWMMSE generates monotonically decreasing objective sequence (see (36) and the following discussion). In addition, due to the compactness of the feasible set of $\{\mathbf{V}, \mathbf{d}\}$ and the continuous mapping among $\mathbf{V}$, $\mathbf{U}$ and $\mathbf{W}$, the objective sequence is lower bounded and thus converges.

We denote $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\}$ as a limit point of the iterates generated by IA-IWMMSE, and Q^* as the corresponding objective value. Denote $\{\mathbf{V}^{(k_t)}, \mathbf{d}^{(k_t)}, \mathbf{s}^{(k_t)}, \mathbf{U}^{(k_t)}, \mathbf{W}^{(k_t)}, \hat{\mathbf{s}}^{(k_t)}\}_{t=1}^{\infty}$ as a subsequence converging to this limit point. We show that $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*\}$ is a stationary solution of (PE) by the following two Lemmas.

Lemma 1: If $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\}$ is a limit point of the iterates generated by IA-IWMMSE, then $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*\}$ is a stationary solution of (PG) with $\hat{\mathbf{s}}$ replaced by $\hat{\mathbf{s}}^*$.

proof: We establish this by showing that $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*\}$ is a KKT point of (PG). To this end, we need to prove that $\mathbf{U}^* = \Gamma(\mathbf{V}^*)$, $\mathbf{W}^* = \Upsilon(\mathbf{V}^*, \mathbf{U}^*)$, and $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*\} \in \Omega(\mathbf{U}^*, \mathbf{W}^*; \hat{\mathbf{s}}^*)$. According to the updating rules in each IA-IWMMSE iteration, i.e., (33), we directly have $\mathbf{U}^* = \Gamma(\mathbf{V}^*)$, $\mathbf{W}^* = \Upsilon(\mathbf{V}^*, \mathbf{U}^*)$, and $\hat{\mathbf{s}}^* = \mathbf{s}^*$, i.e.,

$$\begin{aligned} &\lim_{t\to\infty} \{\mathbf{V}^{(k_t)}, \mathbf{d}^{(k_t)}, \mathbf{s}^{(k_t)}, \mathbf{U}^{(k_t)}, \mathbf{W}^{(k_t)}, \hat{\mathbf{s}}^{(k_t)}\} \\ &= \{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\} \\ &= \{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \Gamma(\mathbf{V}^*), \Upsilon(\mathbf{V}^*, \mathbf{U}^*), \mathbf{s}^*\}. \end{aligned} \tag{35}$$

We next show that $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*\} \in \Omega(\mathbf{U}^*, \mathbf{W}^*; \hat{\mathbf{s}}^*)$. Basically, we have

$$\begin{aligned} &\hat{Q}\left(\mathbf{d}, \mathbf{s}, \mathbf{V}; \mathbf{U}^{(k_t)}, \mathbf{W}^{(k_t)}, \hat{\mathbf{s}}^{(k_t)}\right) \\ &\geq \hat{Q}\left(\mathbf{d}^{(k_{t+1})}, \mathbf{s}^{(k_{t+1})}, \mathbf{V}^{(k_{t+1})}; \mathbf{U}^{(k_t)}, \mathbf{W}^{(k_t)}, \hat{\mathbf{s}}^{(k_t)}\right) \\ &\geq \hat{Q}\left(\mathbf{d}^{(k_{t+1})}, \mathbf{s}^{(k_{t+1})}, \mathbf{V}^{(k_{t+1})}; \mathbf{U}^{(k_{t+1})}, \mathbf{W}^{(k_{t+1})}, \hat{\mathbf{s}}^{(k_t)}\right) \\ &\geq Q\left(\mathbf{d}^{(k_{t+1})}, \mathbf{s}^{(k_{t+1})}, \mathbf{V}^{(k_{t+1})}; \mathbf{U}^{(k_{t+1})}, \mathbf{W}^{(k_{t+1})}\right) \\ &= \hat{Q}\left(\mathbf{d}^{(k_{t+1})}, \mathbf{s}^{(k_{t+1})}, \mathbf{V}^{(k_{t+1})}; \mathbf{U}^{(k_{t+1})}, \mathbf{W}^{(k_{t+1})}, \hat{\mathbf{s}}^{(k_{t+1})}\right). \end{aligned} \tag{36}$$

where the first inequality is due to the fact that $\{\mathbf{V}^{(k_{t+1})}, \mathbf{d}^{(k_{t+1})}, \mathbf{s}^{(k_{t+1})}\}$ is the global optimal solution of (PH), which minimizes $\hat{Q}(\cdot)$ with respect to $\{\mathbf{V}, \mathbf{d}, \mathbf{s}\}$ at $\mathbf{U} = \mathbf{U}^{(k_t)}$, $\mathbf{W} = \mathbf{W}^{(k_t)}$, and $\hat{\mathbf{s}} = \hat{\mathbf{s}}^{(k_t)}$; similarly, the second inequality comes from the fact that $\mathbf{U}^{(k_{t+1})}$ and $\mathbf{W}^{(k_{t+1})}$ are the optimal solutions of the corresponding minimization problems with respect to $\mathbf{U}$ and $\mathbf{W}$, i.e., (PI) and (PJ), respectively; the third inequality is because of that the first-order Taylor expansion is a global overestimate of a concave function; the last equality can be easily verfied in (34). Interestingly, by replacing the first item by $\hat{Q}(\mathbf{d}^{(k_t)}, \mathbf{s}^{(k_t)}, \mathbf{V}^{(k_t)}; \mathbf{U}^{(k_t)}, \mathbf{W}^{(k_t)}, \hat{\mathbf{s}}^{(k_t)})$, (36) also indicates that IA-IWMMSE generates monotonically decreasing objective sequence of (PG).

Taking the limits of both sides of (36) with respect to t, we get[3]

$$\hat{Q}\left(\mathbf{d}, \mathbf{s}, \mathbf{V}; \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\right) \geq \hat{Q}\left(\mathbf{d}^*, \mathbf{s}^*, \mathbf{V}^*; \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\right) \geq Q^*, \tag{37}$$

which implies $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*\} \in \Omega(\mathbf{U}^*, \mathbf{W}^*; \hat{\mathbf{s}}^*)$, or equivalently, the KKT conditions of (PH) are satisfied at $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*\}$ with $\mathbf{U} = \mathbf{U}^*$, $\mathbf{W} = \mathbf{W}^*$ and $\hat{\mathbf{s}} = \hat{\mathbf{s}}^*$.

For the sake of conciseness, we define the superscript $\mathcal{A} \in \{\mathrm{U}, \mathrm{D}\}$ and the subscript $l \in \{c, d\}$ to simplify the notations. For instance, $h_{l,m}^{\mathcal{A}}(d_m, \mathbf{s}_m)$ with $\mathbf{s}_m = \{s_{d,m}^{\mathrm{U}}, s_{d,m}^{\mathrm{D}}, s_{c,m}\}$ may denote the following four functions,

$$h_{d,m}^{\mathrm{U}}(d_m, \mathbf{s}_m) \triangleq \frac{(s_{d,m}^{\mathrm{U}})^2}{B^{\mathrm{U}}(d_m+\epsilon)},\ h_{c,m}^{\mathrm{U}}(d_m, \mathbf{s}_m) \triangleq \frac{s_{c,m}^2}{B^{\mathrm{U}}(1-d_m+\epsilon)},$$
$$h_{d,m}^{\mathrm{D}}(d_m, \mathbf{s}_m) \triangleq \frac{(s_{d,m}^{\mathrm{D}})^2}{B^{\mathrm{D}}(d_m+\epsilon)},\ h_{c,m}^{\mathrm{D}}(d_m, \mathbf{s}_m) \triangleq \frac{s_{c,m}^2}{B^{\mathrm{D}}(1-d_m+\epsilon)}.$$

The KKT conditions of (PH) at $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*\}$ with $\{\mathbf{U}, \mathbf{W}, \hat{\mathbf{s}}\} = \{\mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\}$ are expressed as (38) at the top of next page, where $\xi_{l,m}^{\mathcal{A}}, \tau_m, \kappa_m, \nu_m^{\mathrm{U}}, \nu_{d,m}^{\mathrm{D}}$ and ν_c^{D} are the associated Lagrangian multipliers.

Due to the facts of $\mathbf{U}^* = \Gamma(\mathbf{V}^*)$ and $\mathbf{W}^* = \Upsilon(\mathbf{V}^*, \mathbf{U}^*)$, we must have

$$\begin{cases} w_{l,m}^{\mathcal{A},*} \nabla_{\mathbf{u}_{l,m}^{\mathcal{A}}} e_{l,m}^{\mathcal{A}}(\mathbf{V}^{\mathcal{A},*}, \mathbf{u}_{l,m}^{\mathcal{A},*}) = \mathbf{0}, \\ \qquad\qquad \forall\ \mathcal{A} \in \{\mathrm{U}, \mathrm{D}\},\ l \in \{c, d\},\ m \in \mathcal{M}, \\ \nabla_{w_{l,m}^{\mathcal{A}}} [w_{l,m}^{\mathcal{A},*} e_{l,m}^{\mathcal{A}}(\mathbf{V}^{\mathcal{A},*}, \mathbf{u}_{l,m}^{\mathcal{A},*}) - \log(w_{l,m}^{\mathcal{A},*})] = 0, \\ \qquad\qquad \forall\ \mathcal{A} \in \{\mathrm{U}, \mathrm{D}\},\ l \in \{c, d\},\ m \in \mathcal{M}. \end{cases} \tag{39}$$

To complete the proof, we perform a tactful reformulation. In particular, we multiply the equations in (39) by the Lagrangian

[3]We remark that taking the limit inside the objective value $\hat{Q}(\cdot)$ is legitimate, as the objective function is continuous.

$$
\begin{cases}
\sum\limits_{n\in\mathcal{M}}\sum\limits_{l\in\{c,d\}} \xi^{\mathrm{U}}_{l,n} w^{\mathrm{U},*}_{l,n} \nabla_{\mathbf{v}^{\mathrm{U}}_m} e^{\mathrm{U}}_{l,n}(\mathbf{V}^{\mathrm{U},*}, \mathbf{u}^{\mathrm{U},*}_{l,n}) + 2\nu^{\mathrm{U}}_m \mathbf{v}^{\mathrm{U},*}_m = \mathbf{0},\ \forall m \in \mathcal{M}, \\
\sum\limits_{n\in\mathcal{M}}\sum\limits_{l\in\{c,d\}} \xi^{\mathrm{D}}_{l,n} w^{\mathrm{D},*}_{l,n} \nabla_{\mathbf{v}^{\mathrm{D}}_{d,m}} e^{\mathrm{D}}_{l,n}(\mathbf{V}^{\mathrm{D},*}, \mathbf{u}^{\mathrm{D},*}_{l,n}) + 2\nu^{\mathrm{D}}_{d,m} \mathbf{v}^{\mathrm{D},*}_{d,m} = \mathbf{0},\ \forall m \in \mathcal{M}, \\
\sum\limits_{n\in\mathcal{M}}\sum\limits_{l\in\{c,d\}} \xi^{\mathrm{D}}_{l,n} w^{\mathrm{D},*}_{l,n} \nabla_{\mathbf{v}^{\mathrm{D}}_{c,m}} e^{\mathrm{D}}_{l,n}(\mathbf{V}^{\mathrm{D},*}, \mathbf{u}^{\mathrm{D},*}_{l,n}) + 2\nu^{\mathrm{D}}_{c} \mathbf{v}^{\mathrm{D},*}_{c,m} = \mathbf{0},\ \forall m \in \mathcal{M}, \\
\nabla_{\chi_m} \hat{Q}(\mathbf{d}^*, \mathbf{s}^*, \mathbf{V}^*; \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*) + \sum\limits_{\mathcal{A}\in\{\mathrm{U},\mathrm{D}\}}\sum\limits_{l\in\{c,d\}} \xi^{\mathcal{A}}_{l,m} \nabla_{\chi_m} h^{\mathcal{A}}_{l,m}(d^*_m, \mathbf{s}^*_m) + \tau_m \nabla_{\chi_m} d^*_m - \kappa_m \nabla_{\chi_m} d^*_m = 0, \\
\qquad\qquad \forall\ \chi_m \in \{d_m, s^{\mathrm{U}}_{d,m}, s^{\mathrm{D}}_{d,m}, s_{c,m}\},\ m \in \mathcal{M}, \\
0 \le \nu^{\mathrm{U}}_m \perp p^{\mathrm{U}}_m - \|\mathbf{v}^{\mathrm{U},*}_m\|_2^2 \ge 0, \quad 0 \le \nu^{\mathrm{D}}_{d,m} \perp p^{\mathrm{D}}_m - \|\mathbf{v}^{\mathrm{D},*}_{d,m}\|_2^2 \ge 0,\ \forall\ m \in \mathcal{M}, \\
0 \le \tau_m \perp d^*_m \ge 0,\ 0 \le \kappa_m \perp 1 - d^*_m \ge 0,\ 0 \le \nu^{\mathrm{D}}_c \perp P_B - \sum\limits_{m\in\mathcal{M}} \|\mathbf{v}^{\mathrm{D},*}_{c,m}\|_2^2 \ge 0, \\
0 \le \xi^{\mathcal{A}}_{l,m} \perp \log(w^{\mathcal{A},*}_{l,m}) - w^{\mathcal{A},*}_{l,m} e^{\mathcal{A}}_{l,m}(\mathbf{V}^{\mathcal{A},*}, \mathbf{u}^{\mathcal{A},*}_{l,m}) + 1 - h^{\mathcal{A}}_{l,m}(d^*_m, \mathbf{s}^*_m) \ge 0, \quad \forall\ \mathcal{A} \in \{\mathrm{U},\mathrm{D}\},\ l \in \{c,d\},\ m \in \mathcal{M}.
\end{cases} \tag{38}
$$

multipliers $\xi^{\mathcal{A}}_{l,m}$ and then obtain

$$
\begin{cases}
\xi^{\mathcal{A}}_{l,m} w^{\mathcal{A},*}_{l,m} \nabla_{\mathbf{u}^{\mathcal{A}}_{l,m}} e^{\mathcal{A}}_{l,m}(\mathbf{V}^{\mathcal{A},*}, \mathbf{u}^{\mathcal{A},*}_{l,m}) = \mathbf{0}, \\
\qquad\qquad \forall\ \mathcal{A} \in \{\mathrm{U},\mathrm{D}\},\ l \in \{c,d\},\ m \in \mathcal{M}, \\
\xi^{\mathcal{A}}_{l,m} \nabla_{w^{\mathcal{A}}_{l,m}} [w^{\mathcal{A},*}_{l,m} e^{\mathcal{A}}_{l,m}(\mathbf{V}^{\mathcal{A},*}, \mathbf{u}^{\mathcal{A},*}_{l,m}) - \log(w^{\mathcal{A},*}_{l,m})] = 0, \\
\qquad\qquad \forall\ \mathcal{A} \in \{\mathrm{U},\mathrm{D}\},\ l \in \{c,d\},\ m \in \mathcal{M}.
\end{cases} \tag{40}
$$

We can exactly get the KKT conditions of (PG) at $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*\}$ by directly combining (38) and (40). Then, we claim that $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*\}$ is a stationary solution of (PG). □

Lemma 2: If $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*\}$ is a limit point of the iterates generated by IA-IWMMSE, then it is a stationary solution of (PF).

proof: According to the results of Lemma 1, we know that $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*\}$ is a stationary solution of (PG), and $\hat{\mathbf{s}}^* = \mathbf{s}^*$. Based on the definitions of (34), we obtain

$$
\begin{aligned}
&\nabla_{\chi_m} Q(\mathbf{d}^*, \mathbf{s}^*, \mathbf{V}^*; \mathbf{U}^*, \mathbf{W}^*) \\
&= \nabla_{\chi_m} \hat{Q}(\mathbf{d}^*, \mathbf{s}^*, \mathbf{V}^*; \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^* = \mathbf{s}^*), \\
&\qquad \forall\ \chi_m \in \{d_m, s^{\mathrm{U}}_{d,m}, s^{\mathrm{D}}_{d,m}, s_{c,m}\},\ m \in \mathcal{M}.
\end{aligned} \tag{41}
$$

Replacing $\nabla_{\chi_m} \hat{Q}(\mathbf{d}^*, \mathbf{s}^*, \mathbf{V}^*; \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*)$ by $\nabla_{\chi_m} Q(\mathbf{d}^*, \mathbf{s}^*, \mathbf{V}^*; \mathbf{U}^*, \mathbf{W}^*)$, (38) and (40) still hold. Then, we get the KKT conditions of (PF) at $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\}$, indicating that it is a stationary solution of (PF). □

Utilizing the properties of WMMSE reformulation (Theorem {2, 3} in [28] and Proposition 1 in [29]), we conclude that $\{\mathbf{V}^*, \mathbf{d}^*, \mathbf{s}^*, \mathbf{U}^*, \mathbf{W}^*, \hat{\mathbf{s}}^*\}$ is also a stationary point of (PE).

REFERENCES

[1] J. Lin, Q. Shi, and Q. Li, "Joint device-to-device transmission activation and transceiver design for sum-rate maximization in MIMO interfering channels," in *Proc. IEEE ICASSP*, pp. 3921–3925, Shanghai, Mar. 2016.

[2] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, Nov. 2014.

[3] K. Doppler, M. Rinne, C. Wijting, C.B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, Dec. 2009.

[4] D. Feng, L. Lu, Y.-W. Yi, G.Y. Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 49–55, Apr. 2014.

[5] H. Min, W. Seo, J. Lee, S. Park, and D. Hong, "Reliability improvement using receive mode selection in the device-to-device uplink period underlaying cellular networks," *IEEE Trans. Wireless Communications*, vol. 10, no. 2, pp. 413–418, Feb. 2011.

[6] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 90–96, Jun. 2014.

[7] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design aspects of network assisted device-to-device communications," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[8] C.-H, Yu, K. Doppler, C.B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[9] Z. Liu, T. Peng, S. Xiang, and W. Wang, "Mode selection for device-to-device (D2D) communication under LTE-advanced networks," in *Proc. IEEE ICC*, Ottawa, Canada, June 2012, pp. 5563–5567.

[10] K. Doppler, C.-H. Yu, C. B. Ribeiro, and P. Janis, "Mode selection for device-to-device communication underlaying an LTE-advanced network," in *Proc. IEEE WCNC*, Sydney, Australia, April 2010, pp. 1–6.

[11] L. Song, D. Niyato, Z. Han, and E. Hossain, "Game-theoretic resource allocation methods for device-to-device (D2D) communication," *IEEE Wireless Communications Magazine*, vol. 21, no. 3, pp. 136–144, Jun. 2014.

[12] R. Zhang, X. Cheng, L. Yang, and B. Jiao, "Interference-aware graph based resource sharing for device-to-device communications underlaying cellular networks,", in *Proc. IEEE WCNC*, Shanghai, China, April 2013, pp. 140–145.

[13] L. Su, Y. Ji, P. Wang, and F. Liu, "Resource allocation using particle swarm optimization for D2D communication underlay of cellular networks," in *Proc. IEEE WCNC*, Shanghai, April 2013, pp. 129–133.

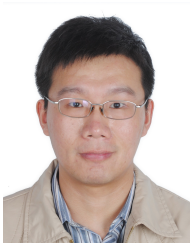
[14] L. Wei, R.Q. Hu, T. He, and Y. Qian, "Device-to-device (D2D) communications underlaying MU-MIMO cellular networks," in *Proc. IEEE GLOBECOM*, Atlanta, Dec. 2013, pp. 4902–4907.

[15] P. Janis, V. Koivunen, C.B. Ribeiro, K. Doppler, and K. Hugl, "Interference-avoiding MIMO schemes for device-to-device radio underlaying cellular networks," in *Proc. IEEE PIMRC*, Tokyo, Japan, Sept. 2009, pp. 2385–2389.

[16] J.C.F. Li, M. Lei, and F. Gao, "Device-to-device (D2D) communication in MU-MIMO cellular networks," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 3583–3587, Anaheim, CA, Dec. 2012.

[17] D. Zhu, W. Xu, H. Zhang, C. Zhao, J.C.F. Li, and M. Lei, "Rate maximized tranceiver optimization for multi-antenna device-to-device communications," in *Proc. IEEE WCNC*, Shanghai, China, April 2013, pp. 4152–4157.

[18] A. Tolli, J. Kaleva, and P. Komulainen, "Mode selection and transceiver design for rate maximization in underlay D2D MIMO systems," in *Proc. IEEE ICC*, London, UK, June 2015, pp. 4822–4827.
[19] N. Reider and G. Fodor, "A distributed power control and mode selection algorithm for D2D communications," *EURASIP Journal on Wireless Communications and Networking*, 2012: 266, pp. 1–25, Aug. 2012.
[20] E. Naghipour and M. Rasti, "A distributed joint power control and mode selection scheme for D2D communications underlaying LET-A networks," in *Proc. IEEE WCNC*, Doha, Qatar, Apr. 2016, pp. 1–6.
[21] H. Tang and Z. Ding, "Mixed mode transmission and resource allocation for D2D Communication," *IEEE Trans. Wireless Communications*, vol. 15, no. 1, pp. 162–175, Jan. 2016.
[22] Q. Ye, M. Al-Shalash, C. Caramanis, and J.G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks" *IEEE Trans. Communications*, vol. 63, no. 2, pp. 441–454, Feb. 2015.
[23] D. Wu, Y. Cai, R.Q. Hu, and Y. Qian, "Dynamic distributed resource sharing for mobile MIMO D2D communications," *IEEE Trans. Wireless Communications*, vol. 14, no. 10, pp. 5417–5429, Oct. 2015.
[24] X. Lin, J.G. Andrews, A. Ghosh, and R. Ratasuk, "An overview on 3GPP device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, Apr. 2014.
[25] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: A survey," *IEEE Communication Surveys & Tutorials*, vol. 17, no. 4, pp.1923–1940, Fourth Quarter 2015.
[26] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in OFDMA cellular networks: A survey and challenges," *IEEE Communication Surveys & Tutorials*, vol. 17, no. 4, pp. 1885–1992, Fourth Quarter 2015.
[27] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: complexity and duality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 20, no. 1, pp. 57–73, Feb. 2008.
[28] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.
[29] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear transceiver design for a MIMO interfering broadcast channel achieving max-min fairness," *Signal Processing*, vol. 93, no. 12, pp. 3327–3340, 2013.
[30] M. Fukushima, "Application of alternating direction method of multipliers to separable convex programming problems," *Computational Optimization and Applications*, vol. 1, no. 1, pp. 93–111, Oct. 1992.
[31] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, UK: Cambridge University Press, 2004.
[32] R. Horst and N.V. Thoai, "DC programming: overview," *J. of Optim. Theory and Appl.*, vol. 103, no. 1, pp. 1–43, 1999.
[33] M. Grant, S. Boyd, "The CVX users' guide," *CVX Research Inc.*, Oct. 24, 2014 [on-line]. http://cvxr.com/cvx/.

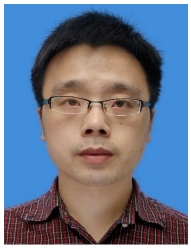


Jingran Lin received the B.S. degree in Computer Communication from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2001, and the M.S. and Ph.D. degrees in Signal and Information Processing from UESTC in 2005 and 2007, respectively.

After his graduation in July 2007, he joined the School of Information and Communication Engineering, UESTC, where he is currently an Associate Professor. From January 2012 to January 2013, he was a Visiting Scholar with the University of Minnesota (Twin Cities), Minneapolis, MN, USA. His research interests include the design and analysis of efficient optimization algorithms for the signal processing problems arising from modern communication systems.

Qingjiang Shi received his Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011. From September 2009 to September 2010, he visited Prof. Z.-Q. (Tom) Luo's research group at the University of Minnesota, Twin Cities. In 2011, he worked as a Research Scientist at Bell Labs China. From 2012, he was with the School of Information and Science Technology at Zhejiang Sci-Tech University. From Feb. 2016 to Mar. 2017, he worked as a research fellow at Iowa State University, USA. From 2018, he is currently a professor with the School of Software Engineering at Tongji University. His interests lie in algorithm design and analysis with applications in machine learning, signal processing and wireless networks. So far he has published more than 40 IEEE journals (five papers were ESI highly cited papers and one was nominated as the best paper award of IEEE Signal Processing Society in 2016) and filed more than 20 national patents.

Dr. Shi is an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He was awarded Golden Medal at the 46th International Exhibition of Inventions of Geneva in 2018, and also was the recipient of the First Prize of Science and Technology Award from China Institute of Communications in 2017, the National Excellent Doctoral Dissertation Nomination Award in 2013, the Shanghai Excellent Doctorial Dissertation Award in 2012, and the Best Paper Award from the IEEE PIMRC'09 conference.



Qiang Li received the B.Eng. and M.Phil. degrees in Communication and Information Engineering from University of Electronic Science and Technology of China (UESTC), Chengdu, China, and the Ph.D. degree in Electronic Engineering from the Chinese University of Hong Kong (CUHK), Hong Kong, in 2005, 2008, and 2012, respectively. From August 2011 to January 2012, he was a Visiting Scholar with the University of Minnesota, Minneapolis, MN, USA. From February 2012 to October 2013, he was a Research Associate with the Department of Electronic Engineering and the Department of Systems Engineering and Engineering Management, CUHK. Since November 2013, he has been with the School of Information and Communication Engineering, UESTC, where he is currently an Associate Professor. His research interests include convex optimization and its application in signal processing with an emphasis on the physical-layer security and full-duplex communications.

He received the First Prize Paper Award in the IEEE Signal Processing Society Postgraduate Forum Hong Kong Chapter in 2010, a Best Paper Award of IEEE PIMRC 2016, and the Best Paper Award of the IEEE Signal Processing Letters 2016.

Dongmei Zhao received the B.S. degree in Communication Engineering from Chongqing University, Chongqing, China, in 2015, and received the M.S. degree in Electronic and Communication Engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018.

She is currently working as a Software Engineer in ZTE Corporation. Her research interests include cooperative communication and communication protocol.