Optimization of MIMO Device-to-Device Networks via Matrix Fractional Programming: A Minorization–Maximization Approach

Kaiming Shen, *Student Member, IEEE*, Wei Yu, *Fellow, IEEE*, Licheng Zhao, and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*— Interference management is a fundamental issue in device-to-device (D2D) communications whenever the transmitter-and-receiver pairs are located in close proximity and frequencies are fully reused, so active links may severely interfere with each other. This paper devises an optimization strategy named FPLinQ to coordinate the link scheduling decisions among the interfering links, along with power control and beamforming. The key enabler is a novel optimization method called matrix fractional programming (FP) that generalizes previous scalar and vector forms of FP in allowing multiple data streams per link. From a theoretical perspective, this paper provides a deeper understanding of FP by showing a connection to the minorization-maximization (MM) algorithm. From an application perspective, this paper shows that as compared to the existing methods for coordinating scheduling in the D2D network, such as FlashLinQ, ITLinQ, and ITLinQ+, the proposed FPLinQ approach is more general in allowing multiple antennas at both the transmitters and the receivers, and further in allowing arbitrary and multiple possible associations between the devices via matching. Numerical results show that FPLinQ significantly outperforms the previous state-of-the-art in a typical D2D communication environment.**

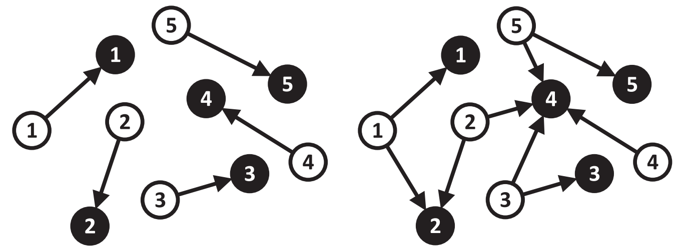
***Index Terms*— Device-to-device (D2D) networks, link scheduling, power control, beamforming, matrix fractional programming (FP), minorization-maximization (MM) algorithm.**

I. Introduction

SPECTRUM sharing in an interference-limited wireless communication environment is one of the most fundamental problems in network engineering, for which no efficient global optimum algorithm is yet available. This problem is challenging especially when a large number of mutually interfering links are present. One essential difficulty lies in deciding which links should be active at any given time, i.e., how to schedule. But the optimal scheduling is also intimately related to power control and beamforming, if the communication links are equipped with multiple antennas, as power and beam pattern have significant effect on the interference. This coordinated scheduling, beamforming, and power control problem is important in the emerging device-to-device (D2D) communication paradigm in which arbitrary peer-to-peer transmissions can take place, but also relevant in traditional cellular networks in which coordination among the cells can significant improve the network performance.

This paper devises a novel optimization technique based on fractional programming (FP) for solving this classic problem. The problem formulation is that of maximizing a weighted sum rate of links across a D2D network, in which the weights account for fairness and the links are selectively activated in order to alleviate interference. In addition, this paper considers a model that allows each transmitter to have the flexibility of associating with one of the multiple receivers, and each receiver to have the flexibility of associating with one of the multiple transmitters, as illustrated in Fig. 1. This overall scheduling, power control, and beamforming problem is a difficult combinatorial and nonconvex optimization, because the scheduling decision of each link depends strongly on the activation states and the transmission parameters (e.g., power and beamforming pattern) of the nearby links.

Motivated by the crucial role of the signal-to-interference-plus-noise ratio (SINR) in the communication system design, this paper proposes a *fractional programming based link scheduling* (FPLinQ) strategy to address the coordinated scheduling, beamforming, and power control problem in an interference network. While the use of FP for scheduling,



(a) Fixed Single Association (b) Flexible Association

Fig. 1. D2D network with white circles denoting the transmitters and black circles denoting the receivers. In the fixed single association model (a), the transmitters have a fixed one-to-one mapping to the receivers. This paper considers a more general setting (b) in which each transmitter has the flexibility of associating with one of the multiple receivers, and each receiver has the flexibility of associating with one of the multiple transmitters.

Manuscript received March 29, 2019; revised August 7, 2019; accepted September 20, 2019; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor C. W. Tan. Date of publication October 8, 2019; date of current version October 15, 2019. This work was supported in part by the Natural Science and Engineering Research Council (NSERC), in part by Huawei Technologies Canada, in part by the Hong Kong RGC 16208917 Research Grant, and in part by the Hong Kong Telecom Institute of Information Technology Visiting Fellowship. This article was presented in part at the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, June 2017 [1]. *(Corresponding author: Kaiming Shen.)*

K. Shen and W. Yu are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: kshen@ece.utoronto.ca; weiyu@ece.utoronto.ca).

L. Zhao and D. P. Palomar are with the Department of Electrical and Computer Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong, China (e-mail: lzhaoai@connect.ust.hk; palomar@ust.hk).

Digital Object Identifier 10.1109/TNET.2019.2943561

beamforming, and power control originated from our previous work [2], [3], the method proposed in [2], [3] is restrictive in the sense that: (i) only scalar and vector cases are treated so that each communication link can only have one single data stream; (ii) the application regime is restricted to the cellular setting in which each user is associated with one single fixed base-station.

The present paper generalizes [2], [3] in several nontrivial directions. The key theoretical development here is a novel matrix FP technique for dealing with ratios involving matrices, in contrast to earlier FP techniques that deal with only the scalar ratio, thus allowing the full capacity of the multiple-input multiple-output (MIMO) channel to be realized with multiple data streams. Furthermore, this paper tackles a more general scheduling problem in which each transmitter/receiver has the flexibility in associating with each other. Moreover, this paper makes a theoretical contribution by interpreting the proposed FP approach as a minorization-maximization (MM) algorithm, thus allowing convergence to be readily established. An interesting finding of this paper is that the FP transforms can be interpreted as novel surrogate functions in the MM context.

A. Related Work

Interference-aware scheduling, power control and beaforming for wireless networks have attracted considerable research interests over the years, e.g., [4]–[12]. In the multiple-antenna cellular network context, the well-known *weighted minimum mean square error* (WMMSE) algorithm [13], [14] is able to attain a stationary point of the joint power control and beamforming problem; furthermore, under some special condition for the single-cell case, the global optimum solution can sometimes be found [15]. But the cellular model is different from the model considered in this paper, because spatial multiplexing can typically be implemented at a cellular base-station, while the D2D model of this paper only allows one-to-one mapping between each transmitter and each receiver.

In the D2D context, there are a vast array of works in the existing literature exploring a variety of different directions, including geometric programming [16], game theory [17], stochastic geometry [18], [19], evolution theory [20], and dynamic programming [21]. While some of existing works [4], [22], [23] adopt a quality-of-service (QoS) model for the scheduling problem, many other works (including this paper) consider maximizing the weighted sum rate across the D2D network, where the fairness is taken into account by appropriate setting of the weights.

This paper is most closely related to a series of works that propose algorithms called FlashLinQ [24], ITLinQ [25], and ITLinQ+ [26], which address the D2D scheduling problem using greedy search while utilizing information theoretic intuition based on generalized degrees-of-freedom (GDoF); we review these algorithms in detail in Section II-C.

An important benchmark method for the problem considered in this paper is the block coordinate descent (BCD) approach, which is proposed for the cellular network in [27], but can also be adopted for the D2D model. However, BCD is prone to being trapped in the local optimum solution, as we discussed in Section II-B. Using the greedy algorithm and the BCD method as the benchmarks, the aim of this paper is to show that an optimization motivated approach based on FP can significantly outperform these state-of-the-art methods.

B. Main Contributions

The main contributions of this paper are summarized below:

- *Multiple-Antenna Flexible-Association D2D Network Model:* This work considers a D2D network setup with multiple antennas at both the transmitters and the receivers, thus each link can carry multiple data streams. Further, the model considered here allows the flexibility among multiple possible associations between the transmitters and the receivers. This is a more general model than the ones considered in the previous works [2], [3], [24]–[26].
- *Matrix FP Transforms:* This paper introduces the matrix FP which treats $\sqrt{\mathbf{A}}^\dagger(\mathbf{x})\mathbf{B}^{-1}(\mathbf{x})\sqrt{\mathbf{A}}(\mathbf{x})$ as a ratio between the positive (semi)definite matrix-valued functions $\mathbf{A}(\mathbf{x})$ and $\mathbf{B}(\mathbf{x})$, whereas the previous FP theory focuses on the scalar ratio $A(\mathbf{x})/B(\mathbf{x})$ between the real-valued functions $A(\mathbf{x}) \geq 0$ and $B(\mathbf{x}) > 0$ [28] or the vector case $\mathbf{a}^\dagger(\mathbf{x})\mathbf{B}^{-1}(\mathbf{x})\mathbf{a}(\mathbf{x})$ with the vector function $\mathbf{a}(\mathbf{x})$ [2], [3]. We extend the FP transforms of [2], [3] to the matrix case.
- *Interpretation of Matrix FP as MM:* This paper shows that, from an MM algorithm perspective, the proposed matrix FP transforms can be thought of as constructing surrogate functions for the original problem. In this sense, this work puts forward a novel way of minorizing the logarithmic objective function and the fractional function, in contrast to the traditional application of MM, which relies on the second-order Taylor expansion.
- *FPLinQ Algorithm:* This paper proposes an efficient FP based numerical algorithm for the iterative optimization of scheduling, beamforming, and power control for a D2D network. It achieves a higher network utility than the previous state-of-the-art. We observe that the direct optimization of these variables, using for example the WMMSE algorithm [13], [14], may incur a premature link turning-off problem. In addition, as compared to FlashLinQ, ITLinQ, and ITLinQ+ [24]–[26], we point out that the information theoretic justification for ITLinQ actually does not guarantee the optimality of scheduling, and further the proposed FPLinQ strategy also has an advantage in that it does not require any tuning of design parameters. The proposed FPLinQ strategy is based on the decoupling of a matrix ratio. There are in fact multiple different possible decoupling strategies, but the one adopted for FPLinQ is best suited for algorithmic implementation.

C. Paper Organization and Notation

The rest of this paper is organized as follows. Section II states the problem formulation for the wireless joint link scheduling, beamforming, and power control problem. Section III introduces the matrix FP. We provide two useful transforms and also connect them to the MM algorithm. Section V derives the proposed FPLinQ algorithm. Section VI provides numerical results to validate the performance of the proposed algorithm. Finally, Section VII concludes the paper.

We use lower case, e.g., s, to denote scalars, bold lower case, e.g., $\mathbf{x}$, to denote vectors, and bold upper case, e.g., $\mathbf{V}$, to denote matrices. We use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{C}$ to denote the set of complex numbers, $\mathbb{H}_{++}$ (or $\mathbb{H}_{+}$) to denote the set of Hermitian positive definite (or semidefinite) matrices, $\boldsymbol{I}$ to denote the identity matrix, and

$\mathbf{0}$ to denote the zero matrix. We use $(\cdot)^\dagger$ to denote matrix conjugate transpose, $\Re\{\cdot\}$ to denote the real part of a complex number, and $\mathrm{tr}(\cdot)$ to denote matrix trace. We use underline to denote a collection of variables, e.g., $\underline{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\underline{\mathbf{Y}} = \{\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}\}$.

II. Joint Scheduling, Beamforming, and Power Control in D2D Network

A. Problem Formulation

Consider a wireless D2D network with a set of transmitters $\mathcal{I}$ and a set of receivers $\mathcal{J}$. We assume that each transmitter may have data to transmit to one or more receivers, and likewise each receiver may wish to receive data from one or more transmitters. Thus, the communication scenario considered in this paper is a generalization of traditional D2D network with fixed single association between each pair of transmitter and receiver to a scenario with multiple possible associations between the transmitters and the receivers as shown in Fig. 1. We assume that in each scheduling time slot, each transmitter (or receiver) can only communicate with at most one of its associated receivers (or transmitters),[1] respectively, so that the mapping between the transmitters and the receivers is one-to-one. The task of scheduling is to identify which set of links over the network to activate in each slot. Further, we assume that the transmitters and the receivers are each equipped with N antennas and permit multiple data streams to be carried via MIMO transmission. The task of beamforming and power control is to design the transmit beamformers for each of these multiple data streams in each active link in the scheduling slot.

Mathematically, let $\mathcal{K}_j \subseteq \mathcal{I}$ be the set of transmitters associated with each particular receiver $j \in \mathcal{J}$. Likewise, let $\mathcal{L}_i \subseteq \mathcal{J}$ be the set of receivers associated with each transmitter $i \in \mathcal{I}$. Let $\mathbf{H}_{ji} \in \mathbb{C}^{N\times N}$ be the channel from transmitter i to receiver j in the scheduling slot. The joint scheduling, beamforming, and power control problem can be written down as that of optimizing two sets of variables: s_j, the index of the transmitter scheduled at receiver j, and $\mathbf{V}_i \in \mathbb{C}^{N\times N}$, the collection of beamforming vectors at transmitter i in each scheduling slot so as to maximize some network wide objective function. Throughout this paper, we assume that the channel state information (CSI) is completely known and network optimization is performed in a centralized manner. We note that this network optimization problem is NP-hard, even under such idealized assumptions [29], [30].

This paper uses the weighted sum rate as the optimization objective in each scheduling slot, where the weights are adjusted from slot to slot in an outer loop in order to maximize a network utility of long-term average rates. We assume that interference is treated as noise, so that the achievable data rate in each scheduling slot can be computed from the receiver's perspective, i.e., for each receiver j, as [31]

$$R_j = \log\left|\boldsymbol{I} + \mathbf{V}_{s_j}^\dagger \mathbf{H}_{js_j}^\dagger \mathbf{F}_j^{-1} \mathbf{H}_{js_j} \mathbf{V}_{s_j}\right| \tag{1}$$

with the interference-plus-noise term

$$\mathbf{F}_j = \sigma^2 \boldsymbol{I} + \sum_{j'\in\mathcal{J}\setminus\{j\}} \mathbf{H}_{js_{j'}} \mathbf{V}_{s_{j'}} \mathbf{V}_{s_{j'}}^\dagger \mathbf{H}_{js_{j'}}^\dagger, \tag{2}$$

[1]Note that the D2D model considered in this paper is more general than the traditional wireless cellular network model of [3] in effectively allowing multiple and arbitrary associations between the base-stations and the mobile terminals, but on the other hand, is also more restrictive in that it does not allow spatial multiplex at either the receiver or the transmitter.

where σ^2 is the power of thermal noise. Given a set of nonnegative weights $w_{ji} \geq 0$, the optimization problem is therefore

$$\underset{\underline{\mathbf{V}},\, \underline{s}}{\text{maximize}} \quad \sum_{j\in\mathcal{J}} w_{js_j} R_j \tag{3a}$$

$$\text{subject to} \quad \mathrm{tr}\left(\mathbf{V}_i^\dagger \mathbf{V}_i\right) \leq P_{\max},\ \forall i \in \mathcal{I}, \tag{3b}$$

$$s_j \in \mathcal{K}_j \cup \{\varnothing\},\ \forall j \in \mathcal{J}, \tag{3c}$$

$$s_j \neq s_{j'} \text{ or } s_j = \varnothing,\ \forall j \neq j', \tag{3d}$$

where we have assumed a per-scheduling-slot and per-node power constraint $P_{\max}$ and $\varnothing$ denotes the decision of not scheduling any transmitter at receiver j. We remark that $\mathbf{H}_{js_j}$, $\mathbf{V}_{s_j}$, and w_{js_j} are set to zero if $s_j = \varnothing$. Constraint (3d) states that the same transmitter cannot be scheduled for more than one receiver at a time, as required by the assumption that the association between the transmitters and the receivers in the D2D network must be one-to-one. Problem (3) involves a discrete optimization over $\underline{s}$ and a nonconvex continuous optimization over $\underline{\mathbf{V}}$, which make it a challenging optimization problem. Below, we first review several conventional approaches including the BCD algorithm and the greedy algorithms.

B. Block Coordinate Descent

The joint scheduling, beamforming, and power allocation problem as formulated in (3) is a mixed discrete-continuous programming problem. To reach a reasonable solution, we could optimize the discrete scheduling variable $\underline{s}$ and the continuous beamforming variable $\underline{\mathbf{V}}$ separately and alternatively in a form of the BCD algorithm [27]. When $\underline{s}$ is held fixed, optimizing $\underline{\mathbf{V}}$ alone in (3) is the conventional beamforming problem for which existing methods (e.g., the WMMSE algorithm [13], [14]) can be applied. When $\underline{\mathbf{V}}$ is held fixed, optimizing $\underline{s}$ alone in (3) can be recognized as a weighted bipartite matching problem which can be solved by standard methods.

However, we point out that the BCD approach is prone to a potentially serious *premature turning-off* problem. Suppose that none of the links related to a particular transmitter i is scheduled at the early stage of the iterative optimization, then the beamforming variable $\mathbf{V}_i$ would be set to zero. As a result, when $\underline{s}$ is optimized via matching for the fixed $\underline{\mathbf{V}}$ in the next iteration, the matching weights related to i would all be equal to zero, so the corresponding links can never be turned back on. Therefore, premature scheduling decisions can adversely affect the overall performance of the algorithm.

C. FlashLinQ, ITLinQ, and ITLinQ+

We further examine the state-of-the-art methods for D2D link scheduling in the literature: FlashLinQ [24], ITLinQ [25], and ITLinQ+ [26]. These works assume that each terminal has a single antenna, and further that each transmitter (or receiver) is only associated with one receiver (or transmitter) respectively, namely the fixed single association case shown in Fig. 1(a).

Because deciding the ON-OFF state for all the links at the same time is difficult, all three algorithms propose to schedule the links in a greedy fashion sequentially, as stated in Algorithm 1. The main difference between FlashLinQ [24], ITLinQ [25], and ITLinQ+ [26] lies in the criterion for deciding whether the new link conflicts with already scheduled ones in Step 3 of Algorithm 1.

Algorithm 1 Sequential Link Selection

Initialize the set of activated links $\mathcal{A}$ to $\emptyset$;
for *each link* (i, j) **do**
  if (i, j) *does not "conflict" with any link in* $\mathcal{A}$ **then**
    Schedule link (i, j) and add it to $\mathcal{A}$;
  end
end

1) FlashLinQ [24]: The FlashLinQ scheme [24] applies a threshold θ to SINR, assuming that adding link i to $\mathcal{A}$ does not cause conflict if and only if all the activated links have their SINRs higher than θ. The performance of FlashLinQ is highly sensitive to the value of θ, but choosing θ properly can be difficult in practice. Further, using the same θ for all the links is usually suboptimal when the weight varies from link to link.

2) ITLinQ [25] and ITLinQ+ [26]: From an information theory perspective, a seminal study [32] on the multi-user Gaussian interference channel provides a sufficient (albeit not necessary) condition for the optimality of treating interference as noise (TIN) for maximizing the GDoF as follows:

$$\log |h_{ji}| \geq \max_{j' \neq j} \{\log |h_{j'i}|\} + \max_{i' \neq i} \{\log |h_{ji'}|\}, \tag{4}$$

where $h_{ji} \in \mathbb{C}$ is the channel of the single-antenna case. We refer to this result as the TIN condition.

The central idea of ITLinQ and ITLinQ+ is to schedule a subset of links that meet this TIN condition. Because the TIN condition in (4) is often too stringent to activate sufficient number of links, ITLinQ and ITLinQ+ both introduce relaxation based on design parameters. Like FlashLinQ, the performance of ITLinQ and ITLinQ+ is heavily dependent on the choice of design parameters, but they are difficult to choose optimally in practice. For example, [26] adopts two different sets of parameters for ITLinQ+ for two different network models. It is often not clear how to adapt ITLinQ and ITLinQ+ to the particular network environment of interest.

It is important to point out that the theoretical basis of ITLinQ and ITLinQ+, i.e., the TIN condition, only helps decide whether for some particular schedule, treating interference as noise is the optimal coding strategy from a GDoF perspective. It does not, however, guarantee that if some schedule satisfies the TIN condition, then it must be the GDoF optimal schedule. Thus, for a particular network, a schedule that does not satisfy the TIN condition can outperform one that does.

This subtle point is illustrated in the three-link D2D network example shown in Fig. 2. Let the desired signal strength be P and interfering signal strength be $P^{0.6}$. At most one link can be activated according to (4), so under the TIN condition, the total GDoF ≤ 1. But, a higher sum GDoF of 1.2 can be achieved by simply activating all the links. Therefore, the TIN condition does not guarantee even the GDoF optimality of a given schedule. Considering further the significant gap between GDoF and the actual achievable rate, ITLinQ and ITLinQ+ can often produce quite suboptimal solutions.

In contrast to FlashLinQ, ITLinQ, and ITLinQ+, this paper takes an optimization theoretic approach of recognizing the optimization objective function as a matrix FP (since it involves a matrix ratio inside a logarithm), then proposes an iterative method via a series of matrix FP transforms.

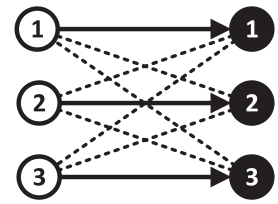


Fig. 2. Power strength is P for each solid signal and is $P^{0.6}$ for each dashed signal. Thus, the sum GDoF equals to 1 if only one link is on, and equals to 1.2 if all links are on.

The proposed iterative method involves the update of all the scheduling variables in one step at the same time and the optimization of beamforming and power variables in subsequent steps. One advantage of this optimization motivated approach as compared to FlashLinQ, ITLinkQ and ITLinQ+ is that it does not require the tuning of any design parameters.

III. Matrix Fractional Programming

To develop the matrix FP transform, we first present the scalar case as proposed in [2], [3], then provide the matrix generalization.

A. Scalar FP Transforms

For scalar FP, the following transform is proposed in [2] in order to facilitate optimization by decoupling the numerators and denominators of the scalar fractional terms in the objective function of an FP.

Proposition 1 (Quadratic Transform [2]): Given a nonempty constraint set $\mathcal{X}$ as well as a sequence of nonnegative functions $A_m(\mathbf{x}) \geq 0$, strictly positive functions $B_m(\mathbf{x}) > 0$, and monotonically nondecreasing functions $f_m(z) : \mathbb{R} \mapsto \mathbb{R}$, for $m = 1, 2, \ldots, M$, the *sum-of-functions-of-ratio* problem

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{m=1}^{M} f_m\left(\frac{A_m(\mathbf{x})}{B_m(\mathbf{x})}\right) \tag{5a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{5b}$$

is equivalent to

$$\underset{\mathbf{x},\, \underline{y}}{\text{maximize}} \quad \sum_{m=1}^{M} f_m\left(2y_m\sqrt{A_m(\mathbf{x})} - y_m^2 B_m(\mathbf{x})\right) \tag{6a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{6b}$$

where $y_m \in \mathbb{R}$ is an auxiliary variable introduced for each ratio term $A_m(\mathbf{x})/B_m(\mathbf{x})$.

The above transform is called the quadratic transform, because it involves a quadratic function of the auxiliary variables. The quadratic transform decouples the numerator and the denominator of the fraction, thereby enabling the iterative optimization between $\mathbf{x}$ and y_m after the transformation. This strategy works well for a variety of optimization problems, including scalar version of the scheduling, beamforming, and power control problem.

Although not immediately recognized at the time [2] was published, the above quadratic transform (at least for the case where the functions are trivial, i.e., $f_m(z) = z$ for each m) is very similar to the earlier work of Benson [33], [34], as restated below.

Proposition 2 (Benson's Transform [33], [34]): Given a nonempty constraint set $\mathcal{X}$ as well as a sequence of nonnegative functions $A_m(\mathbf{x}) \geq 0$ and strictly positive functions

$B_m(\mathbf{x}) > 0$, $m = 1, 2, \ldots, M$, the *sum-of-ratios* problem

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{m=1}^{M} \frac{A_m(\mathbf{x})}{B_m(\mathbf{x})} \tag{7a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{7b}$$

is equivalent to

$$\underset{\mathbf{x},\underline{u},\underline{v}}{\text{maximize}} \quad \sum_{m=1}^{M} \left(2u_m\sqrt{A_m(\mathbf{x})} - v_m B_m(\mathbf{x})\right) \tag{8a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{8b}$$

$$u_m^2 - v_m \leq 0,\ \forall m = 1, 2, \ldots, M, \tag{8c}$$

where $u_m \in \mathbb{R}$ and $v_m \in \mathbb{R}$ are introduced as the auxiliary variables for each ratio term $A_m(\mathbf{x})/B_m(\mathbf{x})$.

The above transform is proposed by Benson [33], [34] in order to facilitate a branch-and-bound search for the global optimum of the sum-of-ratios problem. It can be shown that at the optimum, we must have $u_m^2 = v_m$, thus if we had made them equal *a priori*, this reduces to the quadratic transform of Proposition 1.

In many practical applications, we wish to optimize functions of ratios. If the functions are monotonic, then one can apply the quadratic transform directly as stated in Theorem 1. However in case of logarithmic function, as often encountered in communication system design, a better alternative is to apply the following transform, proposed in [3], to "move" the ratio terms to the outside of logarithm. This has an advantage when discrete (such as scheduling) variables are involved, as it allows matching algorithms to be used for discrete optimization.

Proposition 3 (Lagrangian Dual Transform [3]): Given a nonempty constraint set $\mathcal{X}$ as well as a sequence of nonnegative functions $A_m(\mathbf{x}) \geq 0$, strictly positive functions $B_m(\mathbf{x}) > 0$, and nonnegative weights $w_m \geq 0$, for $m = 1, 2, \ldots, M$, the *sum-of-logarithmic-ratios* problem

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{m=1}^{M} w_m \log\left(1 + \frac{A_m(\mathbf{x})}{B_m(\mathbf{x})}\right) \tag{9a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{9b}$$

is equivalent to

$$\underset{\mathbf{x},\,\underline{\gamma}}{\text{maximize}} \quad f_r(\mathbf{x}, \underline{\gamma}) \tag{10a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{10b}$$

where the new objective function is

$$f_r(\mathbf{x}, \underline{\gamma}) = \sum_{m=1}^{M} w_m \log(1 + \gamma_m) - \sum_{m=1}^{M} \gamma_m w_m + \sum_{m=1}^{M} \frac{(1 + \gamma_m) w_m A_m(\mathbf{x})}{A_m(\mathbf{x}) + B_m(\mathbf{x})} \tag{11}$$

with an auxiliary variable $\gamma_m \in \mathbb{R}$ introduced for each ratio term $A_m(\mathbf{x})/B_m(\mathbf{x})$.

The main result of [3] is that the quadratic transform and the Lagrangian dual transform can be applied together to decouple the ratio terms in the rate expression for wireless cellular networks, thus making the network optimization problem much easier to tackle especially in the presence of discrete scheduling variables. To summarize, two different FP techniques are introduced. Proposition 1 decouples the numerator and denominator of the ratio. Proposition 3 moves the ratio from inside of the logarithm to the outside.

The earlier conference version of this paper [1] further uses the above transforms for scalar FP to solve the optimal scheduling problem in the D2D context, but only the case in which each transmitter or receiver is equipped with a single antenna. This paper aims to develop a matrix extension for the MIMO case.

B. Matrix FP Transforms

The definition of ratio can be naturally generalized to the matrix case. Recall that $\sqrt{\mathbf{A}} \in \mathbb{C}^{n\times n}$ is a square root of matrix $\mathbf{A} \in \mathbb{H}_{+}^{n\times n}$ if $\sqrt{\mathbf{A}}\sqrt{\mathbf{A}}^{\dagger} = \mathbf{A}$. For any pair of $\mathbf{A} \in \mathbb{H}_{+}^{n\times n}$ and $\mathbf{B} \in \mathbb{H}_{++}^{n\times n}$, let $\sqrt{\mathbf{A}}$ be a square root of $\mathbf{A}$, then $\sqrt{\mathbf{A}}^{\dagger}\mathbf{B}^{-1}\sqrt{\mathbf{A}}$ is said to be a matrix ratio between $\mathbf{A}$ and $\mathbf{B}$. The FP transforms of Propositions 1 and 3 can now be generalized. We state these new results in the following.

Theorem 1 (Matrix Quadratic Transform): Given a nonempty constraint set $\mathcal{X}$ as well as a sequence of numerator functions $\mathbf{A}_m(\mathbf{x}) \in \mathbb{H}_{+}^{n\times n}$, denominator functions $\mathbf{B}_m(\mathbf{x}) \in \mathbb{H}_{++}^{n\times n}$, and nondecreasing matrix functions $f_m(\mathbf{Z}) : \mathbb{H}_{+}^{n\times n} \mapsto \mathbb{R}$ in the sense that $f_m(\mathbf{Z}') \geq f_m(\mathbf{Z})$ if $\mathbf{Z}' \succeq \mathbf{Z}$, for $m = 1, 2, \ldots, M$, the *sum-of-functions-of-matrix-ratio* problem

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{m=1}^{M} f_m\left(\sqrt{\mathbf{A}}_m^{\dagger}(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})\right) \tag{12a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{12b}$$

is equivalent to

$$\underset{\mathbf{x},\underline{\mathbf{Y}}}{\text{maximize}} \quad \tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}) \tag{13a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{13b}$$

$$\mathbf{Y}_m \in \mathbb{C}^{n\times n}, \tag{13c}$$

where the new objective function is defined as

$$\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}) = \sum_{m=1}^{M} f_m\Big(2\Re\{\sqrt{\mathbf{A}}_m^{\dagger}(\mathbf{x})\mathbf{Y}_m\} - \mathbf{Y}_m^{\dagger}\mathbf{B}_m(\mathbf{x})\mathbf{Y}_m\Big). \tag{14}$$

Note that (14) implicitly requires that the argument of $f_m(\cdot)$ in (14) is a positive semidefinite matrix.

Proof: To show that (12) is equivalent to (13), we first optimize over $\mathbf{Y}_m$ for fixed $\mathbf{x}$ in (13). This can be done for each term in the summation in $\tilde{f}_q$ separately. Since $f_m(\cdot)$ is assumed to be monotonic, we only need to optimize its argument, which is a quadratic function of $\mathbf{Y}_m$. This optimization has a closed-form solution by completing the square, i.e.,

$$\begin{aligned} &2\Re\{\sqrt{\mathbf{A}}_m^{\dagger}(\mathbf{x})\mathbf{Y}_m\} - \mathbf{Y}_m^{\dagger}\mathbf{B}_m(\mathbf{x})\mathbf{Y}_m \\ &\quad = \sqrt{\mathbf{A}}_m^{\dagger}(\mathbf{x})\mathbf{Y}_m + \mathbf{Y}_m^{\dagger}\sqrt{\mathbf{A}}_m(\mathbf{x}) - \mathbf{Y}_m^{\dagger}\mathbf{B}_m(\mathbf{x})\mathbf{Y}_m \\ &\quad = \sqrt{\mathbf{A}}_m^{\dagger}(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x}) - \mathbf{\Delta}_m^{\dagger}\mathbf{B}_m(\mathbf{x})\mathbf{\Delta}_m, \end{aligned} \tag{15}$$

where $\mathbf{\Delta}_m = \mathbf{Y}_m - \mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})$. We then obtain the optimal $\mathbf{Y}_m^{\star} = \mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})$. Substituting this $\mathbf{Y}_m^{\star}$ in $\tilde{f}_q$ recovers the original problem. ■

The quadratic transform for FP is first developed for the scalar case in Proposition 1, then generalized to the vector case in [2], where the objective function has the form $\sum_{m=1}^{M} f_m(\mathbf{a}_m^\dagger(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\mathbf{a}_m(\mathbf{x}))$, where $\mathbf{a}_m(\mathbf{x}) \in \mathbb{C}^n$, $\mathbf{B}_m(\mathbf{x}) \in \mathbb{S}_{++}^{n\times n}$, and $f_m(z) : \mathbb{R} \mapsto \mathbb{R}$ is a nondecreasing function.

The above result is a further generalization to the matrix case. The scalar FP can be used to model the power control problem for single-antenna links. The vector FP can be used to deal with a special case of MIMO communication [2] where each link has at most one data stream. To reap the full benefit of MIMO, each link needs to carry multiple data streams. In this case, the matrix FP is necessary.

Theorem 2 (Matrix Lagrangian Dual Transform): Given a nonempty constraint set $\mathcal{X}$ as well as a sequence of numerator functions $\mathbf{A}_m(\mathbf{x}) \in \mathbb{H}_+^{n\times n}$, denominator functions $\mathbf{B}_m(\mathbf{x}) \in \mathbb{H}_{++}^{n\times n}$, and nonnegative weights $w_m \geq 0$, for $m = 1, 2, \ldots, M$, the *sum-of-weighted-logarithmic-matrix-ratios* problem

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{m=1}^{M} w_m \log\left|\boldsymbol{I} + \sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})\right| \tag{16a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{16b}$$

is equivalent to

$$\underset{\mathbf{x},\underline{\mathbf{\Gamma}}}{\text{maximize}} \quad f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}) \tag{17a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{17b}$$

$$\mathbf{\Gamma}_m \in \mathbb{H}_+^{n\times n}, \tag{17c}$$

where the new objective function is

$$f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}) = \sum_{m=1}^{M} w_m \Big(\log|\boldsymbol{I}+\mathbf{\Gamma}_m| - \mathrm{tr}(\mathbf{\Gamma}_m) + \mathrm{tr}\Big((\boldsymbol{I}+\mathbf{\Gamma}_m)\cdot \sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\big(\mathbf{A}_m(\mathbf{x}) + \mathbf{B}_m(\mathbf{x})\big)^{-1}\sqrt{\mathbf{A}}_m(\mathbf{x})\Big)\Big). \tag{18}$$

Proof: First, using the Woodbury matrix identity, i.e., $(\mathbf{D} + \mathbf{UCV})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{V}\mathbf{D}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{D}^{-1}$, we can rewrite (18) as

$$f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}) = \sum_{m=1}^{M} w_m \Big(\log|\boldsymbol{I}+\mathbf{\Gamma}_m| + n - \mathrm{tr}\Big((\boldsymbol{I}+\mathbf{\Gamma}_m)\cdot \big(\boldsymbol{I} + \sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})\big)^{-1}\Big)\Big). \tag{19}$$

We then consider the optimization of the above new form of f_r. Note that the optimization over $\mathbf{\Gamma}_m$ can be done separately for each term of the summation. Since each of the terms is concave over $\mathbf{\Gamma}_m$ when $\mathbf{x}$ is fixed, the optimal $\mathbf{\Gamma}_m$ can be determined by setting $\partial f_r/\partial\mathbf{\Gamma}_m$ to zero, i.e.,

$$(\boldsymbol{I}+\mathbf{\Gamma}_m)^{-1} - \Big(\boldsymbol{I} + \sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})\Big)^{-1} = \mathbf{0}. \tag{20}$$

Note that the derivative $\partial f_r/\partial\mathbf{\Gamma}_m$ exists in this case because f_r is a spectral function [35]. Thus, we obtain the optimal $\mathbf{\Gamma}_m^\star = \sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\mathbf{B}_m^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_m(\mathbf{x})$. Substituting this $\mathbf{\Gamma}_m^\star$ in (19) recovers the original problem, thereby establishing the theorem. ∎

Observe that the proposed matrix quadratic transform of Theorem 1 can be applied to decouple the ratio terms of f_r in (18) to further transform the matrix FP, as stated in the corollary below.

Corollary 1: The problem (16) is equivalent to

$$\underset{\mathbf{x},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}}{\text{maximize}} \quad f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}, \underline{\mathbf{Y}}) \tag{21a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{21b}$$

$$\mathbf{\Gamma}_m \in \mathbb{H}_+^{n\times n}, \tag{21c}$$

$$\mathbf{Y}_m \in \mathbb{C}^{n\times n}, \tag{21d}$$

where the new objective function is displayed in (22) at the bottom of this page. Note that $\Re\{\cdot\}$ can be dropped for the term $\sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\mathbf{Y}_m$ because of trace.

Proof: Treating $f_m(\mathbf{Z}) = \mathrm{tr}\big((\boldsymbol{I}+\mathbf{\Gamma}_m)\mathbf{Z}\big)$ as the nondecreasing function, $\sqrt{w_m}\sqrt{\mathbf{A}}_m(\mathbf{x})$ as the square root of the numerator, and $\mathbf{A}_m(\mathbf{x})+\mathbf{B}_m(\mathbf{x})$ as the denominator, we apply the matrix quadratic transform of Theorem 1 to the last term of f_r in (18) to obtain the above reformulation. ∎

Note that the new objective function f_q is a linear function of each $\sqrt{w_m}\sqrt{\mathbf{A}}_m(\mathbf{x})$ and $\mathbf{B}_m(\mathbf{x})$, while keeping all other terms fixed. This facilitates algorithm design for solving the matrix FP problem. We also remark that there are also other ways of applying the matrix quadratic transform to f_r in (18) by choosing different matrix ratios and functions $f_m(\cdot)$. The advantage of the above decomposition as compared to the alternatives is discussed in detail in Section V-F.

IV. Fractional Programming Transform as Minorization Maximization

An important theoretical observation of this paper is that the matrix FP transform proposed above can be recast in the MM framework. First, we give a brief introduction to MM. Consider a general optimization problem:

$$\underset{\mathbf{x}}{\text{maximize}} \quad f(\mathbf{x}) \tag{23a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{23b}$$

where $f(\mathbf{x})$ is not assumed to be concave. Because of the nonconvexity, it is not always easy to solve the problem directly. The core idea behind the MM algorithm is to successively solve a sequence of *well-chosen* approximations of the original problem [36], [37]. Specifically, at point $\hat{\mathbf{x}} \in \mathcal{X}$, the MM algorithm approximates problem (23) as

$$\underset{\mathbf{x}}{\text{maximize}} \quad g(\mathbf{x}|\hat{\mathbf{x}}) \tag{24a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{24b}$$

where $g(\mathbf{x}|\hat{\mathbf{x}})$ is referred to as the *surrogate function* and is defined by the following two conditions:

- C1: $g(\mathbf{x}|\hat{\mathbf{x}}) \leq f(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$;
- C2: $g(\hat{\mathbf{x}}|\hat{\mathbf{x}}) = f(\hat{\mathbf{x}})$.

$$f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}, \underline{\mathbf{Y}}) = \sum_{m=1}^{M} \Big(w_m \log|\boldsymbol{I}+\mathbf{\Gamma}_m| - w_m \mathrm{tr}\big(\mathbf{\Gamma}_m\big) + \mathrm{tr}\Big((\boldsymbol{I}+\mathbf{\Gamma}_m)\Big(2\sqrt{w_m}\sqrt{\mathbf{A}}_m^\dagger(\mathbf{x})\mathbf{Y}_m - \mathbf{Y}_m^\dagger\big(\mathbf{A}_m(\mathbf{x})+\mathbf{B}_m(\mathbf{x})\big)\mathbf{Y}_m\Big)\Big)\Big). \tag{22}$$

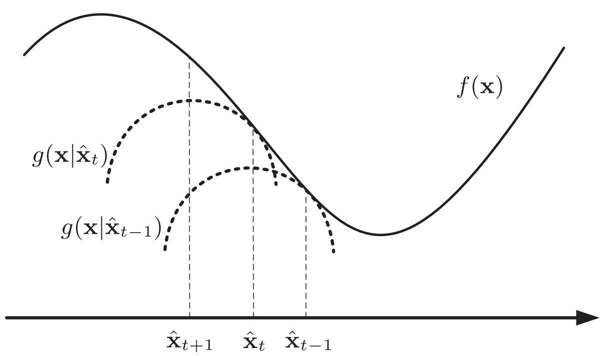


Fig. 3. The iterative optimization by the MM algorithm. Observe that $f(\hat{\mathbf{x}})$ is monotonically nondecreasing after each iteration.

The MM algorithm updates $\hat{\mathbf{x}}$ iteratively as follows:

$$\hat{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}|\hat{\mathbf{x}}_t), \tag{25}$$

where subscript t is the iteration index. Note that the function value of $f(\hat{\mathbf{x}})$ is nondecreasing after each iteration because

$$f(\hat{\mathbf{x}}_{t+1}) \overset{(a)}{\geq} g(\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t) \overset{(b)}{\geq} g(\hat{\mathbf{x}}_t|\hat{\mathbf{x}}_t) \overset{(c)}{=} f(\hat{\mathbf{x}}_t), \tag{26}$$

where (a) follows by C1, (b) follows by the optimality of $\hat{\mathbf{x}}_{t+1}$ in (25), and (c) follows by C2. This is illustrated in Fig. 3.

The following proposition gives a convergence analysis of the MM algorithm.

Proposition 4: Let $\hat{\mathbf{x}}_t$ be the solution produced by the MM update (25) after t iterations. The function value $f(\hat{\mathbf{x}}_t)$ converges in a nondecreasing fashion in t. Further, the variable $\hat{\mathbf{x}}_t$ converges to a stationary point solution to the original optimization problem (23) if the following three conditions are satisfied: (i) $f(\mathbf{x})$ is continuous over a convex closed set $\mathcal{X}$; (ii) $g(\mathbf{x}|\hat{\mathbf{x}})$ is continuous in $(\mathbf{x}, \hat{\mathbf{x}})$; (iii) $f(\mathbf{x})$ and $g(\mathbf{x}|\hat{\mathbf{x}})$ are differentiable with respect to $\mathbf{x}$ given $\hat{\mathbf{x}}$.

Proof: The non-decreasing convergence of $f(\hat{\mathbf{x}})$ is already verified in (26). Further, combining the above condition (iii) with the conditions C1 and C2, we obtain that $f(\mathbf{x})$ and $g(\mathbf{x}|\hat{\mathbf{x}})$ have the same gradient with respect to $\mathbf{x}$ at $\mathbf{x} = \hat{\mathbf{x}}$. This result, along with the above conditions (i) and (ii), guarantees that $\hat{\mathbf{x}}_t$ converges to a stationary point solution to the original optimization problem (23) according to [36]. We remark that the proof can be adapted to the case where $\mathbf{x}$ is a complex variable; the argument is similar to that of [38]. ∎

The MM algorithm is a framework rather than an algorithmic prescription, because the algorithm depends on the specific choice of the surrogate function. If $f(\cdot)$ is twice differentiable, its second order Taylor expansion is often the first candidate to check to see whether it is suitable as a surrogate function. For more general functions, many of the ingenious ways of constructing a surrogate function have been documented in [37].

The main point of this section is that the proposed matrix FP transforms can be interpreted in the MM framework as a way of constructing surrogate functions of the original problems, as stated below.

Theorem 3: Consider the matrix quadratic transform in Theorem 1, if we consider the optimal $\mathbf{Y}^\star_m$ as a function of $\hat{\mathbf{x}}$ and substitute it into $\tilde{f}_q$ in (14), then the new objective function $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$, where

$$\mathbf{Y}_m(\hat{\mathbf{x}}) = \mathbf{B}_m^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}}) \tag{27}$$

is a surrogate function of the objective function of the optimization problem (12).

Proof: We use $f_{\mathrm{I}}(\mathbf{x})$ to denote the objective function in (12a). Substitute $\mathbf{Y}_m(\hat{\mathbf{x}}) = \mathbf{B}_m^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}})$ back in $\tilde{f}_q$. We aim to show that $g(\mathbf{x}|\hat{\mathbf{x}}) = \tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$ is a surrogate function of $f_{\mathrm{I}}(\mathbf{x})$.

As already shown in the proof of Theorem 1, $\underline{\mathbf{Y}}(\mathbf{x})$ is the optimum solution for the maximization of $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}})$ over $\underline{\mathbf{Y}}$ when $\mathbf{x}$ is fixed. So, $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}})) \leq \tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\mathbf{x})), \forall \hat{\mathbf{x}}, \mathbf{x}$. Further, it can be seen that $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\mathbf{x})) = f_{\mathrm{I}}(\mathbf{x})$ for any $\mathbf{x}$.

Thus, for each fixed $\hat{\mathbf{x}}$, we have $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}})) \leq f_{\mathrm{I}}(\mathbf{x})$, $\forall \mathbf{x}$, and $\tilde{f}_q(\hat{\mathbf{x}}, \underline{\mathbf{Y}}(\hat{\mathbf{x}})) = f_{\mathrm{I}}(\hat{\mathbf{x}})$, thus verifying the conditions C1 and C2 for $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$ to be a surrogate function of $f_{\mathrm{I}}(\mathbf{x})$. ∎

Theorem 4: Consider the matrix Lagrangian dual transform in Theorem 2, if we consider the optimal $\mathbf{\Gamma}^\star_m$ as a function of $\hat{\mathbf{x}}$ and substitute it into f_r in (18), then the new objective function $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$, where

$$\mathbf{\Gamma}_m(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}^\dagger_m(\hat{\mathbf{x}})\mathbf{B}_m^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}}) \tag{28}$$

is a surrogate function of the objective function of the optimization problem (16).

Proof: We use $f_{\mathrm{II}}(\mathbf{x})$ to denote the objective function in (16a). We substitute $\mathbf{\Gamma}_m(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}^\dagger_m(\hat{\mathbf{x}})\mathbf{B}_m^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}})$ back in f_r, and aim to show that $g(\mathbf{x}|\hat{\mathbf{x}}) = f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$ is a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$.

As shown in the proof of Theorem 2, $\underline{\mathbf{\Gamma}}(\mathbf{x})$ is the optimal solution to maximizing $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}})$ over $\underline{\mathbf{\Gamma}}$ when $\mathbf{x}$ is fixed, so $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) \leq f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\mathbf{x})), \forall \mathbf{x}, \hat{\mathbf{x}}$. Also, it holds true that $f_r(\hat{\mathbf{x}}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) = f_{\mathrm{II}}(\hat{\mathbf{x}})$, $\forall \hat{\mathbf{x}}$. Combining the above results, we see that the conditions C1 and C2 are satisfied, thus $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$ is a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$. ∎

Corollary 2: Consider the transform in Corollary 1, if we consider the optimal $\mathbf{\Gamma}^\star_m$ and the optimal $\mathbf{Y}^\star_m$ as two functions of $\hat{\mathbf{x}}$, and substitute them into f_q, then the new objective function $f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}), \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$, where

$$\mathbf{\Gamma}_m(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}^\dagger_m(\hat{\mathbf{x}})\mathbf{B}_m^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}}) \tag{29}$$

and

$$\mathbf{Y}_m(\hat{\mathbf{x}}) = \big(\mathbf{A}_m(\hat{\mathbf{x}}) + \mathbf{B}_m(\hat{\mathbf{x}})\big)^{-1}\big(\sqrt{w_m}\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}})\big), \tag{30}$$

is a surrogate function of the objective function of the optimization problem (16).

Proof: Again, let $f_{\mathrm{II}}(\mathbf{x})$ be the objective function in (16a). We introduce two new variables $\hat{\mathbf{x}}$ and $\hat{\hat{\mathbf{x}}}$, and substitute $\mathbf{\Gamma}_m(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}^\dagger_m(\hat{\mathbf{x}})\mathbf{B}_m^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_m(\hat{\mathbf{x}})$ and $\mathbf{Y}_m(\hat{\hat{\mathbf{x}}}) = \big(\mathbf{A}_m(\hat{\hat{\mathbf{x}}}) + \mathbf{B}_m(\hat{\hat{\mathbf{x}}})\big)^{-1}\big(\sqrt{w_m}\sqrt{\mathbf{A}}_m(\hat{\hat{\mathbf{x}}})\big)$ back in f_q and f_r. Let $g_1(\mathbf{x}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}}) = f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}), \underline{\mathbf{Y}}(\hat{\hat{\mathbf{x}}}))$, and $g_2(\mathbf{x}|\hat{\mathbf{x}}) = f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$.

According to Theorem 4, $g_2(\mathbf{x}|\hat{\mathbf{x}})$ is a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$ in the sense that $g_2(\mathbf{x}|\hat{\mathbf{x}}) \leq f_{\mathrm{II}}(\mathbf{x})$ and $g_2(\hat{\mathbf{x}}|\hat{\mathbf{x}}) = f_{\mathrm{II}}(\hat{\mathbf{x}}), \forall \mathbf{x}, \hat{\mathbf{x}}$. According to Theorem 3, $g_1(\mathbf{x}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}})$ is a surrogate function with respect to f_r in the sense that $g_1(\mathbf{x}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}}) \leq f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$ and $g_1(\hat{\hat{\mathbf{x}}}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}}) = f_r(\hat{\hat{\mathbf{x}}}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})), \forall \mathbf{x}, \hat{\mathbf{x}}, \hat{\hat{\mathbf{x}}}$.

Combining these results and fixing $\hat{\hat{\mathbf{x}}} = \hat{\mathbf{x}}$, we obtain $g_1(\mathbf{x}|\hat{\mathbf{x}}, \hat{\mathbf{x}}) \leq f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) = g_2(\mathbf{x}|\hat{\mathbf{x}}) \leq f_{\mathrm{II}}(\mathbf{x}), \forall \mathbf{x}$ and $g_1(\hat{\mathbf{x}}|\hat{\mathbf{x}}, \hat{\mathbf{x}}) = f_r(\hat{\mathbf{x}}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) = g_2(\hat{\mathbf{x}}|\hat{\mathbf{x}}) = f_{\mathrm{II}}(\hat{\mathbf{x}})$, thereby verifying the conditions C1 and C2 for $f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}), \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$ to be a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$. ∎

This MM interpretation of the FP transforms provides a theoretical basis for the proposed FPLinQ strategy for joint scheduling, beamforming and power control. Note that the above results carry over to Propositions 1 and 3 for the scalar FP case, so the approach of [2], [3] can be interpreted as an MM algorithm as well.

V. Joint Scheduling, Power Control and Beamforming Using FPLinQ

A. Iterative Optimization via Matrix FP

We propose to solve the joint scheduling and beamforming problem (3) iteratively by first reformulating it using Corollary 1. Specifically, after specializing the variable $\mathbf{x}$ in (16) to be the variables $(\underline{\mathbf{V}}, \underline{s})$ in (3), we obtain the following reformulation:

Theorem 5: The joint beamforming and link scheduling problem (3) is equivalent to

$$\begin{aligned} \underset{\underline{s},\underline{\mathbf{V}},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}}{\text{maximize}} \quad & f_q(\underline{s}, \underline{\mathbf{V}}, \underline{\mathbf{\Gamma}}, \underline{\mathbf{Y}}) && (31a)\\ \text{subject to} \quad & (3b), (3c), (3d), \\ & \mathbf{\Gamma}_j \in \mathbb{H}_+^{N\times N}, && (31b)\\ & \mathbf{Y}_j \in \mathbb{C}^{N\times N}, && (31c) \end{aligned}$$

where the new objective function f_q is shown in (32a) as displayed at the bottom of the page.

Proof: The reformulating steps directly follow Corollary 1. We remark that f_q can be rewritten as in (32b), which enables an efficient optimization by matching. ■

We now optimize over the variables of the new problem (31) in an iterative manner. First, when $\underline{s}$ and $\underline{\mathbf{V}}$ are both held fixed, the auxiliary variables $\underline{\mathbf{\Gamma}}$ and $\underline{\mathbf{Y}}$ can be optimally determined as

$$\mathbf{\Gamma}_j^\star = \mathbf{V}_{s_j}^\dagger \mathbf{H}_{js_j}^\dagger \mathbf{F}_j^{-1} \mathbf{H}_{js_j} \mathbf{V}_{s_j} \tag{33}$$

and

$$\mathbf{Y}_j^\star = \left(\mathbf{F}_j + \mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{V}_{s_j}^\dagger \mathbf{H}_{js_j}^\dagger\right)^{-1} \sqrt{w_{js_j}}\, \mathbf{H}_{js_j}\mathbf{V}_{s_j}. \tag{34}$$

We remark that the implicit constraints as stated in Theorem 1 are automatically satisfied by the above optimal solution of the auxiliary variable $\mathbf{Y}_j^\star$.

It remains to optimize the beamforming variable $\underline{\mathbf{V}}$ and the scheduling variable $\underline{s}$. The key idea is to formulate the problem as a bipartite weighted matching problem, which is described in detail below.

B. Scheduling and Beamforming via Bipartite Matching

We consider the objective function f_q of the form (32b). The key observation is that the beamformer of each link (if it is scheduled) can be optimally determined from f_q, even without knowing the scheduling decisions for the nearby links. To formalize this idea, let $\tilde{\mathbf{V}}_{ji}$ be the tentative value of $\mathbf{V}_i^\star$ if link (i, j) is scheduled. By completing the square in f_q, we can compute $\tilde{\mathbf{V}}_{ji}$ as

$$\begin{aligned} \tilde{\mathbf{V}}_{ji} = & \left(\mu_{ji}\boldsymbol{I} + \sum_{j'\in\mathcal{J}} \mathbf{H}_{j'i}^\dagger \mathbf{Y}_{j'}(\boldsymbol{I}+\mathbf{\Gamma}_{j'})\mathbf{Y}_{j'}^\dagger \mathbf{H}_{j'i}\right)^{-1} \cdot \\ & \sqrt{w_{ji}}\, \mathbf{H}_{ji}^\dagger \mathbf{Y}_j(\boldsymbol{I}+\mathbf{\Gamma}_j), \end{aligned} \tag{35}$$

where μ_{ji} is a Lagrangian multiplier for the power constraint (3b), optimally determined as

$$\mu_{ji}^\star = \min\{\mu_{ji} \geq 0 : \mathrm{tr}(\tilde{\mathbf{V}}_{ji}^\dagger \tilde{\mathbf{V}}_{ji}) \leq P_{\max}\}, \tag{36}$$

which can be computed efficiently by bisection search since $\tilde{\mathbf{V}}_{ji}$ is monotonically decreasing with μ_{ji}. The solution $\tilde{\mathbf{V}}_{ji}$ in (35) has the same structure as an MMSE beamformer.

We now turn to the question of which $\tilde{\mathbf{V}}_{ji}$ should be chosen to be $\mathbf{V}_i$ so as to maximize f_q. This is akin to a scheduling step of choosing the best transmitter i for each receiver j. The key is to recognize this question as a weighted bipartite matching problem:

$$\begin{aligned} \underset{\underline{q}}{\text{maximize}} \quad & \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{K}_j} \lambda_{ji} q_{ji} && (37a)\\ \text{subject to} \quad & \sum_{i\in\mathcal{K}_j} q_{ji} \leq 1, && (37b)\\ & \sum_{j\in\mathcal{L}_i} q_{ji} \leq 1, && (37c)\\ & q_{ji} \in \{0, 1\}, && (37d)\\ & q_{ji} = 0 \ \text{ if } i \notin \mathcal{K}_j \text{ or } j \notin \mathcal{L}_i, && (37e) \end{aligned}$$

where the weight λ_{ji} is evaluated by (38) as displayed at the bottom of the next page, and q_{ji} is the matching variable between the associated transmitters and receivers. This weighted bipartite matching problem can be solved optimally in polynomial time by using well-known approaches such as the Hungarian algorithm [39] or the auction algorithm [40].

Note that (37) is typically a *sparse* matching problem, since most pairs of $(i, j) \in \mathcal{I}\times\mathcal{J}$ are not associated, so the auction algorithm is likely to be more efficient than the Hungarian algorithm. The matching variable q_{ji} indicates $\mathbf{V}_i$ should be set to which of the $\tilde{\mathbf{V}}_{ji}$. Mathematically, $\underline{\mathbf{V}}$ is recovered as

$$\mathbf{V}_i^\star = \begin{cases} \tilde{\mathbf{V}}_{ji}, \text{ if } \ q_{ji} = 1 \text{ for some } j; \\ \mathbf{0}, \quad \text{otherwise}. \end{cases} \tag{39}$$

After updating $\underline{\mathbf{V}}$, the final step is to update the scheduling variable $\underline{s}$ for the fixed $\underline{\mathbf{V}}$. This is again a weighted bipartite

$$\begin{aligned} & f_q(\underline{s}, \underline{\mathbf{V}}, \underline{\mathbf{\Gamma}}, \underline{\mathbf{Y}}) \\ & = \sum_{j\in\mathcal{J}} \Big(w_{js_j}\log|\boldsymbol{I}+\mathbf{\Gamma}_j| - w_{js_j}\mathrm{tr}(\mathbf{\Gamma}_j) + \mathrm{tr}\Big((\boldsymbol{I}+\mathbf{\Gamma}_j)\Big(2\sqrt{w_{js_j}}\,\mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{Y}_j^\dagger - \mathbf{Y}_j^\dagger\big(\mathbf{F}_j + \mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{V}_{s_j}^\dagger\mathbf{H}_{js_j}^\dagger\big)\mathbf{Y}_j\Big)\Big)\Big) && (32a)\\ & = \sum_{j\in\mathcal{J}} \bigg(w_{js_j}\log|\boldsymbol{I}+\mathbf{\Gamma}_j| - w_{js_j}\mathrm{tr}(\mathbf{\Gamma}_j) + \mathrm{tr}\Big(2\sqrt{w_{js_j}}\,(\boldsymbol{I}+\mathbf{\Gamma}_j)\mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{Y}_j^\dagger - \sum_{j'\in\mathcal{J}}(\boldsymbol{I}+\mathbf{\Gamma}_{j'})\mathbf{Y}_{j'}^\dagger\mathbf{H}_{j's_j}\mathbf{V}_{s_j}\mathbf{V}_{s_j}^\dagger\mathbf{H}_{j's_j}^\dagger\mathbf{Y}_{j'}\Big)\bigg) \\ & \quad + \sum_{j\in\mathcal{J}} \sigma^2 \mathbf{Y}_j^\dagger(\boldsymbol{I}+\mathbf{\Gamma}_j)\mathbf{Y}_j. && (32b) \end{aligned}$$

matching problem, but now since $\mathbf{V}_i$ is fixed, this amounts to choosing the best receiver j for each transmitter i:

$$\underset{\underline{q}}{\text{maximize}} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{L}_i} w_{ji}r_{ji}q_{ji} \tag{40a}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{K}_j} q_{ji} \le 1, \tag{40b}$$

$$\sum_{j\in\mathcal{L}_i} q_{ji} \le 1, \tag{40c}$$

$$q_{ji} \in \{0,1\}, \tag{40d}$$

$$q_{ji} = 0 \ \text{ if } i \notin \mathcal{K}_j \text{ or } j \notin \mathcal{L}_i, \tag{40e}$$

where $w_{ji}r_{ji}$ is the weighted achievable rate if the receiver j is scheduled for transmitter i under fixed $\mathbf{V}_i$. Note that since $\underline{\mathbf{V}}$ is fixed, r_{ij} can be computed independently of the schedule, using an expression similar to (1). This problem can again be solved in polynomial time. The optimal schedule is then determined from the optimal q_{ij} as

$$s_j^\star = \begin{cases} i, & \text{if } \ q_{ji} = 1 \text{ for some } i; \\ \varnothing, & \text{otherwise.} \end{cases} \tag{41}$$

We note that the reason for having two sets of matching is because we allow a general network model in which each transmitter may associate with multiple receivers and each receiver may associate with multiple transmitters. For simpler D2D model such as the one in Fig. 1(a), these two matching steps would not have been necessary, as in [1].

Algorithm 2 Proposed FPLinQ Strategy for D2D Link Scheduling With Power Control and Beamforming

1 Initialize all the variables to feasible values;
2 **repeat**
3   Update $\underline{\mathbf{\Gamma}}$ according to (33);
4   Update $\underline{\mathbf{Y}}$ according to (34);
5   Solve the weighted bipartite matching (37);
6   Update $\underline{\mathbf{V}}$ according to (39);
7   Update $\underline{s}$ by the weighted bipartite matching (41);
8 **until** *the weighted sum rate converges*;

Combining all the above steps together yields the FPLinQ strategy. Algorithm 2 summarizes the overall approach.

A desirable feature of FPLinQ as compared to FlashLinQ, ITLinQ and ITLinQ+ is that no tuning of design parameters is needed. But, FPLinQ is also somewhat more difficult to implement in a distributed fashion than FlashLinQ, ITLinQ, and ITLinQ+, because it additionally requires the update of the auxiliary variables $\underline{\mathbf{\Gamma}}$ and $\underline{\mathbf{Y}}$ per iteration.

C. Alleviating Premature Turning-Off

It is worthwhile to take a deeper look into Algorithm 2 to understand how FPLinQ is able to alleviate the *premature turning-off* problem. FPLinQ differs from the BCD method mainly in Step 5, where the beamforming variable $\underline{\mathbf{V}}$ is optimized for the new objective function f_q instead of the weighted sum-rate objective function. Taking a close look at (37) and (39), we can see that the update of each $\mathbf{V}_i$ at Step 5 of FPLinQ is not affected by the current value of $\underline{s}$. From the MM interpretation, we see that when updating $\underline{\mathbf{V}}$ for fixed $\underline{s}$, FPLinQ is actually using the surrogate function to mimic the original objective function so that the optimization over $\underline{\mathbf{V}}$ no longer relies on $\underline{s}$. In comparison to the BCD method, this less aggressive update of $\underline{\mathbf{V}}$ by FPLinQ allows the existing OFF-transmitters to be reactivated, thereby alleviating the premature turning-off issue as mentioned in Section II-B.

D. Convergence Analysis

We now examine the convergence behavior of the proposed algorithm by utilizing the MM interpretation as a tool.

Theorem 6: The weighted sum rate across all the D2D links is nondecreasing after each iteration of Algorithm 2, so the objective function of the optimization problem is guaranteed to converge. Furthermore, at convergence, for fixed $\underline{s}$, the solution $\underline{\mathbf{V}}$ is a stationary point of the problem (3).

Proof: We prove convergence based on the MM interpretation of the FP transforms. The Step 3 and Step 4 of the algorithm construct the surrogate functions as defined in Theorem 4 and Theorem 3. Step 5 of Algorithm 2 performs the maximization step of the MM algorithm, so the weighted sum rate must be nondecreasing after Step 5, by (26). Step 6 further optimizes the link schedule, so the weighted sum rate is nondecreasing after Step 6. Since the optimization objective is nondecreasing and is bounded above, Algorithm 2 must converge in objective value.

The weighted sum rate is a differentiable function over $\underline{\mathbf{V}}$ under fixed $\underline{s}$. Further, the conditions of Proposition 4 are satisfied. So, at convergence, the solution of $\underline{\mathbf{V}}$ given by Algorithm 2 must be a stationary point according to the proof of Proposition 4. ■

We remark that proving the convergence of Algorithm 2 without the MM interpretation would have been much more cumbersome.

Regarding the rate of convergence, the *spectral radius*

$$\rho = 1 - \min_{\mathbf{u}\neq 0} \frac{\mathbf{u}^\dagger \cdot \nabla^2 \ f(\mathbf{x}) \cdot \mathbf{u}}{\mathbf{u}^\dagger \cdot \nabla^2 \ g(\mathbf{x}|\hat{\mathbf{x}}_\infty) \cdot \mathbf{u}} \tag{43}$$

has been proposed to compare the different MM algorithms [41]. In principle, ρ reflects how well the surrogate function $g(\mathbf{x}|\hat{\mathbf{x}})$ approximates the original objective function $f(\mathbf{x})$ in terms of the second moment—smaller ρ indicates tighter approximation and thus faster convergence. However, this type of analysis has limited value in our problem case because: (i) it requires the updating function of $\mathbf{x}$ to be differentiable whereas our problem involves discrete variables; (ii) computing ρ entails solving a difficult nonconvex problem; (iii) it only characterizes the local convergence in the proximity of $\mathbf{x}_\infty$. For these reasons, the rate of convergence will be compared numerically for the different examples in Section VI.

E. Complexity Analysis

We now analyze the complexity of FPLinQ (i.e., Algorithm 2). We assume that there are a total of L D2D links

$$\lambda_{ji} = w_{ji}\log|\boldsymbol{I}+\mathbf{\Gamma}_j| - w_{ji}\mathrm{tr}(\mathbf{\Gamma}_j) + \mathrm{tr}\left(2\sqrt{w_{ji}}\,(\boldsymbol{I}+\mathbf{\Gamma}_j)\mathbf{Y}_j^\dagger\mathbf{H}_{ji}\tilde{\mathbf{V}}_{ji} - \sum_{j'\in\mathcal{J}}(\boldsymbol{I}+\mathbf{\Gamma}_{j'})\mathbf{Y}_{j'}^\dagger\mathbf{H}_{j'i}\tilde{\mathbf{V}}_{ji}\tilde{\mathbf{V}}_{ji}^\dagger\mathbf{H}_{j'i}^\dagger\mathbf{Y}_{j'}\right). \tag{38}$$

TABLE I
COMPARISON OF LINK SCHEDULING ALGORITHMS FOR D2D NETWORKS

	FPLinQ	FlashLinQ [24]	ITLinQ [25]	ITLinQ+ [26]	BCD [27]
Scheduling & Association	Flexible	Single	Single	Single	Flexible
Power Control	✓	✗	✗	✓	✓
Beamforming	✓	✗	✗	✗	✓
Tuning Parameters	Not Needed	Required	Required	Required	Not Needed
Convergence with Fixed Schedule	Stationary Point	–	–	–	Stationary Point
Computational Complexity	$O(L^2(N^4+\log L))$	$O(L^2)$	$O(L^2)$	$O(L^2)$	$O(L^2(N^4+\log L))$
Communication Complexity	$O(N^2L^2)$	$O(L^2)$	$O(L^2)$	$O(L^2)$	$O(N^2L^2)$
Link Reactivation	✓	✗	✗	✗	✗

in the network; each transmitter/receiver is associated with a small number (i.e., constant number) of neighboring devices, so that $|\mathcal{I}| = O(L)$ and $|\mathcal{J}| = O(L)$. To ease the analysis, we assume that FPLinQ runs for a fixed number of iterations.

Communication Complexity: In each iteration of FPLinQ, each transmitter i requires the tuple $(\underline{\mathbf{\Gamma}}, \underline{\mathbf{Y}}, \underline{s})$ to update $\mathbf{V}_i$, while every receiver j requires $\underline{\mathbf{V}}$ to update $\mathbf{\Gamma}_j$ and $\mathbf{Y}_j$. Each of $\mathbf{V}_i, \mathbf{\Gamma}_j, \mathbf{Y}_j$ is an $N \times N$ matrix. Further, the channel coefficients from $O(L^2)$ direct and interfering channels are needed, with each channel being an $N \times N$ matrix. Thus, the total communication complexity of these updates is $O(N^2L^2)$. The two matchings in Step 5 and Step 6 require the matching weights of all the links, thus introducing a communication complexity of $O(L)$. The overall communication complexity of FPLinQ is then $O(N^2L^2)$. In the single-antenna single-association case, the communication complexity of FPLinQ in each iteration is $O(L^2)$; in comparison, the communication complexity of each step of FlashLinQ, ITLinQ, and ITLinQ+ is also $O(L^2)$, as they all require the $O(L^2)$ channel coefficients.

Computational Complexity: We first consider the update steps of FPLinQ prior to matching, which as analyzed in [3] has a per-iteration computational complexity of $O(N^4\ L^2)$. The matching step can be performed using the auction algorithm [40], which has a computational complexity of $O(L|\mathcal{I}|\log|\mathcal{I}| + L|\mathcal{J}|\log|\mathcal{J}|) = O(L^2\log(L))$. Thus, the overall per-iteration computational complexity of FPLinQ is $O(N^4L^2 + L^2\log(L))$. In the single-antenna single-association case, the per-iteration computational complexity of FPLinQ reduces to $O(L^2\log L)$, while the total computational complexities of FlashLinQ, ITLinQ, and ITLinQ+ are all equal to $O(L^2)$.

We observe that the computational complexity of FPLinQ is sensitive to the number of antennas N (mainly due to the matrix inverse). Overall, asymptotically, FPLinQ has the same communication complexity, but higher computational complexity than the greedy based approaches—FlashLinQ, ITLinQ, and ITLinQ+. Note that although the joint scheduling and power control problem is NP-hard in general [29], [30], recent results nevertheless show that scalable implementation is feasible for a metropolitan-scale network with thousands of terminals [12], [42]. In particular, [42] uses the scalar FP method of [2], [3].

Table I summarizes the comparison between the proposed FPLinQ algorithm and the main benchmarks. The main advantage of FPLinQ is that it allows for flexible association, guarantees convergence without needing tuning parameters, while alleviating the potential pre-mature turn-off problem.

F. Different Ways to Decouple the Ratios

In the derivation of FPLinQ, we decouple the matrix ratios of f_r shown in (42) at the bottom of the page in a particular form, but such decoupling is not unique. There exist other ways to decouple the ratio.

Recall that the proposed reformulation in Theorem 5 follows the proof of Corollary 1 by treating $(\boldsymbol{I} + \mathbf{\Gamma}_m)$ as the fixed weight, with $f_m(\mathbf{Z}) = \mathrm{tr}((\boldsymbol{I} + \mathbf{\Gamma}_m)\mathbf{Z})$, where $\mathbf{Z}$ is a matrix ratio, i.e., $(\boldsymbol{I} + \mathbf{\Gamma})\mathbf{Z}$ is

$$(\boldsymbol{I}+\mathbf{\Gamma})\boxed{\sqrt{w}\mathbf{V}^\dagger\mathbf{H}^\dagger}\underset{\sim\sim\sim\sim\sim\sim\sim}{\left(\mathbf{F}+\mathbf{H}\mathbf{V}\mathbf{V}^\dagger\mathbf{H}^\dagger\right)}^{-1}\boxed{\sqrt{w}\mathbf{H}\mathbf{V}} \tag{43}$$

as in (42). Here, the boxed component represents the numerator and the underlined component the denominator; all the subscripts are omitted for notational simplicity.

The matrix ratio in (43) can also be decoupled in other ways. For instance, we could have included the term $(\boldsymbol{I} + \mathbf{\Gamma})$ in the numerator, i.e.,

$$\boxed{\sqrt{\mathbf{A}}^\dagger}\underset{\sim\sim\sim\sim\sim\sim\sim}{\left(\mathbf{F}+\mathbf{H}\mathbf{V}\mathbf{V}^\dagger\mathbf{H}^\dagger\right)}^{-1}\boxed{\sqrt{\mathbf{A}}}, \tag{44}$$

where

$$\mathbf{A} = w\mathbf{H}\mathbf{V}(\boldsymbol{I}+\mathbf{\Gamma})\mathbf{V}^\dagger\mathbf{H}^\dagger. \tag{45}$$

In fact, the above decoupling is exactly what [2] and [3] use when treating scalar FP problems. However, the inclusion of the $(\boldsymbol{I} + \mathbf{\Gamma})$ term would result in an extra matrix decomposition step when computing the matrix square root, hence the resulting algorithm would be somewhat computationally more complex.

Alternatively, we could have excluded w from $\mathbf{A}$, i.e.,

$$(\boldsymbol{I}+\mathbf{\Gamma})w\boxed{\mathbf{V}^\dagger\mathbf{H}^\dagger}\underset{\sim\sim\sim\sim\sim\sim\sim}{\left(\mathbf{F}+\mathbf{H}\mathbf{V}\mathbf{V}^\dagger\mathbf{H}^\dagger\right)}^{-1}\boxed{\mathbf{H}\mathbf{V}}. \tag{46}$$

The above pattern yields yet another different f_q. It turns out that optimizing $\underline{\mathbf{V}}$, $\underline{\mathbf{\Gamma}}$, and $\underline{\mathbf{Y}}$ iteratively for this particular

$$f_r(\underline{s}, \underline{\mathbf{V}}, \underline{\mathbf{\Gamma}}) = \sum_{j\in\mathcal{J}} w_{js_j}\left(\log|\boldsymbol{I}+\mathbf{\Gamma}_j| - \mathrm{tr}(\mathbf{\Gamma}_j) + \mathrm{tr}\left((\boldsymbol{I}+\mathbf{\Gamma}_j)\mathbf{V}_{s_j}^\dagger\mathbf{H}_{js_j}^\dagger\left(\mathbf{F}_j+\mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{V}_{s_j}^\dagger\mathbf{H}_{js_j}^\dagger\right)^{-1}\mathbf{H}_{js_j}\mathbf{V}_{s_j}\right)\right). \tag{42}$$

f_q is exactly the WMMSE algorithm [13], [14] for beamforming. (This connection to WMMSE has been shown for the vector FP case in [3].) As a corollary, this implies that there is a connection between the WMMSE algorithm and the MM algorithm as well! However, since w_{js_j} contains the scheduling decision, this approach leads us to the situation that $\mathbf{V}_i$'s are updated only for the ON-links, thus it suffers from the premature turning-off problem.

VI. Simulation Results

We validate the performance of FPLinQ through comparison with the benchmark methods for a D2D network in a 1km$\times$1km square area where the D2D links are randomly located. Following [24]–[26], we adopt the short-range outdoor channel model ITU-1411 and use a 5MHz-wide frequency band centered at 2.4GHz. Moreover, the antenna height of each device is 1.5m; the antenna gain is 2.5dBi; the noise power spectrum density is -169dBm/Hz; the noise figure is 7dB; the maximum transmit power is 20dBm; the shadowing is modeled as a Gaussian random variable in decibel with the standard deviation of 10; the distance between the transmitter and receiver of each link is uniformly distributed between 2m and 65m.

The first simulation setting follows [24]–[26]: given a set of links with single-antenna transmitters/receivers and fixed single association (as shown in Fig. 1), the aim is to maximize the sum rate across the links. We use FlashLinQ [24], ITLinQ [25], and ITLinQ [26] as benchmarks. The BCD method is equivalent to FPLinQ in this single-association case. Because the benchmark methods do not have power control, for fair comparison, we modify FPLinQ slightly to restrict the power to be either zero or the maximum, i.e., round each $\mathbf{V}_i$ to $\{0, \sqrt{P_{\max}}\}$. This new version of FPLinQ without power control is referred to as "FPLinQ (no pc)". Further, we introduce two baselines: one is to activate all the links and the other is to activate the links greedily to meet the TIN condition.

Fig. 4 shows the sum rate versus the total number of D2D links. Observe that ITLinQ+ outperforms ITLinQ, and ITLinQ outperforms FlashLinQ, as expected from the previous literature [25], [26]. Without power control, FPLinQ (no pc) significantly outperforms FlashLinQ, ITLinQ, and ITLinQ+, especially when the D2D links are densely located in the area. In particular, observe that Greedy TIN is even worse than simply scheduling all the links because it is too conservative about the effect of interference. Further, as suggested in [26], we run ITLinQ+ and the power control algorithm (e.g., the WMMSE method) alternatively in order to account for joint scheduling and power control; this method is referred to as "ITLinQ (pc)". However, the performance of ITLinQ+ with power control is still inferior to that of FPLinQ and even that of FPLinQ (no pc).

The above simulation setting is only concerned with sum rate, as the weights are all set to 1. We now consider a more demanding setting that takes priority weights into account. In this simulation, the weights are updated using the proportional fairness criterion, which is equivalent to maximizing the log-utility of the average link rates in the long run [43]. The network setting follows the previous simulation; the total number of links is fixed at 100. Fig. 5 compares the cumulative distribution of the link rates; the upper part of Table II compares the log-utility values. As we can see

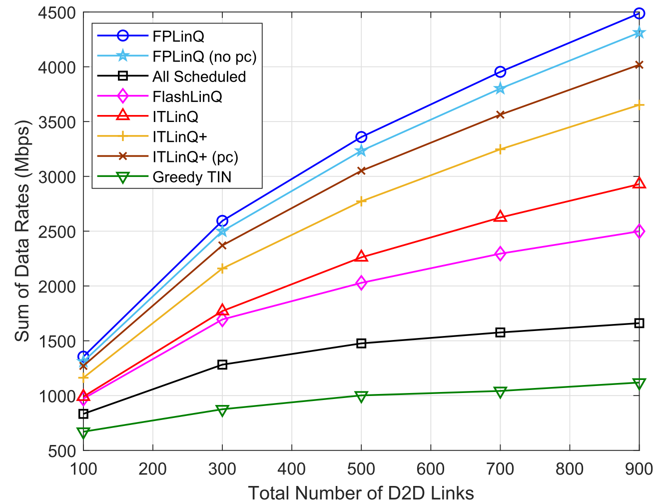


Fig. 4. Sum-rate maximization for the single-association D2D network.

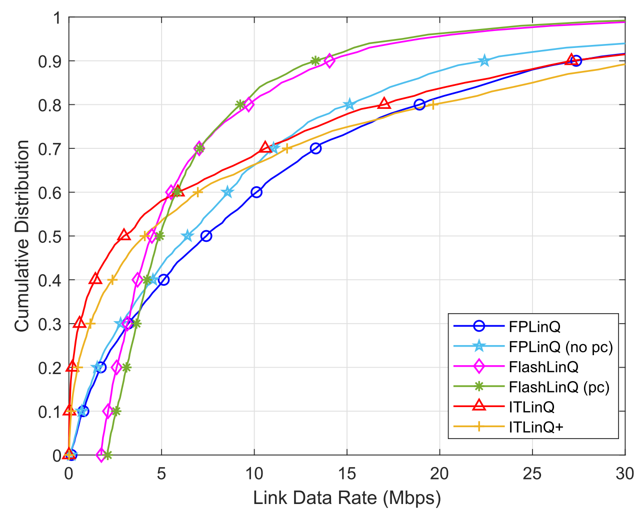


Fig. 5. Log-utility maximization for the single-association D2D network.

in Fig. 5, FPLinQ (no pc) strikes a better balance between the high-rate regime and the low-rate regime than ITLinQ and ITLinQ+. Surprisingly, FlashLinQ performs much better than ITLinQ and ITLinQ+ in this simulation; its performance is even slightly better than FPLinQ (no pc) according to Table II. In particular, observe in Fig. 5 that the low-rate links benefit the most from FlashLinQ, so FlashLinQ is fairly effective in protecting the low-rate links from strong interference, but its threshold value must be chosen carefully. Further, the benefit from the low-rate links comes at a cost for high-rate links. Overall, when we include power control and compare FPLinQ with a new benchmark method that combines FlashLinQ and power control in an alternative fashion, referred to as "FlashLinQ (pc)", FPLinQ outperforms FlashLinQ (pc) in network utility, when scheduling is optimized along with transmit powers, as shown in Table II.

Finally, we consider the flexible association case. We first generate 100 disjoint D2D links as before, but also generate two extra transmitters randomly for each receiver, and further let one third of the transmitters connect with one additional geographically closest receiver (excluding the already connected one). In this setup, we frequently encounter the situation that multiple transmitters contend for the same receiver, so the premature turning-off problem is very likely to occur.

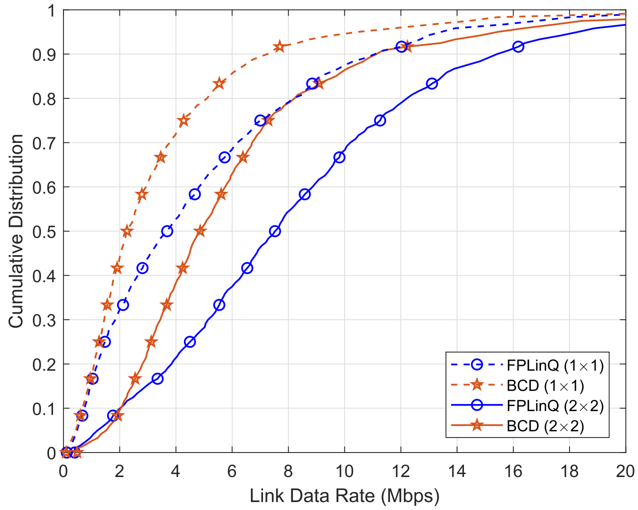


Fig. 6. Log-utility maximization for the flexible-association D2D network: FPLinQ vs. BCD.

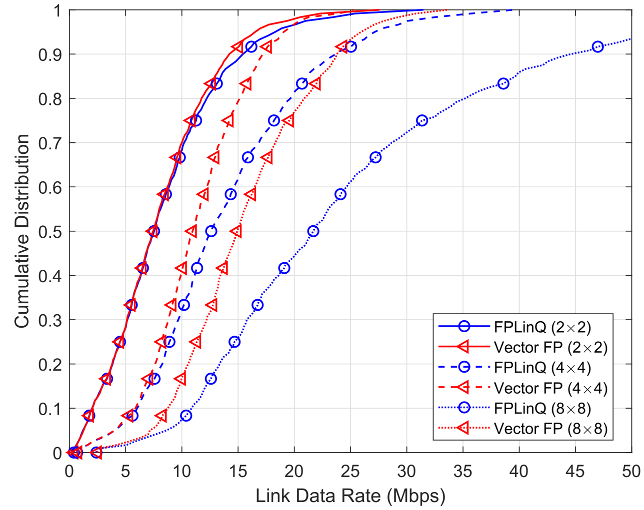


Fig. 7. Log-utility maximization for the flexible-association D2D network: FPLinQ vs. Vector FP.

TABLE II
SUM LOG-UTILITY OVER D2D NETWORKS

Fixed Single Association	Log Utility
FPLinQ	177.6
FPLinQ (no pc)	162.3
FlashLinQ	163.0
FlashLinQ (pc)	170.6
ITLinQ	57.0
ITLInQ+	109.5
Flexible Association	**Log Utility**
BCD (1 × 1)	99.6
BCD (2 × 2)	186.4
FPLinQ (1 × 1)	139.3
FPLinQ (2 × 2)	224.4
FPLinQ (4 × 4)	298.9
FPLinQ (8 × 8)	369.0
Vector FP (2 × 2)	223.3
Vector FP (4 × 4)	279.0
Vector FP (8 × 8)	321.5

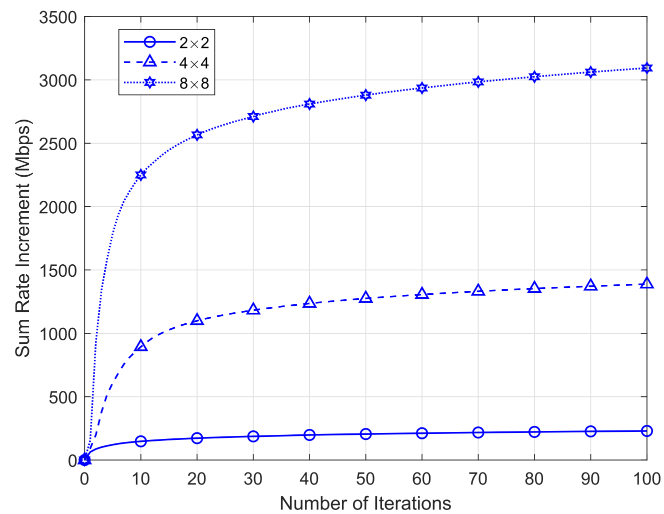


Fig. 8. Convergence of FPLinQ in maximizing the sum rate for the flexible-association D2D network.

We again optimize the log-utility by updating the link weights according to the proportional fairness criterion. FPLinQ is compared with the BCD method for both the single-antenna case and the 2×2 MIMO case (i.e., when each device terminal has 2 antennas). Note that FlashLinQ, ITLinQ, and ITLinQ+ are not applicable here, because they do not handle MIMO. Fig. 6 shows the cumulative distribution function of link rates, and the lower part of Table II summarizes the log-utility results. It can be seen that FPLinQ significantly outperforms BCD. In fact, as shown in Fig. 6, FPLinQ improves upon the BCD method by more than 50% for the 50th percentile link rate, in both the single-antenna case and the MIMO case. The corresponding log-utility of FPLinQ is also much higher. These results show that the premature turning-off can be fairly detrimental to the performance of D2D system in the flexible association case, thus making the proposed FPLinQ strategy a preferred strategy.

One of the key advantages of the proposed matrix FP strategy is its ability to accommodate multiple data streams in each MIMO link. In the next simulation, we evaluate the gain of multiple data-stream transmission over the single data-stream transmission. Toward this end, we compare FPLinQ (with matrix FP) against the vector FP method (also called multidimensional FP in [3]). The vector FP algorithm is the same as Algorithm 2 except that each transmit beamformer $\mathbf{V}_i \in \mathbb{C}^N$ is a complex vector instead of a matrix, so at most one data stream can be transmitted on each link. Fig. 7 shows the cumulative distribution function of link rates under different MIMO settings. It can be seen that while the gain of FPLinQ as compared to the vector FP is marginal in the 2×2 MIMO case, as more antennas are deployed at each terminal, the multiple data-stream transmission by FPLinQ starts to significantly outperform. The above observations is also evident from the lower part of Table II. Therefore, if the number of antenna N is small (e.g., 2), then using the vector FP in Algorithm 2 is more suited because of its lower complexity; on the other hand, if N is large (e.g., 8), then using FPLinQ with multiple data-stream transmission can boost the overall network throughput significantly.

Finally, Fig. 8 shows the convergence speed of FPLinQ when applied to maximizing the sum rate for the flexible-association D2D network with 400 links. Under the three MIMO settings (i.e., 2×2, 4×4, and 8×8), FPLinQ is observed to have fairly rapid convergence rate. Taking the 2×2 case for example, we observe from Fig. 8 that the majority of sum rate increment is obtained after the first 10 iterations. We also see that the convergence of FPLinQ is slower when more antennas are deployed at each terminal. But, as shown in the 8×8 case in Fig. 8, we can already reap most of the rate gain after about 40-60 iterations.

VII. Conclusion

This work proposes an interference-aware spectrum sharing strategy named FPLinQ to coordinate the scheduling decisions along with beamforming and power control across the wireless D2D links. The key step is to treat the weighted sum-rate maximization as a matrix FP problem and to use a sequence of matrix FP transforms to allow iterative optimization of scheduling and beamforming. We show that FPLinQ is closely related to the MM algorithm, thus its convergence is guaranteed. As compared to the existing methods, FPLinQ does not involve tuning of design parameters and does not suffer from the premature turning-off problem. The numerical results show that FPLinQ outperforms the state-of-the-art methods in terms of sum-rate maximization and log-utility maximization.

References

[1] K. Shen and W. Yu, "FPLinQ: A cooperative spectrum sharing strategy for device-to-device communications," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2323–2327.

[2] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[3] K. Shen and W. Yu, "Fractional programming for communication systems—Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, May 2018.

[4] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 74–85, Jan. 2004.

[5] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 302–315, Apr. 2006.

[6] Y. Xi and E. M. Yeh, "Distributed algorithms for spectrum allocation, power control, routing, and congestion control in wireless networks," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Sep. 2007, pp. 180–189.

[7] C. Joo, X. Lin, and N. B. Shroff, "Understanding the capacity region of the greedy maximal scheduling algorithm in multihop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 4, pp. 1132–1145, Aug. 2009.

[8] J. Huang, V. G. Subramanian, R. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 2, pp. 226–234, Feb. 2009.

[9] C. W. Tan, S. Friedland, and S. Low, "Spectrum management in multiuser cognitive wireless networks: Optimality and algorithm," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 421–430, Feb. 2011.

[10] C. W. Tan, M. Chiang, and R. Srikant, "Fast algorithms and performance bounds for sum rate maximization in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 706–719, Jun. 2013.

[11] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2027–2038, Oct. 2015.

[12] Z. Zhou and D. Guo, "1000-cell global spectrum management," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Jul. 2017, Art. no. 20.

[13] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[14] 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 Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[15] C. W. Tan, M. Chiang, and R. Srikant, "Maximizing sum rate and minimizing MSE on multiuser downlink: Optimality, fast algorithms and equivalence via max-min SINR," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6127–6143, Dec. 2011.

[16] J. Papandriopoulos and J. S. Evans, "SCALE: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3711–3724, Aug. 2009.

[17] A. Asadi and V. Mancuso, "On the compound impact of opportunistic scheduling and D2D communications in cellular networks," in *Proc. ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, Nov. 2013, pp. 279–288.

[18] Z. Liu, T. Peng, H. Chen, and W. Wang, "Optimal D2D user allocation over multi-bands under heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 1339–1344.

[19] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[20] P. Cheng, L. Deng, H. Yu, Y. Xu, and H. Wang, "Resource allocation for cognitive networks with D2D communication: An evolutionary approach," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 2671–2676.

[21] F. Malandrno, C. Casetti, C.-F. Chiasserini, and Z. Limani, "Fast resource scheduling in HetNets with D2D support," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr./May 2014, pp. 1536–1544.

[22] M. Sheng *et al.*, "On-demand scheduling: Achieving QoS differentiation for D2D communications," *IEEE Commun. Mag.*, vol. 53, no. 7, pp. 162–170, Jul. 2015.

[23] X. Lin, R. Ratasuk, and A. Ghosh, "Network-assisted device-to-device scheduling in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.

[24] X. Wu *et al.*, "FlashLinQ: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.

[25] N. Naderializadeh and A. S. Avestimehr, "ITLinQ: A new approach for spectrum sharing in device-to-device communication systems," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1139–1151, Jun. 2014.

[26] X. Yi and G. Caire, "Optimality of treating interference as noise: A combinatorial perspective," *IEEE Trans. Inf. Theory*, vol. 62, no. 8, pp. 4654–4673, Aug. 2016.

[27] W. Yu, T. Kwon, and C. Shin, "Multicell coordination via joint scheduling, beamforming, and power spectrum adaptation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 1–14, Jul. 2013.

[28] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Found. Trends Commun. Inf. Theory*, vol. 11, nos. 3–4, pp. 185–396, Jun. 2015.

[29] G. Sharma, R. R. Mazumdar, and N. B. Shroff, "On the complexity of scheduling in wireless networks," in *Proc. ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, Sep. 2006, pp. 227–238.

[30] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[31] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[32] C. Geng, N. Naderializadeh, S. Avestimehr, and S. A. Jafar, "On the optimality of treating interference as noise," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1753–1767, Apr. 2015.

[33] H. P. Benson, "On the global optimization of sums of linear fractional functions over a convex set," *J. Optim. Theory Appl.*, vol. 121, no. 1, pp. 19–39, 2004.

[34] H. P. Benson, "Solving sum of ratios fractional programs via concave minimization," *J. Optim. Theory Appl.*, vol. 135, no. 1, pp. 1–17, 2007.

[35] A. S. Lewis, "Derivatives of spectral functions," *Math. Oper. Res.*, vol. 21, no. 3, pp. 576–588, Aug. 1996.

[36] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[37] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[38] J. Song, P. Babu, and D. P. Palomar, "Optimization methods for designing sequences with low autocorrelation sidelobes," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3998–4009, Aug. 2015.

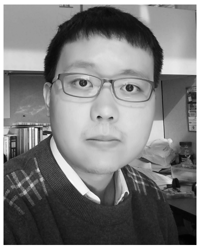
[39] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[40] D. P. Bertsekas, "The auction algorithm: A distributed relaxation method for the assignment problem," *Ann. Oper. Res.*, vol. 14, no. 1, pp. 105–123, Dec. 1988.

[41] T. T. Wu and K. Lange, "The MM alternative to EM," *Stat. Sci.*, vol. 25, no. 4, pp. 492–505, 2010.

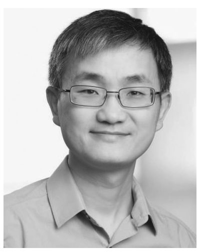
[42] Z. Zhou and D. Guo, “A centralized metropolitan-scale radio resource management scheme,” Aug. 2018, *arXiv:1808.02582*. Accessed: Aug. 9, 2018. [Online]. Available: https://arxiv.org/abs/1808.02582

[43] H. J. Kushner and P. A. Whiting, “Convergence of proportional-fair sharing algorithms under general conditions,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.



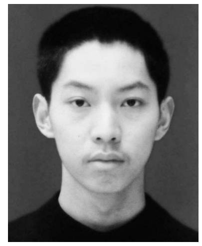
Kaiming Shen (S’13) received the B.Eng. degree in information security and the B.S. degree in mathematics from Shanghai Jiao Tong University, Shanghai, China, in 2011, and the M.A.Sc. degree in electrical and computer engineering from the University of Toronto, ON, Canada, in 2013, where he is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department. His current research interests include nonconvex optimization, data science, wireless communications, and information theory.

Mr. Shen received the Chinese Government Award for Outstanding Students Abroad in 2018 and the Shahid U. H. Qureshi Memorial Scholarship in 2019.



Wei Yu (S’97–M’02–SM’08–F’14) received the B.A.Sc. degree in computer engineering and mathematics from the University of Waterloo, Waterloo, ON, Canada, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1998 and 2002, respectively. Since 2002, he has been with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, Canada, where he is currently a Professor and holds a Canada Research Chair (Tier 1) in information theory and wireless communications. His main research interests include information theory, optimization, wireless communications, and broadband access networks.

Dr. Yu is a Fellow of the Canadian Academy of Engineering and a member of the College of New Scholars, Artists and Scientists of the Royal Society of Canada. He was a recipient of the Steacie Memorial Fellowship in 2015, the IEEE Marconi Prize Paper Award in Wireless Communications in 2019, the IEEE Communications Society Award for Advances in Communication in 2019, the IEEE Signal Processing Society Best Paper Award in 2017 and 2008, the *Journal of Communications and Networks* Best Paper Award in 2017, the IEEE Communications Society Best Tutorial Paper Award in 2015, the IEEE International Conference on Communications Best Paper Award in 2013, the McCharles Prize for Early Career Research Distinction in 2008, the Early Career Teaching Award from the Faculty of Applied Science and Engineering, University of Toronto, in 2007, and the Early Researcher Award from Ontario in 2006. He currently serves as the Vice President for the IEEE Information Theory Society. He is an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2010 to 2013, as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2009 to 2011, and as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2004 to 2007. He served as the Chair of the Signal Processing for Communications and Networking Technical Committee, IEEE Signal Processing Society, from 2017 to 2018. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2016. He is recognized as a Highly Cited Researcher.



Licheng Zhao received the B.S. degree in information engineering from Southeast University (SEU), Nanjing, China, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST), in 2014 and 2018, respectively. His research interests are in optimization theory and fast algorithms, with applications in signal processing, machine learning, and financial engineering.

Daniel P. Palomar (S’99–M’03–SM’08–F’12) received the Electrical Engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He was a Fulbright Scholar at Princeton University, Princeton, NJ, USA, from 2004 to 2006. He is currently a Professor with the Department of Electronics and Computer Engineering and also with the Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology (HKUST), Hong Kong, which he joined in 2006. He had previously held several research appointments, namely, at King’s College London (KCL), London, U.K.; Stanford University, Stanford, CA, USA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome “La Sapienza”, Rome, Italy; and Princeton University. His current research interests include applications of optimization theory and signal processing in financial systems and big data analytics.

Dr. Palomar is a Fellow of the IEEE, a recipient of the 2004/06 Fulbright Research Fellowship, the 2004 and 2015 (coauthor) Young Author Best Paper Awards by the IEEE Signal Processing Society, the 2015-16 HKUST Excellence Research Award, the 2002/03 Best Ph.D. Prize in Information Technologies and Communications by UPC, the 2002/03 Rosina Ribalta first prize for the Best Doctoral Thesis in Information Technologies and Communications by the Epson Foundation, and the 2004 prize for the Best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT. He has been a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING 2016 Special Issue on “Financial Signal Processing and Machine Learning for Electronic Trading,” an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the *IEEE Signal Processing Magazine* 2010 Special Issue on “Convex Optimization for Signal Processing,” the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on “Game Theory in Communication Systems,” and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on “Optimization of MIMO Transceivers for Realistic Communication Networks.”