An Optimization-Based User Scheduling Framework for mmWave Massive MU-MIMO Systems

Victoria Palhares and Christoph Studer

Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland
e-mail: palhares@iis.ee.ethz.ch and studer@ethz.ch

***Abstract*—We propose a novel user equipment (UE) scheduling framework for millimeter-wave (mmWave) massive multiuser (MU) multiple-input multiple-output (MIMO) wireless systems. Our framework determines (sub)sets of UEs that should transmit simultaneously in a given time slot by approximately solving a nonconvex optimization problem using forward-backward splitting. Our UE scheduling framework is flexible in the sense that it (i) supports a variety of cost functions, including post-equalization mean square error and sum rate, and (ii) enables precise control over the minimum and maximum number of resources the UEs should occupy. We demonstrate the efficacy of our framework using realistic mmWave channel vectors generated with a commercial ray-tracer. We show that our UE scheduler outperforms a range of existing scheduling methods and closely approaches the performance of an exhaustive search.**

I. Introduction

Millimeter-wave (mmWave) communication combined with massive multiuser (MU) multiple-input multiple-output (MIMO) is expected to be a core component in sixth generation (6G) wireless systems [1]. Combining these two technologies not only promises large beamforming gains to combat the high path loss at mmWave frequencies but also enables high-bandwidth data transmission to multiple user equipments (UEs) in the same time-frequency resource.

In order to maximize the quality-of-service (QoS) for all UEs in the network, resource allocation strategies, such as power control and UE scheduling, are necessary. Power control mitigates the near-far problem between the UEs transmitting to an infrastructure basestation (BS). UE scheduling distributes the UEs' requests either in time/frequency or spatially, with the goal of minimizing interference in scenarios where UEs have similar channel impulse responses—this can happen in congested scenarios with many transmitting UEs in close vicinity. The literature focuses almost exclusively on greedy UE scheduling algorithms, where one UE is selected at a time to join a set of scheduled UEs [2], [3]. Such approaches have difficulties scheduling UEs over multiple time slots and fail to consider the scheduling problem globally, which results in error propagation and inevitably leads to sub-optimal QoS.

The work of VP and CS was supported in part by an ETH Research Grant. The work of CS was supported in part by ComSenTer, one of six centers in JUMP, a SRC program sponsored by DARPA. The work of CS was supported in part by the U.S. NSF under grants CNS-1717559 and ECCS-1824379.

We thank Haochuan Song and Seyed Hadi Mirfarshbafan for discussions on FBS and mmWave channels, respectively. We thank Gian Marti and Sueda Taner for comments on an early version of this paper. We also thank Remcom for providing us with a license for the Wireless InSite ray-tracing software.

A. Contributions

We propose a novel UE scheduling framework for mmWave MU-MIMO systems that identifies subsets of UEs that should transmit simultaneously in a given time slot. We formulate a global optimization problem, which supports a variety of cost functions, including post-linear minimum mean-square error (LMMSE) equalization mean square error (MSE) and post-LMMSE equalization sum rate. Our framework also enables precise control over the minimum and maximum number of resources each UE should occupy. In order to efficiently find approximate solutions to the nonconvex UE scheduling problem, we approximately solve a relaxed version using forward-backward splitting (FBS). We use mmWave channel vectors from a commercial ray-tracer to demonstrate that our framework outperforms a range of existing algorithms in terms of bit error rate (BER) and average per-UE rates, often closely approaching the performance of an exhaustive search (ES).

B. Relevant Prior Art

UE scheduling has been studied extensively for mmWave communication [3]–[6] and massive MU-MIMO systems [2], [7]–[10]. Many of the proposed algorithms follow a greedy approach, such as semiorthogonal UE selection (SUS) [2], channel structure-based scheduling (CSS) [3], greedy max-sum rate scheduling (greedy) [3], and chordal distance-based UE scheduling (chordal) [3]. All of these methods iteratively select UEs to greedily maximize a predefined cost function. In stark contrast, we address the UE scheduling problem globally, i.e., we determine the transmitting UEs for all time slots jointly by solving a single optimization problem. Our framework is flexible in (i) the choice of the cost function and (ii) the constraints on the maximum/minimum number of UEs that can transmit per time slot as well as the maximum/minimum number of time slots that a UE is allowed to transmit. Furthermore, most methods proposed in the literature aim at optimizing the signal-to-interference-plus-noise ratio (SINR), chordal distance, or channel capacity [10]. Our framework also allows other, practically relevant cost functions, such as the post-equalization MSE or sum rate. Finally, we note that the majority of results for mmWave systems focus on hybrid

analog-digital beamforming architectures [3]–[6]. In contrast, our framework considers all-digital BS architectures, which are recently gaining popularity [11], [12].

C. Notation

Upper case and lower case bold symbols denote matrices and vectors, respectively. We use $A_{i,j}$ for the element on the ith row and jth column of $\mathbf{A}$, $\mathbf{a}_j$ as the jth column of $\mathbf{A}$, and a_i as the ith element of vector $\mathbf{a}$. We define $\mathbf{I}_M$, $\mathbf{1}_{L\times M}$, and $\mathbf{0}_{L\times M}$ as the $M\times M$ identity matrix, $L\times M$ all-one matrix, and $L\times M$ all-zero matrix, respectively. The superscript $(\cdot)^H$ denotes the Hermitian transpose. A diagonal matrix with $\mathbf{a}$ on the main diagonal is denoted by $\mathrm{diag}(\mathbf{a})$. The Euclidean and Frobenius norms are denoted by $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$, respectively. Expectation is $\mathbb{E}[\cdot]$ and $\overset{e}{\leq}$ represents element-wise less-than-or-equal-to.

II. Prerequisites

A. System Model

We consider the uplink of a mmWave massive MU-MIMO system in which U single-antenna UEs transmit data to an all-digital BS equipped with a B-antenna uniform linear array (ULA). We consider a block-fading scenario with $t=1,\ldots,T$ time slots and frequency-flat channels with input-output relation

$$\mathbf{y}[t]=\mathbf{H}\mathbf{s}[t]+\mathbf{n}[t]\,. \tag{1}$$

Here, $\mathbf{y}[t]\in\mathbb{C}^B$ is the receive vector at time slot t, $\mathbf{s}[t]\in\mathbb{C}^U$ contains the transmit signals from all U UEs, and $\mathbf{n}[t]\in\mathbb{C}^B$ models noise whose entries are i.i.d. circularly-symmetric complex Gaussian with variance N_0. The effective channel matrix $\mathbf{H}=\tilde{\mathbf{H}}\mathbf{\Delta}$ in (1) combines the effect of the MIMO channel matrix $\tilde{\mathbf{H}}\in\mathbb{C}^{B\times U}$ and the power control matrix $\mathbf{\Delta}=\mathrm{diag}(\delta_1,\ldots,\delta_U)$, whose entries are given by [13]

$$\delta_u^2=\min\Big\{\|\tilde{\mathbf{h}}_u\|_2^2,10^{\frac{\eta}{10}}\min\nolimits_{u'=1,\ldots,U}\|\tilde{\mathbf{h}}_{u'}\|_2^2\Big\}/\|\tilde{\mathbf{h}}_u\|_2^2. \tag{2}$$

In (2), $\tilde{\mathbf{h}}_u$ stands for the uth column of $\tilde{\mathbf{H}}$ and $\eta\geq 0$ determines the maximum dynamic range between the weakest and strongest UE receive power in decibel (dB).

B. UE Scheduling

To formalize the UE scheduling problem, we define a binary-valued UE scheduling matrix $\mathbf{C}\in\{0,1\}^{U\times T}$, where $C_{u,t}=1$ and $C_{u,t}=0$ indicate that UE u during time slot t is active and inactive, respectively. Furthermore, we define a diagonal mask matrix $\mathbf{D}_{\mathbf{C}}[t]=\mathrm{diag}(\mathbf{c}_t)\in\{0,1\}^{U\times U}$, where $\mathbf{c}_t$ is the tth column of $\mathbf{C}$. Through multiplication with $\mathbf{s}[t]$, this mask matrix describes which UEs transmit symbols and which UEs are idle in time slot t. We absorb the effect of the mask matrix in the effective channel as $\mathbf{H}[t]=\mathbf{H}\mathbf{D}_{\mathbf{C}}[t]$, which now depends on t. Our goal is to develop a framework that determines a UE scheduling matrix $\mathbf{C}$ by minimizing a given cost function $F(\mathbf{C})$ subject to constraints that specify the minimum and maximum number of resources that UEs are allowed to occupy.

III. UE Scheduling Framework

A. Scheduling as an Optimization Problem

The proposed UE scheduling framework consists of an application-specific cost function $F(\mathbf{C})$ and a set of constraints that determine the resource utilization in time and space. We wish to solve the following optimization problem:

$$\underset{\mathbf{C}\in\{0,1\}^{U\times T}}{\text{minimize}}\ F(\mathbf{C})\ \text{ subject to }\ \mathbf{C}\in\mathcal{C}_U\cap\mathcal{C}_T, \tag{3}$$

with the two constraint sets

$$\mathcal{C}_U=\big\{U_{\min}\mathbf{1}_{1\times T}\overset{e}{\leq}\mathbf{1}_{1\times U}\mathbf{C}\overset{e}{\leq}U_{\max}\mathbf{1}_{1\times T}\big\} \tag{4}$$

$$\mathcal{C}_T=\big\{T_{\min}\mathbf{1}_{U\times 1}\overset{e}{\leq}\mathbf{C}\mathbf{1}_{T\times 1}\overset{e}{\leq}T_{\max}\mathbf{1}_{U\times 1}\big\}. \tag{5}$$

The set $\mathcal{C}_U$ in (4) determines the minimum $U_{\min}$ and maximum $U_{\max}$ number of UEs allowed to transmit simultaneously per time slot; the set $\mathcal{C}_T$ in (5) determines the minimum $T_{\min}$ and maximum $T_{\max}$ number of time slots each UE is allowed to transmit. Due to the discrete nature of the UE scheduling matrix $\mathbf{C}$, the problem in (3) is of combinatorial nature. For example, in a scenario with a total number of $U=32$ UEs with 16 UEs transmitting in the first time slot and the remaining 16 UEs in the second time slot, an ES would need to test over 600 million scheduling matrices. Evidently, approximate methods to solve the scheduling problem in (3) are necessary.

B. Problem Relaxation

To arrive at an efficient UE scheduling algorithm, we relax the binary-valued $\{0,1\}$ entries in $\mathbf{C}$ to the continuous range $[0,1]$ as follows: $\mathbf{C}\in[0,1]^{U\times T}$. While this relaxation enables the use of computationally efficient gradient-descent-based methods, it no longer enforces that the solutions are in $\{0,1\}$. To mitigate this issue, we augment the cost function of the relaxed optimization problem with the following regularizer, which promotes binary-valued scheduling matrices $\mathbf{C}$ [14]:

$$R(\mathbf{C})=-\sum_{t=1}^{T}\sum_{u=1}^{U}\alpha|C_{u,t}-0.5|^2\,. \tag{6}$$

Here, $\alpha\geq 0$ is a regularization parameter, where larger values enforce binary-valued solutions more strictly. By utilizing $\tilde{F}(\mathbf{C})=F(\mathbf{C})+R(\mathbf{C})$ as a new cost function, we are now able to deploy numerical optimization methods to efficiently determine binary-valued solutions to the relaxed problem in (3). Since the augmented cost function is nonconvex (even if the original cost function $F(\mathbf{C})$ is convex), numerical methods may only converge to a local minimum. Thus, to improve the solution quality, we perform multiple random initializations and use the solution candidate $\mathbf{C}^*$ with the lowest cost $F(\mathbf{C}^*)$.

C. UE Scheduling via Forward-Backward Splitting (FBS)

To solve (3) for the set $\mathbf{C}\in[0,1]^{U\times T}$ and the augmented cost function $\tilde{F}(\mathbf{C})=F(\mathbf{C})+R(\mathbf{C})$, we can use FBS [15]. This technique performs the following step for iterations $i=1,2,\ldots$ until a stopping-condition is met:

$$\mathbf{C}^{(i+1)}=\mathrm{prox}_{\mathcal{C}_U\cap\mathcal{C}_T}\Big(\mathbf{C}^{(i)}-\tau^{(i)}\nabla\tilde{F}\Big(\mathbf{C}^{(i)}\Big)\Big). \tag{7}$$

Here, the proximal operator $\mathrm{prox}_{\mathcal{C}_U \cap \mathcal{C}_T}(\cdot)$ is the orthogonal projection onto $\mathcal{C}_U \cap \mathcal{C}_T$, $\tau^{(i)}$ are suitably-chosen step sizes, and $\nabla \tilde{F}(\cdot)$ is the gradient of $\tilde{F}(\cdot)$. FBS is initialized with a matrix $\mathbf{C}^{(1)} \in [0,1]^{U \times T}$ drawn uniformly at random and the entries of the matrix in the last iteration $\mathbf{C}^{(I_{\max})}$ are quantized to $\{0,1\}$ to satisfy the constraint sets $\mathcal{C}_U$ and $\mathcal{C}_T$.

IV. Proximal Operator

A. Douglas-Rachford Splitting (DRS)

We now outline the implementation of the proximal operator $\mathrm{prox}_{\mathcal{C}_U \cap \mathcal{C}_T}(\cdot)$ in (7). This operator is the orthogonal projection onto an intersection of two simplexes. To determine a solution that lies within the two sets, we utilize Douglas-Rachford splitting (DRS) [16]. DRS alternatively projects onto the sets $\mathcal{C}_U$ and $\mathcal{C}_T$ until a solution matrix is found. Concretely, we solve the following optimization problem

$$\mathrm{prox}_{\mathcal{C}_U \cap \mathcal{C}_T}(\mathbf{Z}) = \arg\min_{\mathbf{X}} \ \psi(\mathbf{X}, \mathbf{Z}) + \xi(\mathbf{X}, \mathbf{Z}), \tag{8}$$

with the two functions

$$\psi(\mathbf{X}, \mathbf{Z}) = \frac{\beta}{2}\|\mathbf{X} - \mathbf{Z}\|_{\mathrm{F}}^2 + \mathcal{X}_{\mathcal{C}_U}(\mathbf{X}) \tag{9}$$

$$\xi(\mathbf{X}, \mathbf{Z}) = \frac{\beta}{2}\|\mathbf{X} - \mathbf{Z}\|_{\mathrm{F}}^2 + \mathcal{X}_{\mathcal{C}_T}(\mathbf{X}). \tag{10}$$

Here, $\beta \geq 0$ is a tuning parameter that can be used to accelerate convergence and the constraints are enforced using indicator functions $\mathcal{X}_{\mathcal{C}_U}$ and $\mathcal{X}_{\mathcal{C}_T}$, which we define as follows:

$$\mathcal{X}_{\mathcal{C}}(\mathbf{X}) = \begin{cases} 0 & \mathbf{X} \in \mathcal{C} \\ \infty & \text{otherwise.} \end{cases} \tag{11}$$

DRS iteratively carries out the two steps [16]

$$\mathbf{V}^{(k+1)} = \mathrm{prox}_{\psi}\left(\mathbf{G}^{(k)}\right) = \arg\min_{\mathbf{X} \in \mathcal{C}_U} \left\| \mathbf{X} - \frac{\beta \mathbf{Z} + \mathbf{G}^{(k)}}{\beta + 1} \right\|_{\mathrm{F}}^2 \tag{12}$$

$$\mathbf{G}^{(k+1)} = \mathrm{prox}_{\xi}\left(2\mathbf{V}^{(k+1)} - \mathbf{G}^{(k)}\right) + \mathbf{G}^{(k)} - \mathbf{V}^{(k+1)}, \tag{13}$$

and is initialized with $\mathbf{G}^{(1)} = \mathbf{0}_{U \times T}$. After convergence or a maximum number of iterations, DRS outputs the matrix $\mathbf{V}^{(K_{\max})}$. The details of this algorithm will be provided in [17].

B. Projection with Inequality Constraints

The steps in (12) and (13) project each column onto the simplexes. Since both of the sets $\mathcal{C}_U$ and $\mathcal{C}_T$ are described by inequalities, we need a projection onto a generic simplex with inequality constraints. Given a vector $\mathbf{q} \in \mathbb{R}^M$, our objective is to find a solution vector $\mathbf{p}^* \in \mathcal{C}^M$ closest to $\mathbf{q}$ as follows:

$$\underset{\mathbf{p} \in \mathbb{R}^M}{\text{minimize}} \ \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2^2 \tag{14a}$$

$$\text{subject to } l_{\min} \leq \sum_{i=1}^{M} p_i \leq l_{\max}, \ 0 \leq p_i \leq 1, \forall i. \tag{14b}$$

To solve this subproblem efficiently, we use a line search approach that requires sorting of $\mathbf{q}$ to determine the vector $\mathbf{p}^*$ satisfying the inequality constraints. The Karush-Kuhn-Tucker (KKT) conditions and the algorithm will be detailed in [17].

V. Cost Functions

We now discuss the two cost functions considered in this work. Due to space constraints, the gradients as well as other cost functions will be discussed in future work [17].

A. Post-LMMSE Equalization MSE

As our first cost function, we consider the post-LMMSE MSE. We define this cost function as

$$F(\mathbf{C}) = \sum_{t=1}^{T} \mathbb{E}\left[\left\| \mathbf{D}_{\mathbf{C}}[t]\, \mathbf{s}[t] - \mathbf{D}_{\mathbf{C}}[t]\, \mathbf{W}[t]^H\, \mathbf{y}[t] \right\|_2^2 \right], \tag{15}$$

where $\mathbf{W}[t] = \mathbf{H}[t]\left(\mathbf{H}[t]^H\, \mathbf{H}[t] + \frac{N_0}{E_s}\mathbf{I}_U\right)^{-1}$ is the LMMSE equalization matrix in time slot t and E_s is the transmit signal energy. Note that this cost function only accounts for errors from the UEs that transmit in a given time slot t. The gradient for this cost function will be detailed in [17].

B. Post-LMMSE Equalization Sum Rate

As our second cost function, we consider the sum of achievable rates over all UEs and time slots after an LMMSE equalizer. We define this cost function as

$$F(\mathbf{C}) = -\sum_{t=1}^{T}\sum_{u=1}^{U} \log_2(1 + \mathit{SINR}_u(\mathbf{c}_t)), \tag{16}$$

where the SINR of the uth UE in time slot t is given by

$$\mathit{SINR}_u(\mathbf{c}_t) = \frac{\left|\mathbf{w}_u[t]^H \mathbf{h}_u[t]\right|^2}{\sum_{u'=1, u' \neq u}^{U} \left|\mathbf{w}_u[t]^H \mathbf{h}_{u'}[t]\right|^2 + \frac{N_0}{E_s}\|\mathbf{w}_u[t]\|_2^2}. \tag{17}$$

Here, $\mathbf{w}_u[t]$ is the uth column of $\mathbf{W}[t]$, and $\mathbf{h}_u[t]$ is the uth column of the effective, masked channel matrix $\mathbf{H}[t] = \mathbf{H}\mathbf{D}_{\mathbf{C}}[t]$. The gradient for this cost function will be detailed in [17].

VI. Simulation Results

A. Simulation Setup

We consider a mmWave massive MU-MIMO system with a carrier frequency of 60 GHz and a bandwidth of 100 MHz. At the BS, we consider an ULA with $\lambda/2$ antenna spacing. The BS and UE antennas are omnidirectional and are at a height of 10 m and 1.65 m, respectively. The UEs transmit 16-QAM symbols and the per-UE power control dynamic range is set to $\eta = 6$ dB. We evaluate our framework using mmWave channel vectors generated with Wireless InSite [18]. The channel vectors are generated for $22,448$ UE positions in an area of $109.7\,\mathrm{m} \times 164.7\,\mathrm{m}$. The scenario is depicted in Fig. 1; the BS location is shown with a blue circle. To generate a channel realization, U UE positions are selected uniformly and independently at random. For channel estimation, we use BEACHES [11]. In Table I, we list the four different scenarios considered in what follows and their respective number of random initializations. We set $T_{\min} = T_{\max} = T_S$ and $U_{\min} = U_{\max} = U_S$. For the gradient step, we set $\tau^{(i)} = \tau$ to a constant stepsize.

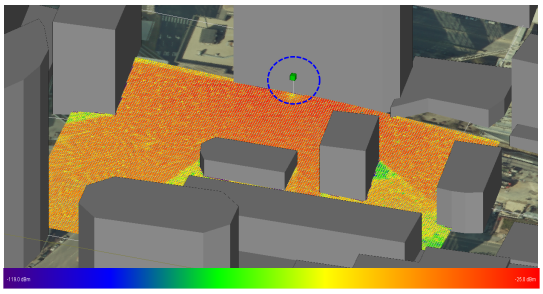


Fig. 1. Simulated scenario with 22,448 UE positions and a 32 antenna BS. Colors indicate receive power from −119 dBm (blue) to −25 dBm (red).

TABLE I
EVALUATED SCENARIOS

Scenario	B	U	T	T_S	U_S	Initializations
S1	16	16	2	1	8	80
S2	32	32	2	1	16	10
S3	32	64	2	1	32	3
S4	32	64	4	1	16	3

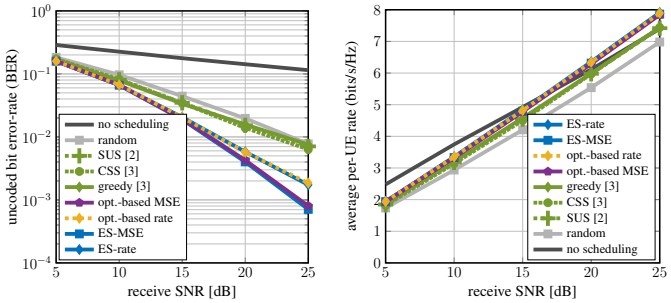
B. Performance Metrics and Baseline Algorithms

We consider the uncoded BER and average per-UE rate, both evaluated after utilizing an LMMSE equalizer. In our simulation results, we refer to the post-LMMSE equalization MSE as "opt.-based MSE" and to the post-LMMSE equalization sum rate as "opt.-based rate." In order to compare our framework with baseline methods, we also simulate the "SUS" [2], "CSS" [3], and "greedy" [3] algorithms. For all of the mentioned baseline methods, we schedule the UEs for the first time slot and then schedule the remaining unscheduled UEs in the subsequent time slots until all of the UE requests are spread over T time slots. We also consider a baseline, which picks the subsets of UEs uniformly at random (called "random") and a "no scheduling" baseline, in which all of the UEs are scheduled in all the time slots. For Scenario S1, we also consider an ES, which tests every possible scheduling matrix $\mathbf{C}$ and selects the one that minimizes the given cost function.

C. Simulation Results

Fig. 2 shows the simulation results for Scenario S1. In Fig. 2(a), we see that our UE scheduling framework, both with the "opt.-based MSE" and "opt.-based rate" cost functions are comparable to the optimal baselines "ES-MSE" and "ES-rate," reaching a BER of less than 0.1% at a signal-to-noise ratio (SNR) of 25 dB. In Fig. 2(b), we observe the same behavior: the performance of our UE scheduling framework is nearly indistinguishable from that of an ES. Compared to existing baseline algorithms, our "opt.-based MSE" and "opt.-based rate" methods achieve superior performance. For example, compared to "CSS," "opt.-based rate" realizes a gain of over 4 dB at 1% BER. Compared with the "no scheduling," "opt.-based MSE" and "opt.-based rate" are superior in terms of BER for all SNR values and in terms of average per-UE rate at high SNR.

Fig. 3(a), Fig. 3(b), and Fig. 3(c) show simulation results for Scenarios S2, S3, and S4 in terms of BER while Fig. 3(d), Fig. 3(e), and Fig. 3(f) present the same scenarios in terms of average per-UE rate. In Fig. 3(a), we see that our "opt.-based MSE" and "opt.-based rate" UE scheduling algorithms have a gain of at least 3 dB when compared to the best baseline algorithm at 1% BER. In terms of average per-UE rate, we see in Fig. 3(d) that the performance of our framework outperforms existing algorithms, albeit only slightly. In Fig. 3(b) and Fig. 3(e), we consider a system in which the number of scheduled UEs is equal to the number of BS antennas. Even with the use of UE scheduling, we observe that we cannot provide reasonable QoS, especially in terms of uncoded BER where it can provide a BER of approximately 10% at an SNR of 25 dB with the best scheduling method. This behavior is mainly due to the use of an LMMSE equalizer. To improve performance, one could spread the UE requests over more time slots, so that $U_S < B$, which is what we have done in Scenario S4. We see in Fig. 3(c) and Fig. 3(f), that the performance of our "opt.-based MSE" and "opt.-based rate" algorithms is superior to the baseline methods, with a gain of at least 3 dB at 1% BER, for example.



(a) BER (b) Average per-UE rate

Fig. 2. (a) BER and (b) average per-UE rate for Scenario S1 with 10^2 channel realizations and 10^5 transmissions. The proposed UE scheduling algorithm is comparable to an ES for the MSE and sum rate criteria, in terms of BER and average per-UE rate. Our methods can reach a BER of less than 0.1% at a SNR of 25 dB. Compared to baseline methods from the literature, our framework provides a gain over 4 dB at 1% BER, for example.

Not surprisingly, we see that utilizing the "opt.-based MSE" cost function provides better BER performance than "opt.-based rate." Furthermore, we observe that in all scenarios, the "no scheduling" baseline yields extremely poor BER performance, which is due to the suboptimal LMMSE equalizer. When it comes to the average per-UE rate, as the UE to BS antenna ratio $\frac{U}{B}$ or the SNR increases, the performance of "no scheduling" decreases. No scheduling can be beneficial at low SNR in terms of average per-UE rate, but realizing these gains would require sophisticated forward error correction strategies.

VII. CONCLUSIONS

We have proposed a novel optimization-based UE scheduling framework for mmWave massive MU-MIMO systems. The framework supports a range of cost functions and constraints that specify the UE resource allocation. For a small system with $B = 16$ and $U = 16$, we have shown that the performance

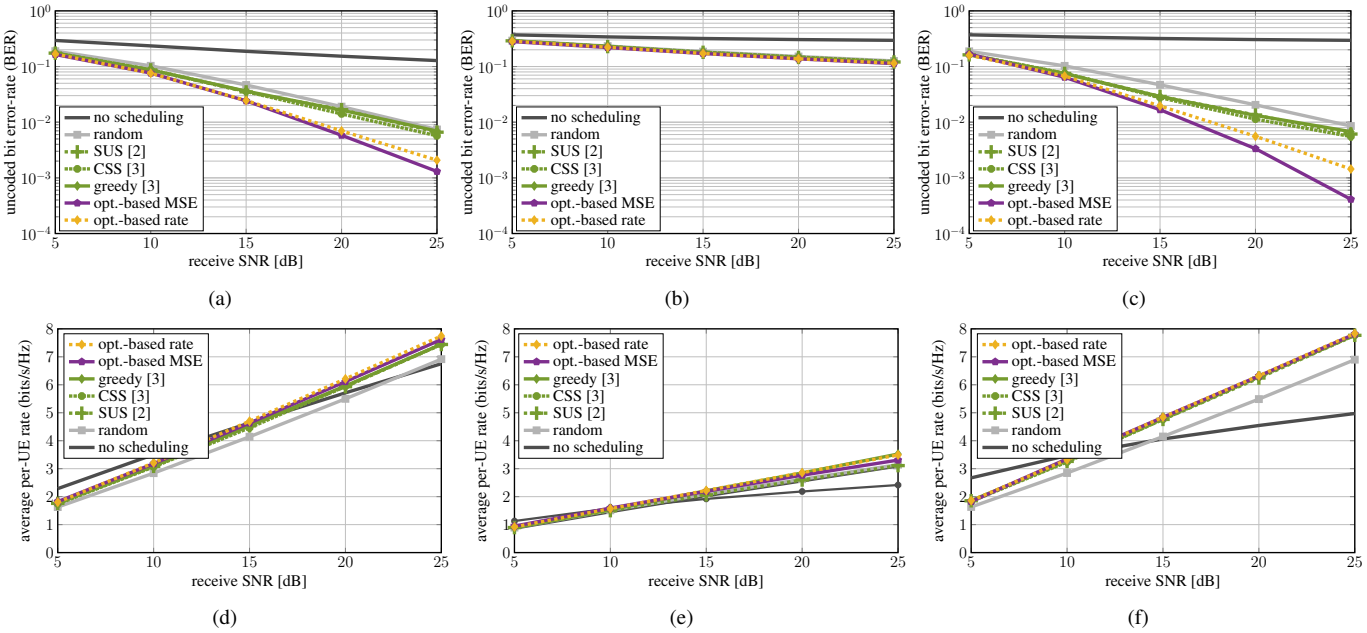


Fig. 3. (a), (b) and (c) are the BER performances of Scenarios S2, S3, and S4, respectively while (d), (e), and (f) are the average per-UE rates of Scenarios S2, S3, and S4, respectively. These simulations consider 10^2 channel realizations and 10^5 transmissions. In (a) and (c), where $U_S < B$, "opt.-based MSE" and "opt.-based rate" outperform baseline methods by 3 dB at 1% BER. In (d) and (f), our proposed algorithms outperform existing methods and the "no scheduling" baseline at high SNR. In (b) and (e), where $U_S = B$, even the use of scheduling does not provide acceptable QoS.

of the proposed methods is comparable to that of an exhaustive search, whereas existing baseline algorithms perform (often significantly) worse. In larger systems with $U_S < B$, we have shown that the proposed methods outperform existing methods in terms of BER and average per-UE rate. In scenarios where $U_S \geq B$, even our framework together with an LMMSE equalizer cannot achieve acceptable performance. Thus, one can schedule the UE requests over more time slots so that $U_S < B$, as it has been shown in Fig. 3(c) and Fig. 3(f).

In [17], we will introduce a range of other cost functions as well as their respective gradients. In addition, we will present the complete derivation of the projection on the simplex with inequality constraints and include a complexity analysis.

REFERENCES

[1] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: the next wireless revolution?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.

[2] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[3] J. Choi, G. Lee, and B. L. Evans, "User scheduling for millimeter wave hybrid beamforming systems with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2401–2414, Mar. 2019.

[4] H. Wu, D. Liu, W. Wu, C. Na, and M. Liu, "A low complexity two-stage user scheduling scheme for mmWave massive MIMO hybrid beamforming systems," in *IEEE Int. Conf. Comput. Commun. (ICCC)*, Dec. 2017, pp. 945–951.

[5] J. Zhu, Q. Li, Z. Liu, H. Chen, and H. Vincent Poor, "Enhanced user grouping and power allocation for hybrid mmWave MIMO-NOMA systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, Sep. 2021.

[6] X. Gao, X. Wu, Z. Zhang, and D. Liu, "Low complexity joint user scheduling and hybrid beamforming for mmWave massive MIMO systems," in *IEEE Int. Symp. Personal, Indoor, Mobile Radio Commun. (PIMRC)*, Aug. 2020, pp. 1–6.

[7] G. Lee and Y. Sung, "A new approach to user scheduling in massive multi-user MIMO broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1481–1495, Dec. 2017.

[8] A. Farsaei, A. Alvarado, F. M. J. Willems, and U. Gustavsson, "An improved dropping algorithm for line-of-sight massive MIMO with max-min power control," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 1109–1112, Apr. 2019.

[9] Z. Jiang, S. Chen, S. Zhou, and Z. Niu, "Joint user scheduling and beam selection optimization for beam-based massive MIMO downlinks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2190–2204, Jan. 2018.

[10] B. Hu, C. Hua, C. Chen, X. Ma, and X. Guan, "MUBFP: Multiuser beamforming and partitioning for sum capacity maximization in MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 233–245, Mar. 2016.

[11] S. H. Mirfarshbafan, A. Gallyas-Sanhueza, R. Ghods, and C. Studer, "Beamspace channel estimation for massive MIMO mmWave systems: Algorithm and VLSI design," *IEEE Trans. Circuits Syst. I*, vol. 67, no. 12, pp. 5482–5495, Sep. 2020.

[12] S. Dutta, C. N. Barati, D. Ramirez, A. Dhananjay, J. F. Buckwalter, and S. Rangan, "A case for digital beamforming at mmWave," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 756–770, Feb. 2020.

[13] H. Song, T. Goldstein, X. You, C. Zhang, O. Tirkkonen, and C. Studer, "Joint channel estimation and data detection in cell-free massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, Nov. 2021.

[14] O. Castañeda, T. Goldstein, and C. Studer, "VLSI designs for joint channel estimation and data detection in large SIMO wireless systems," *IEEE Trans. Circuits Syst. I*, vol. 65, no. 3, pp. 1120–1132, Oct. 2018.

[15] T. Goldstein, C. Studer, and R. G. Baraniuk, "A field guide to forward-backward splitting with a FASTA implementation," *arXiv*, vol. abs/1411.3406, 2014. [Online]. Available: http://arxiv.org/abs/1411.3406

[16] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Trans. of the American Mathematical Society*, vol. 82, no. 2, pp. 421–439, Jul. 1956.

[17] V. Palhares and C. Studer, "An optimization-based user scheduling framework for multiuser systems," *work in progress*, 2022.

[18] (2021). [Online]. Available: https://www.remcom.com/wireless-insite-em-propagation-software/