KTH Electrical Engineering

Iterative Precoder Design and User Scheduling for Block-Diagonalized Systems

In IEEE Transactions on Signal Processing, vol. 60, no. 7, pp. 3726–3739, July 2012

© 2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

DOI: 10.1109/TSP.2012.2192433

LE-NAM TRAN, MATS BENGTTSSON AND BJÖRN OTTERSTEN

Stockholm 2012

Signal Processing
ACCESS Linnaeus Centre
KTH Royal Institute of Technology

Iterative Precoder Design and User Scheduling for Block-Diagonalized Systems

Le-Nam Tran, *Member, IEEE*, Mats Bengtsson, *Senior Member, IEEE*, and Björn Ottersten, *Fellow, IEEE*

Abstract—The block diagonalization (BD) scheme is a low-complexity suboptimal precoding technique for multiuser MIMO downlink channels, which completely precancels the multiuser interference. Accordingly, the precoder of each user lies in the null space of other users' channel matrices. In this paper, we propose an iterative algorithm using QR decompositions (QRDs) to compute the precoders. Specifically, to avoid dealing with a large concatenated matrix, we apply the QRD to a sequence of matrices of lower dimensions. One problem of BD schemes is that the number of users that can be simultaneously supported is limited due to zero interference constraints. When the number of users is large, a set of users must be selected, and selection algorithms should be designed to exploit the multiuser diversity gain. Finding the optimal set of users requires an exhaustive search, which has too high computational complexity to be practically useful. Based on the iterative precoder design, this paper proposes a low-complexity user selection algorithm using a greedy method, in which the precoders of selected users are recursively updated after each selection step. The selection metric of the proposed scheduling algorithm relies on the product of the squared row norms of the effective channel matrices, which is related to the eigenvalues by the Hadamard and Schur inequalities. An asymptotic analysis is provided to show that the proposed algorithm can achieve the optimal sum rate scaling of the MIMO broadcast channel. The numerical results show that the proposed algorithm achieves a good trade-off between sum rate performance and computational complexity. When users suffer different channel conditions, providing fairness among users is of critical importance. To address this problem, we also propose two fair scheduling (FS) algorithms, one imposing fairness in the approximation of the data rate, and another directly imposing fairness in the product of the squared row norms of the effective channel matrices.

Index Terms—Multiuser MIMO, downlink channel, asymptotic analysis, block diagonalization, user scheduling, fairness scheduling.

Manuscript received June 28, 2011; revised December 12, 2011; accepted February 22, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Love.

Copyright (c) 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Le-Nam Tran was with Signal Processing Laboratory, ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden. He is now with Centre for Wireless Communications and Department of Communications Engineering, University of Oulu, Oulu, Finland. Email: le.nam.tran@ee.oulu.fi.

Mats Bengtsson and Björn Ottersten are with Signal Processing Laboratory, ACCESS Linnaeus Centre, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. E-mail: {mats.bengtsson,bjorn.ottersten}@ee.kth.se.

Björn Ottersten is also with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, L-1359 Luxembourg-Kirchberg, Luxembourg (email: bjorn.ottersten@uni.lu).

This research has been supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 228044.

I. INTRODUCTION

Multiple input multiple output (MIMO) systems have drawn a lot of attention over the last decade thanks to their capability of boosting the channel capacity without the need of additional bandwidth or power [1], [2]. MIMO systems exploit spatial degrees of freedom provided by multiple antennas at both ends of a communication link to support simultaneous multiple data streams. The pioneering work in the area focused on point-to-point transmission, but more recently, investigations on MIMO communication have shifted from single-user to multiuser contexts [3]. In this paper, we consider the downlink of a single cell, where a multi-antenna base station (BS) wants to send data to multiple multi-antenna receivers simultaneously. Recent information theoretic studies have proved that dirty paper coding (DPC) is the capacity-achieving transmission technique for MIMO downlink channels [4]–[6]. However, implementing DPC is still a challenging task because of its successive nature of encodings and decodings, especially when the number of users is large. Furthermore, DPC is very sensitive to channel errors.

To avoid the complexity of DPC, there is a large interest in developing low-complexity precoding methods. In [7], a zero-forcing beamforming (ZFBF) for the case where all receivers have a single antenna was proposed, which is also known as channel inversion in [8]. Block diagonalization, introduced in [9], can be considered as an extension of ZFBF to cope with the case where receivers employ multiple antennas. Unlike DPC, BD completely presubtracts the multiuser interference, and thus decomposes a multiuser MIMO (MU-MIMO) channel into a set of parallel single-user MIMO (SU-MIMO) channels. This is accomplished by designing the precoding matrix of each user to lie in the null space of the remaining users' channel matrices. The traditional precoder design is based on the singular value decomposition (SVD) of a concatenated matrix, which requires high complexity [9]. In this paper, we assume perfect channel knowledge at the transmitter. First, we propose a low-complexity iterative precoder design based on the QRD. The idea behind the proposed method is that the precoding matrix of each user is designed to sequentially suppress the interference caused by other users. In this way, the precoders can be found by recursively applying the QRD to a sequence of matrices of lower dimensions. The computational analysis shows that our proposed method has lower complexity than the SVD-based method in [9].

The condition for the existence of the null space requires that the number of users is limited by the number of transmit antennas [9]–[12]. Suppose that the maximum number of

supportable users is $K_{\max}$. When K , the number of users, is larger than $K_{\max}$, the BS should select a set of users that maximizes a utility function. The best solution to this problem is to search over all possible subsets of users which is of size $\sum_{i=1}^{K_{\max}} \binom{K}{i}$. When K becomes large, the size of the search space is of the order $\mathcal{O}(K_{\max}^K)$, and thus the resulting complexity grows exponentially with K . To reduce the complexity, greedy algorithms are commonly employed in a large number of existing low-complexity user selection approaches [10]–[15]. In contrast to full search methods where a set of users are considered together, the general concept of greedy algorithms is to select one user, at each iteration, that best fits a selection criterion. For example, the c-algorithm proposed in [10] iteratively selects from the set of unselected users the user that provides the highest total throughput together with the previously selected users. It can be easily understood that this algorithm exhibits a good performance as the selection criterion is based directly on the sum rate. However, the frequent use of the SVD to obtain the singular values of efficient channel matrices and the water-filling process at each iteration is highly complex. In [10], the n-algorithm, which is based on the Frobenius norm of the effective channel matrix, was also introduced. The n-algorithm can greatly reduce the complexity but has a worse performance because the Frobenius norm is generally not a reliable quantity to reflect the true capacity of a wireless channel. Further improvements of the n-algorithm have been proposed in [11], [16]. In [12], a low-complexity user selection algorithm based on an upper bound on the capacity was proposed. The idea is that the sum rate of a MU-MIMO downlink channel is upper bounded by that of a SU-MIMO channel, which is formed by letting users cooperate. However, the gap between the sum rate of the BD scheme and this bound is large since the BD scheme is a suboptimal precoding technique. Other works in the area use a volume based-algorithm which requires the computation of the determinant of matrices [14], [15]. Also these algorithms suffer from a high computational burden.

Considering the problem of choosing a set of users that provides the maximal sum rate, we propose a greedy low-complexity scheduling algorithm, in which the product of squared row norms of the effective channels is used as the selection metric. This idea arises from the fact that there is a close relation between eigenvalues and row norms of a Hermitian matrix by the Hadamard and Schur inequalities [17]. In each selection step, the proposed algorithm seeks a user, from the set of users that are not selected yet, that provides the maximum product of squared row norms of effective channels. In this way, we avoid using any SVD, and the resulting user selection and precoder design rely only on QR decompositions. When adding one user to the set of selected users, the precoders of these selected users must be redesigned since the precoders of already selected users should eliminate the interference caused by the newly added user. This problem applies to any user selection algorithms for the BD approach based on a greedy method. Instead of finding the null space of a temporary aggregate channel matrix of high dimensions as in existing algorithms [10], [11], [15], we apply a QR decomposition to matrices of lower dimensions.

This is an extension of the proposed iterative QRD-based precoder design to solve the scheduling problem. Particularly, this recursive design can be utilized to further reduce the complexity of existing greedy user selection algorithms. Note that the precoders of selected users in the n-algorithm in [10] are redesigned by applying SVD to the aggregate channel matrix without referring to the precoders in the previous step. We show by an asymptotic analysis that, when the number of users is large, the proposed user selection algorithm can achieve a sum rate that has the same scaling law as that of DPC.

One practical matter that any scheduling algorithm should take into account is how to maintain fairness among all users. This problem is important for the case, where users have different fading conditions. A scheduling algorithm that maximizes the sum rate is more likely to select users with favorable channel conditions, and thus leaves users with poor conditions starved. The difficulty in addressing the fairness issue with a greedy method is that the data rate of each user is unknown until the whole selection process terminates. Thus, a quantity that approximates the data rate in each selection step should be defined [11], [18]. In this paper, two FS algorithms are proposed. In the first FS algorithm, we approximate the data rate of each user in each selection step by summing the logarithms of the squared row norms of the corresponding effective channels. To further lower the complexity, in the second FS algorithm, we directly employ the fairness in the product of the squared row norms. In other words, the second FS algorithm aims at maintaining fairness among the products of the squared row norms, rather than fairness among the actual rates of users.

The rest of the paper is organized as follows. Section II overviews MU-MIMO downlink channels with block diagonalization, and presents a novel precoder design for the BD method based on QRD. This section also discusses the dimensionality constraint and problem of user selection for the BD scheme. Section III presents the proposed greedy user scheduling algorithm for sum rate maximization based on product of squared row norms of effective channels. The computational complexity analysis of the proposed scheduling algorithm is also provided in this section. Two FS algorithms are described in Section IV to cope with the problem of how to maintain fairness among users. An asymptotic analysis for a large number of users is presented in Section V, and numerical results through various simulation setups and insights are given in Section VI. Section VII concludes the paper with some remarks.

Notation: Standard notations in the literature are used in this paper. Bold lower and upper case letters represent vectors and matrices, respectively; $\mathbf{H}^H$ and $\mathbf{H}^T$ are Hermitian and normal transpose of $\mathbf{H}$, respectively; $\|\mathbf{H}\|_F$ is the Frobenius norm of $\mathbf{H}$. For an $M \times N$ matrix $\mathbf{H}$, we denote by $\gamma(\mathbf{H})$ the product of the squared row norms of $\mathbf{H}$, i.e., $\gamma(\mathbf{H}) = \prod_{m=1}^M \sum_{n=1}^N |h_{m,n}|^2$; $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$; $\mathcal{N}(\mathbf{H})$ denotes the null space of $\mathbf{H}$; $[x]_+ = \max(0, x)$.

II. MU-MIMO DOWNLINK CHANNELS WITH BLOCK DIAGONALIZATION

A. System model

Consider a single cell MU-MIMO downlink channel with a BS and K users. The number of antennas at the BS and the k th user is denoted by N and $M_k \geq 1$, respectively. We denote this downlink channel by the notation $N \times \{M_1, M_2, \dots, M_K\}$. In our work, data from the BS to each user is transmitted by using all available antennas at both ends. In other words, the transmit and/or receive antenna selection are not considered. Accordingly, the channel matrix between the BS and the k th user is represented by $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$, which is assumed to be perfectly known at the transmitter.

Let $\mathcal{S} \subset \{1, 2, \dots, K\}$ denote a set of users to whom the BS simultaneously transmits data over the same radio resources (time, frequency). For the k th user, let $\mathbf{x}_k \in \mathbb{C}^{M_k \times 1}$ be the vector of transmitted symbols. In linear precoding methods, $\mathbf{x}_k$ is multiplied with a precoder $\mathbf{T}_k \in \mathbb{C}^{N \times M_k}$. We assume that $E[\mathbf{x}_k \mathbf{x}_k^H] = \mathbf{I}$ and the power allocation pattern for user k is absorbed in the covariance matrix $\mathbf{\Omega}_k = \mathbf{T}_k \mathbf{T}_k^H$. The received signal at the k th user is given by

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{T}_k \mathbf{x}_k + \sum_{j < k} \mathbf{H}_k \mathbf{T}_j \mathbf{x}_j + \sum_{j > k} \mathbf{H}_k \mathbf{T}_j \mathbf{x}_j + \mathbf{n}_k, \quad (1)$$

where $\mathbf{n}_k \in \mathbb{C}^{M_k \times 1}$ is the white complex-Gaussian noise vector with zero mean and covariance matrix $\mathbf{I}_{M_k}$. The summations of the right-hand side of (1) represent the multiuser interference.

Transmission strategies for MU-MIMO downlink channels are characterized by how we treat the interference between users in (1). For example, in DPC schemes, the data transmitted to each user is sequentially encoded in such a way that the k th user only sees the signals intended for users $k+1$ through $\mathcal{S}$ as interference. That is, the encoding process completely cancels the noncausal interference ($\sum_{j < k} \mathbf{H}_k \mathbf{T}_j \mathbf{x}_j$) caused by users 1 through $k-1$. Although DPC was proved to be the capacity achieving transmission strategy for MIMO downlink channels, the actual algorithm of DPC, which is based on nonlinear precoding, may be too complex for practical applications [19]. Hence, it is desirable to develop a low-complexity precoding scheme.

B. Iterative Precoder Design for BD

Block diagonalization, proposed in [9], is a suboptimal low-complexity solution that completely cancels the multiuser interference in (1). The precoder $\mathbf{T}_k$ is a cascade of two precoding matrices $\mathbf{B}_k$ and $\mathbf{\Pi}_k$, i.e., $\mathbf{T}_k = \mathbf{B}_k \mathbf{\Pi}_k$, where $\mathbf{B}_k$ is designed to remove the interference, and $\mathbf{\Pi}_k$ is used for power allocation to optimize a performance measure, such as the sum rate under a total power constraint, or the total transmit power under individual rate constraints [9]. First, $\mathbf{B}_k$ is designed such that $\mathbf{H}_j \mathbf{B}_k = \mathbf{0}$ for all $k \neq j$. In [9], an approach based on SVD was introduced to find $\mathbf{B}_k$. Specifically, given a set of users $\mathcal{S}$, let us consider the matrix

$$\bar{\mathbf{H}}_k = [\mathbf{H}_1^H \ \mathbf{H}_2^H \ \dots \ \mathbf{H}_{k-1}^H \ \mathbf{H}_{k+1}^H \ \dots \ \mathbf{H}_{|\mathcal{S}|}^H]^H \in \mathbb{C}^{\sum_{i \neq k} M_i \times N}. \quad (2)$$

Then the precoding matrix $\mathbf{B}_k$ for user k lies in $\mathcal{N}(\bar{\mathbf{H}}_k)$. Denote the SVD of $\bar{\mathbf{H}}_k$ as

$$\bar{\mathbf{H}}_k = \bar{\mathbf{U}}_k \bar{\mathbf{\Sigma}}_k [\bar{\mathbf{V}}_k^{(1)} \ \bar{\mathbf{V}}_k^{(0)}]^H, \quad (3)$$

where $\bar{\mathbf{V}}_k^{(1)}$ consists of the right singular vectors of $\bar{\mathbf{H}}_k$, and $\bar{\mathbf{V}}_k^{(0)} \in \mathbb{C}^{N \times \bar{M}_k}$, where $\bar{M}_k = N - \sum_{i \neq k} M_i$, forms a basis of $\mathcal{N}(\bar{\mathbf{H}}_k)$. To satisfy the zero-interference constraint, we can simply choose $\mathbf{B}_k = \bar{\mathbf{V}}_k^{(0)}$, and BD schemes decompose a MU-MIMO channel into a set of parallel SU-MIMO channels.

The method to design precoders mentioned above employs SVD to find the null space of a complex matrix of size $(\sum_{i \neq k} M_i) \times N$ for each $k \in \mathcal{S}$, which is computationally costly. In this paper, we propose an iterative method based on QRD to design the precoding matrices for the BD systems. Comparing the two decomposition methods, the complexity of the QRD is lower and deterministic, whereas that of the SVD is data dependent. Other related works often use Gram-Schmidt orthogonalization (GSO) procedure to find the null space [10], [11]. However, we note that the best algorithms for QRD are not based on GSO but rather on Givens rotations or Householder transformations. As a result of applying the QRD to a matrix $\mathbf{H} \in \mathbb{C}^{m \times n}$ with $m < n$, we can write $\mathbf{H}$ as

$$\mathbf{H} = [\mathbf{L}_{m \times m} \ \mathbf{0}_{m \times (n-m)}] \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_{\text{null}}^H \end{bmatrix}, \quad (4)$$

where $\mathbf{L} \in \mathbb{C}^{m \times m}$ is an upper triangular matrix, $\mathbf{Q}_1 \in \mathbb{C}^{m \times n}$ contains the orthonormal basis of the row space of $\mathbf{H}$, and $\mathbf{Q}_{\text{null}} \in \mathbb{C}^{n \times (n-m)}$ forms an orthonormal basis of $\mathcal{N}(\mathbf{H})$, i.e., $\mathbf{H} \mathbf{Q}_{\text{null}} = \mathbf{0}$, and $\mathbf{Q}_{\text{null}}^H \mathbf{Q}_{\text{null}} = \mathbf{I}$. For brevity, we denote by $\text{null}(\mathbf{H})$ an orthonormal basis of $\mathcal{N}(\mathbf{H})$.

To be specific, let us illustrate the proposed iterative design based on the QRD of the precoding matrix for user 1, which is denoted by $\mathbf{B}_1^{(i)}$ at the i th iteration. In the first step, we initially set $\mathbf{B}_1^{(1)} = \mathbf{I}_N$. In the second step, let $\mathbf{B}_1^{(2)} = \text{null}(\mathbf{H}_2) \in \mathbb{C}^{N \times (N-M_2)}$.¹ Suppose that we have obtained $\mathbf{B}_1^{(i)} \in \mathbb{C}^{N \times (N - \sum_{k=2}^i M_k)}$ after the i th iteration such that $\mathbf{H}_k \mathbf{B}_1^{(i)} = \mathbf{0}$ for all $1 < k \leq i$. In the $(i+1)$ th step, we should find $\mathbf{B}_1^{(i+1)}$ that $\mathbf{H}_k \mathbf{B}_1^{(i+1)} = \mathbf{0}$ for all $1 < k \leq i+1$. The following lemma demonstrates how to obtain $\mathbf{B}_1^{(i+1)}$ from $\mathbf{B}_1^{(i)}$. The idea of this iterative design is based on the algorithm to find the intersection of two null spaces, introduced in [20]. We notice that the same concept was also used in [16] for user selection. Herein, we elaborate the algorithm to design the precoders for BD without user selection first, i.e., assuming the BS can support all users, and then extend this idea to address the problem of user scheduling with a new selection metric.

Lemma 1. Let $\mathbf{B}_1^{(i)} \in \mathbb{C}^{N \times (N - \sum_{k=2}^i M_k)}$ be the precoding matrix of user 1 after i iterations, then $\mathbf{B}_1^{(i+1)}$ can be found as

$$\mathbf{B}_1^{(i+1)} = \mathbf{B}_1^{(i)} \mathbf{G}, \quad (5)$$

where the columns of $\mathbf{G}$ span $\mathcal{N}(\mathbf{H}_{i+1} \mathbf{B}_1^{(i)})$. Moreover, $\mathbf{B}_1^{(i+1)}$ will be orthonormal if $\mathbf{B}_1^{(i)}$ and $\mathbf{G}$ are orthonormal.

¹The order by which the interference cancellation is performed depends on the performance metric of interest. For the sum rate maximization problem, the user ordering is irrelevant [7], [9].

Proof: After i steps, we already have $\mathbf{H}_k \mathbf{B}_1^{(i+1)} = \mathbf{H}_k \mathbf{B}_1^{(i)} \mathbf{G} = \mathbf{0}$ for all $1 < k \leq i$. Thus, we only need to find $\mathbf{G}$ such that it suppresses the interference caused by $\mathbf{H}_{i+1}$. Thus, the condition for $\mathbf{G}$ is written as

$$\mathbf{H}_{i+1} \mathbf{B}_1^{(i+1)} = \mathbf{H}_{i+1} \mathbf{B}_1^{(i)} \mathbf{G} = \mathbf{0}. \quad (6)$$

Eq. (6) holds if $\mathbf{G}$ lies in $\mathcal{N}(\mathbf{H}_{i+1} \mathbf{B}_1^{(i)})$, which completes the proof. ■

In our work, we choose $\mathbf{G} = \text{null}(\mathbf{H}_{i+1} \mathbf{B}_1^{(i)})$. Note that $\mathbf{H}_{i+1} \mathbf{B}_1^{(i)}$ of size $M_{i+1} \times (N - \sum_{k=2}^i M_k)$, and thus $\mathbf{G} \in \mathbb{C}^{(N - \sum_{k=2}^i M_k) \times (N - \sum_{k=2}^{i+1} M_k)}$. Thus, the number of columns of $\mathbf{B}_1^{(i)}$ is reduced by M_i after every iteration. Unlike the precoder design in [9] where the precoding matrix of a user is achieved by finding a basis for the null space of a concatenated matrix, the iterative method proposed in this paper designs the precoding matrix by applying the QRD to a sequence of matrices of lower dimensions. In particular, the precoder design of all other users in $\mathcal{S}$ can be carried out in parallel to get rid of redundant computations. For completeness, we summarize the procedure to design precoding matrices for all users in $\mathcal{S}$ in Algorithm 1. In the initialization step, we set the precoding matrix of the first user to be $\mathbf{I}_N$, i.e., $\mathbf{B}_1^{(1)} = \mathbf{I}_N$. In the second step, the precoding matrix of user 2 is simply found by setting $\mathbf{B}_2^{(2)} = \mathbf{B}_1^{(1)} \times \text{null}(\mathbf{H}_1 \mathbf{B}_1^{(1)}) = \text{null}(\mathbf{H}_1)$, and the precoding matrix of user 1 is updated by $\mathbf{B}_1^{(2)} = \mathbf{B}_1^{(1)} \times \text{null}(\mathbf{H}_2 \mathbf{B}_1^{(1)}) = \text{null}(\mathbf{H}_2)$, following Lemma 1, to cancel the interference induced by $\mathbf{H}_2$. Assume that we already obtained $\{\mathbf{B}_k^{(i)}\}_{k=1}^i$, the precoding matrices of users 1 through i , after i steps. Consequently, it holds that $\mathbf{H}_j \mathbf{B}_i^{(i)} = \mathbf{0}$ for all $j = 1, 2, \dots, i-1$. In the $(i+1)$ th step, the precoding matrix for the $(i+1)$ th is designed such that $\mathbf{H}_j \mathbf{B}_{i+1}^{(i+1)}$ for all $j = 1, 2, \dots, i$, i.e., $\mathbf{B}_{i+1}^{(i+1)}$ must lie in the intersection of $\mathcal{N}(\mathbf{H}_j)$, $j = 1, 2, \dots, i$. Since $\mathbf{B}_i^{(i)}$ has already belonged to the intersection of $\mathcal{N}(\mathbf{H}_j)$, for $j = 1, 2, \dots, i-1$, we can compute $\mathbf{B}_{i+1}^{(i+1)}$ as $\mathbf{B}_{i+1}^{(i+1)} = \mathbf{B}_i^{(i)} \times \text{null}(\mathbf{H}_i \mathbf{B}_i^{(i)})$, which is shown in line 4. The precoding matrices for already-added users are updated as $\mathbf{B}_m^{(i+1)} = \mathbf{B}_m^{(i)} \times \text{null}(\mathbf{H}_{i+1} \mathbf{B}_m^{(i)})$ shown in line 6, according to Lemma 1. This process continues until all precoders are found.

Algorithm 1 Iterative precoder design for BD schemes

Initialization

- 1: $i := 1$.
- 2: $\mathbf{B}_1^{(1)} = \mathbf{I}_N$.

Main loop

- 3: **repeat**
 - 4: $\mathbf{B}_{i+1}^{(i+1)} = \mathbf{B}_i^{(i)} \times \text{null}(\mathbf{H}_i \mathbf{B}_i^{(i)})$.
 - 5: **for** $m = 1$ to i **do**
 - 6: Update $\mathbf{B}_m^{(i+1)} = \mathbf{B}_m^{(i)} \times \text{null}(\mathbf{H}_{i+1} \mathbf{B}_m^{(i)})$.
 - 7: **end for**
 - 8: $i := i + 1$.
 - 9: **until** $i = |\mathcal{S}|$
-

Algorithm 1 outputs the precoding matrix $\mathbf{B}_k \in \mathbb{C}^{N \times \bar{M}_k}$ that forms an orthonormal basis of $\mathcal{N}(\bar{\mathbf{H}}_k)$ given in (2) for the k th user. Since columns of $\mathbf{B}_k$ and $\mathbf{V}_k^{(0)}$ both span $\mathcal{N}(\bar{\mathbf{H}}_k)$,

i.e., $\text{span}(\mathbf{B}_k) = \text{span}(\mathbf{V}_k^{(0)}) = \mathcal{N}(\bar{\mathbf{H}}_k)$, it is easy to see that the proposed iterative precoder design and the SVD-based method in [9] yield the same results. As will be shown later in Section III-B, the computational complexity of the iterative design presented above is lower than that of the SVD-based method introduced in [9]. Moreover, this iterative method allows for a low-complexity user scheduling algorithm for BD schemes, which is described in Section III.

Next, the precoder design for BD schemes is completed by finding $\mathbf{\Pi}_k$ to optimize a performance metric under some constraints. For example, the achievable sum rate of the BD scheme with a total power constraint P is given by

$$\begin{aligned} R^{\text{BD}}(\mathcal{S}) &= \max_{\mathbf{\Omega}_k: \sum_{k=1}^{|\mathcal{S}|} \text{tr}(\mathbf{\Omega}_k) \leq P} \sum_{k=1}^{|\mathcal{S}|} \log |\mathbf{I} + \mathbf{H}_k \mathbf{B}_k \mathbf{\Omega}_k \mathbf{B}_k^H \mathbf{H}_k^H| \\ &= \max_{\mathbf{\Omega}_k: \sum_{k=1}^{|\mathcal{S}|} \text{tr}(\mathbf{\Omega}_k) \leq P} \sum_{k=1}^{|\mathcal{S}|} \log |\mathbf{I} + \tilde{\mathbf{H}}_k \mathbf{\Omega}_k \tilde{\mathbf{H}}_k^H|, \end{aligned} \quad (7)$$

where $\mathbf{\Omega}_k = \mathbf{\Pi}_k \mathbf{\Pi}_k^H$, and $\tilde{\mathbf{H}}_k = \mathbf{H}_k \mathbf{B}_k$ is the effective channel matrix of the k th user. The maximization in (7) is a convex optimization problem, which is equivalent to finding the capacity of a SU-MIMO channel with channel given by

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H}_1 \mathbf{B}_1 & & \\ & \ddots & \\ & & \mathbf{H}_{|\mathcal{S}|} \mathbf{B}_{|\mathcal{S}|} \end{bmatrix}. \quad (8)$$

Due to the block diagonal structure of $\tilde{\mathbf{H}}$, the sum rate and the resulting power allocation matrix $\mathbf{\Pi}_k$ can be easily found by applying the water-filling algorithm to the eigenvalues of $\{\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H\}_{k=1}^{|\mathcal{S}|}$. More specifically, let $\boldsymbol{\lambda}_k = [\lambda_{k,1} \ \lambda_{k,2} \ \dots \ \lambda_{k,L_k}]$ be the vector of $L_k = \min(M_k, \bar{M}_k)$ eigenvalues of $\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H$. Then the sum rate of BD is expressed as

$$R^{\text{BD}}(\mathcal{S}) = \sum_{k=1}^{|\mathcal{S}|} R_k = \sum_{k=1}^{|\mathcal{S}|} \sum_{i=1}^{L_k} [\log_2 \mu \lambda_{k,i}]_+, \quad (9)$$

where $R_k = \sum_{i=1}^{L_k} [\log_2 \mu \lambda_{k,i}]_+$ is the data rate of user k , and μ is the water-level which is determined by the power constraint

$$\sum_{k=1}^{|\mathcal{S}|} \sum_{i=1}^{L_k} \left[\mu - \frac{1}{\lambda_{k,i}} \right]_+ = P. \quad (10)$$

In the sequel, we mainly focus on the problem of sum rate maximization. However, we note that other problems such as the power control problem, i.e., total power minimization under individual rate constraints, can be easily solved as well (see [9] for more details).

C. User selection

In the iterative design above, we have assumed that $\mathcal{N}(\bar{\mathbf{H}}_k)$ in (2) exists for every user k . The dimensionality constraint for the existence of $\mathcal{N}(\bar{\mathbf{H}}_k)$ puts a limit on the number of simultaneously supportable users. Recall that $\bar{\mathbf{H}}_k$ is of dimension $(\sum_{j=1, j \neq k}^{|\mathcal{S}|} M_j) \times N$, and the BS is able to transmit data to user k if $\mathcal{N}(\bar{\mathbf{H}}_k)$ has a dimension greater than zero. Assuming all the rows of $\bar{\mathbf{H}}_k$ are independent, this condition is equivalent

to $(\sum_{j=1, j \neq k}^{|\mathcal{S}|} M_j) < N$ because $\text{rank}(\tilde{\mathbf{H}}_k) = \sum_{j=1, j \neq k}^{|\mathcal{S}|} M_j$. For all users, the condition is given by

$$N > \max_{k \in \mathcal{S}} \sum_{j=1, j \neq k}^{|\mathcal{S}|} M_j. \quad (11)$$

The condition in (11) does not imply that $N \geq \sum_k^{|\mathcal{S}|} M_k$, and, in fact, there are many cases that $N < \sum_k^{|\mathcal{S}|} M_k$. As an example, the BD scheme can support all users of the downlink channel $6 \times \{3, 2, 2\}$. For such cases, the generalized channel inversion method proposed in [21], [22] is not applicable since the right inverse of $\mathbf{H}_S = [\mathbf{H}_1^H \ \mathbf{H}_2^H \ \dots \ \mathbf{H}_{|\mathcal{S}|}^H]^H$ does not exist. Suppose $M_k = M$ for all k , BS can send data to at most $K_{\max} = \lceil N/M \rceil$ users at a time, where $\lceil x \rceil$ is the ceiling function, i.e the smallest integer not less than x . When K is large, there are possibly many supportable sets $\mathcal{S}$ that satisfy the dimensionality constraint, i.e., $|\mathcal{S}| \leq K_{\max}$. Consider the sum rate problem, the multiuser diversity gain is achieved by selecting an optimal set of users that maximizes the sum rate in (9). Mathematically, the sum rate of the BD scheme with multiuser diversity is given by

$$R^{\text{BD}} = \max_{\substack{\mathcal{S} \subset \{1, 2, \dots, K\} \\ |\mathcal{S}| \leq K_{\max}}} R^{\text{BD}}(\mathcal{S}). \quad (12)$$

We easily see that the optimal set of users is found by searching over all possible sets. However, the complexity of an exhaustive search grows exponentially with the number of users. Thus, significant research efforts have been made to find suboptimal low-complexity algorithms that can achieve a significant fraction of the best performance provided by a full search method. In this paper, we propose a low-complexity user selection for BD, based on a greedy method.

III. PROPOSED GREEDY USER SELECTION ALGORITHMS

The problem of user selection for BD schemes has been studied extensively in the literature, for single-antenna receivers in [13], [18], [23]–[25], and for multiple-antenna receivers in [10]–[12], [14]. A general approach to deal with multiple-antenna receivers is to treat each antenna as a separate user as in [18]. In this regard, user scheduling is investigated in connection with receive antenna selection (RAS) [11], or with eigenmode selection [23], [24]. On the one hand such approaches efficiently exploit spatial multiplexing gains, but on the other hand they require more feedback information to be exchanged. Herein, we consider user scheduling for MIMO downlink channel without receive antenna or eigenmode selection.

The performance of a greedy user selection algorithm mainly depends on the selection metric. It can be easily understood that the c-algorithm proposed in [10] exhibits a good performance as the selection criterion is based directly on the sum rate. Note that, when x is relatively small, we have $\ln(1+x) \approx x$. Accordingly, in the low SNR regime, the sum rate of the BD method in (9) is closely related to the sum of eigenvalues of $\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H$, which is identical to the sum of the Frobenius norm of $\tilde{\mathbf{H}}_k$, and thus the Frobenius norm can be a reliable indicator for the sum rate of the BD scheme in this situation.

In this section, we propose a greedy low-complexity scheduling algorithm for the sum-rate maximization, in which the product of squared row norms of the effective channels is used as the selection metric in each user selection step. The motivation of the proposed scheduling algorithm is as follows. We can see that the data rate of the k th user in the BD scheme is related to eigenvalues of $\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H$, where $\tilde{\mathbf{H}}_k$ is the effective channel matrix. If the total transmit power is high enough, the data rate of an individual user is a monotonically increasing function of the determinant (i.e., product of eigenvalues) of the effective channel matrix. This is also true even for a low or moderate transmit power, since only the dominant eigenvalues will be used. Naturally, we can use the determinant as a criterion for user selection [14], [15]. Nevertheless, computing the determinant of a matrix requires high complexity. For a Hermitian matrix, the determinant is smaller than or equal to the product of the diagonal elements (Hadamard inequality [17]). Moreover, the vector of diagonal entries is majorized by the vector of eigenvalues (Schur inequality [17]). The strong connection between eigenvalues and diagonal elements is also used in [26]. Note that the i th diagonal element of $\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H$, i.e., $[\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H]_{i,i}$ is actually the i th squared row norm of $\tilde{\mathbf{H}}_k$.

A. Proposed algorithm

At each iteration in the selection process of a greedy method, the precoders of previously selected users must be redesigned to eliminate the interference induced by the newly added user. This procedure is computationally costly. Motivated by the iterative QRD-based design presented in Section II-B, we introduce a recursive method that can efficiently compute the precoders of selected users, thereby significantly reducing the complexity.

The main operations of the proposed algorithm are summarized in Algorithm 2, where Φ_n and Ω_n store the set of remaining and selected users, respectively, $\mathbf{B}_k^{(n)}$ denotes the precoding matrix of the k th user in the set of selected users, and $\mathbf{Q}$ constitutes the null space of the rows of selected users at the n th selection step. In the initialization step, the proposed algorithm finds a user that yields the maximum product of squared row norms of the channel matrix. Then, the precoding matrix of this selected user is initially set to be the identity matrix, and $\mathbf{Q}$ is an orthonormal basis for the null space of the selected channel matrix. At the n th selection step, for each unselected user, the proposed algorithm

- immediately finds the temporary precoding matrix, and effective channel matrix for that user, because the precoding matrix is actually an orthonormal basis for the null space of the rows of all selected channel matrices, which is already stored in $\mathbf{Q}$ (line 7 in Algorithm 2),
- calculates the resulting effective channel matrices for the previously selected users to cancel the interference induced by the currently considered user (line 11 in Algorithm 2), and
- finds the user that produces the maximum product of squared row norms of effective channels together with those selected users.

After one user is added to the set of selected users, we update the null space of all selected channel matrices (line 14 in Algorithm 2) and the precoding matrices for the previously selected users (line 16 in Algorithm 2). This process repeats until the maximum number of users that the BS can support is reached. Specifically, after step 2, $K_{\max} = \lceil N/M \rceil$ users are selected because the proposed algorithm first aims at maximizing the multiplexing gain, i.e., attempting to serve as many users as possible. In fact, the selection process should terminate when the total throughput starts to drop if more users are added. This is true for the case where users' channel matrices are highly correlated. In such cases, it is better to schedule a number of users less than $K_{\max}$. Thus, in the finalization step, we apply the c-algorithm in [10] to find the final set of selected users and the corresponding sum rate. In this regard, steps 1 and 2 can be viewed as a *coarse* selection process where the algorithm attempts to search for a set of good candidates, and step 3 is a *fine* selection process.

Algorithm 2 Product of squared row norms-based user selection algorithm for sum rate maximization.

Step 1: Initialization

- 1: Let $\Phi_1 = \{1, 2, \dots, K\}$ and $\Omega_1 = \emptyset$.
- 2: Select the first user $u_1 = \arg \max_{k \in \Phi_1} \gamma(\mathbf{H}_k)$.
- 3: Precoding matrix of u_1 : $\mathbf{B}_{u_1}^{(1)} = \mathbf{I}_{N_t}$.
- 4: Let $\mathbf{Q} := \text{null}(\mathbf{H}_{u_1})$.
- 5: Update $\Phi_2 := \Phi_1 - u_1$; $\Omega_2 = u_1$; $n := 2$.

Step 2: Main loop

- 6: **for** each $m \in \Phi_n$ **do**
- 7: Set $\mathbf{B}_m = \mathbf{Q}$.
- 8: Calculate the effective channel matrix $\tilde{\mathbf{H}}_m^{(n)}$ of user m :
 $\tilde{\mathbf{H}}_m^{(n)} = \mathbf{H}_m \mathbf{B}_m$.
- 9: **for** each $k \in \Omega_n$ **do**
- 10: Set $\mathbf{G}_m = \text{null}(\mathbf{H}_m \mathbf{B}_k^{(n-1)})$.
- 11: Update the effective channel matrices of previously selected users $\tilde{\mathbf{H}}_k^{(n)} = \tilde{\mathbf{H}}_k^{(n-1)} \mathbf{G}_m$.
- 12: **end for**
- 13: Select the user that maximizes $u_n = \arg \max_{m \in \Phi_n} \gamma_m \gamma_1 \dots \gamma_{n-1}$ where $\gamma_k = \gamma(\tilde{\mathbf{H}}_k^{(n)})$.
- 14: Let $\mathbf{Q}$ be an orthonormal basis for all rows of the channel matrices of already selected users.
- 15: **for** each $k \in \Omega_n$ **do**
- 16: Update the precoders of previously selected users due to the newly added user: $\mathbf{B}_k^{(n)} = \mathbf{B}_k^{(n-1)} \mathbf{G}_{u_n}$.
- 17: Store $\tilde{\mathbf{H}}_k^{(n)}$, $k \in \Omega_n$ to the next iteration.
- 18: **end for**
- 19: Update: $\Phi_{n+1} := \Phi_n - u_n$; $\Omega_{n+1} = \Omega_n \cup u_n$; $n := n + 1$.
- 20: **end for**

Step 3: Finalization

- 21: Apply the c-algorithm in [10] to find the final set of selected users and the corresponding sum rate.
-

Remark 1. As a matter of fact, the effective channels of selected users $\tilde{\mathbf{H}}_k^{(n)}$ are not necessary to be stored after each iteration since we can write $\tilde{\mathbf{H}}_k^{(n)} = \mathbf{H}_k \mathbf{B}_k^{(n)}$. However, this

computation involves multiplication an $M \times N$ matrix and an $N \times (N - \sum_{j=1}^n M_j)$ matrix as mentioned in Section II-B, which approximately needs $8MN(N - \sum_{j=1}^n M_j)$ flops [10].² However, according to the proposed iterative precoder design, the dimensions of $\mathbf{B}_k^{(n)}$, $\tilde{\mathbf{H}}_k^{(n)}$, and $\mathbf{G}$ are reduced after every iteration. Thus, in Algorithm 2, the effective channel matrices are updated recursively as $\tilde{\mathbf{H}}_k^{(n)} = \tilde{\mathbf{H}}_k^{(n-1)} \mathbf{G}$, and this computation only requires $8M(N - \sum_{j=1}^{n-1} M_j)(N - \sum_{j=1}^n M_j)$ flops which needs lower complexity. Recall that the n-algorithm in [10] finds a null space of an $(n-1)M \times N$ matrix at the n th selection step.

It is commonly known that the sum rate of the BD approach depends on the orthogonality of users. The multiuser diversity gain comes from the fact that, for a large number of users, we can find a set of users which are nearly orthogonal to each other. The complexity of the proposed algorithm can be further reduced by dropping the users who have large correlation with the set of selected users. In fact, this idea was used to develop the ZFBF-SUS algorithm in [18] for single-antenna receivers, where the channel matrix becomes a vector, and the orthogonality between vectors is well defined in the literature. In our work, we deal with MIMO channels, and a quantity needs to be defined to measure the degree of the orthogonality between two matrices. For this purpose, we simply define the orthogonality between two matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ as

$$0 \leq \varepsilon(\mathbf{H}_1, \mathbf{H}_2) = \frac{\|\mathbf{H}_1 \mathbf{H}_2^H\|_F}{\|\mathbf{H}_1\|_F \|\mathbf{H}_2\|_F} \leq 1. \quad (13)$$

This quantity is called the spatial correlation coefficient in [11]. Suppose $\mathbf{Q}$ is an orthonormal basis for all rows of the channel matrices of selected users after n steps. In the $(n+1)$ th step, we constrain the search for a user in the reduction set

$$\Phi_{n+1} = \{k \in \Phi_n | \varepsilon(\mathbf{H}_k, \mathbf{Q}) > \zeta\}, \quad (14)$$

where $0 \leq \zeta \leq 1$ is a predetermined correlation threshold. The physical meaning is that when the channel matrix of a user in the set of remaining users is highly correlated with the null space of the space spanned by the rows of selected users' channel matrices, that user will be kept in future selection steps. The choice of ζ affects both the complexity and sum rate performance of the proposed algorithm. When ζ is high, many users are ignored after each iteration, and much computational complexity can be saved. However, the achievable throughput is degraded since we may accidentally remove some potential users. Conversely, when ζ is small, very few users are excluded, and the sum rate converges to that with $\zeta = 0$, where no additional user is excluded after every iteration. We discuss the optimal range for ζ with numerical results in Sect. V.

B. Complexity analysis

In this subsection, we analyze the complexity of Algorithms 1 and 2. The complexity is measured by the number of flops as in [10], [12], denoted as ψ . Although flop counting is a

²A flop is equal to a real floating point operation [20]. A real addition, multiplication, or division is counted as one flop. A complex addition and multiplication have two flops and six flops, respectively.

crude measurement of the true computational complexity, it captures the order of the computation load. The number of flops for typical operations is given as follows. One complex addition and multiplication take 2 and 6 flops, respectively. Multiplication of a matrix $\mathbf{A} \in \mathbb{C}^{m \times p}$ and a matrix $\mathbf{B} \in \mathbb{C}^{p \times n}$ requires $8mpn$ flops (mpn complex multiplications and mpn complex additions). The number of flops needed to compute QRD of a real matrix of size $m \times n$, $m \geq n$ with fast Given transformations is $2n^2(m - n/3)$ [20]. For a complex matrix of the same size, we approximate the number of flops for a QRD by $4n^2(3m - n)$, i.e., treating every operation as complex multiplication.

For simplicity, we show the complexity for the case where all users have the same number of antennas, i.e., $M_k = M$ for all k , and the maximum number of users that BS can support simultaneously is $K_{\max} = N/M$. First, we compare the complexity of Algorithm 1 to that of the SVD-based method when N and M are fixed, and there are $K = K_{\max}$ users, which are all scheduled by the BS. For the SVD-based method, it takes $(24M(K_{\max} - 1)N^2 + 48M^2(K_{\max} - 1)^2N + 54M^3(K_{\max} - 1)^3)$ flops to design the precoder of each user [10]. Thus, the total number of flops required for the SVD-based method is easily computed as

$$\psi_{\text{SVD}} = K_{\max} \{ (24M(K_{\max} - 1)N^2 + 48M^2(K_{\max} - 1)^2N + 54M^3(K_{\max} - 1)^3) \}. \quad (15)$$

Since $\mathbf{B}_1^{(1)}$ is initialized with the identity matrix, we only need to perform the QRD to $\mathbf{H}_1$ and $\mathbf{H}_2$ to find $\mathbf{B}_2^{(2)}$ and $\mathbf{B}_1^{(2)}$, respectively. For $i \geq 2$, to compute $\mathbf{B}_{i+1}^{(i+1)}$, we need $8MN[N - (i-1)M]$ flops for $\mathbf{H}_{i+1}\mathbf{B}_i^{(i)}$, $4M^2[3(N - (i-1)M) - M]$ flops for $\text{null}(\mathbf{H}_{i+1}\mathbf{B}_i^{(i)})$, and $8N[N - (i-1)M](N - iM)$ flops for $\mathbf{B}_i^{(i)} \times \text{null}(\mathbf{H}_{i+1}\mathbf{B}_i^{(i)})$. The same number of flops is needed to update $\mathbf{B}_m^{(i+1)}$ for $m = 1, 2, \dots, i$. Thus, the complexity of Algorithm 1 is given by

$$\begin{aligned} \psi_{\text{Algo. 1}} = & 8M^2(3N - M) + \sum_{i=2}^{K_{\max}-1} (i+1) \{ 8MN[N \\ & - (i-1)M] + 4M^2[3(N - (i-1)M) - M] \\ & + 8N[N - (i-1)M](N - iM) \}. \end{aligned} \quad (16)$$

In Fig 1, we plot the number of flops of the SVD-based method and Algorithm 1 with the number of transmit antennas, where $M_k = 2$ for all k . As we can see, a substantial complexity reduction can be obtained by Algorithm 1, compared to the SVD-based method.

Next, we analyze the complexity of Algorithm 2, where the number of users K is larger than $K_{\max}$. In fact, the proposed algorithm may terminate with the number of selected users less than $K_{\max}$ if adding more users reduces the sum rate. Thus, it is difficult to provide an exact expression for the complexity. Here, we consider the complexity for the case where BS selects $K_{\max}$ users. As mentioned above, the complexity of the proposed algorithm is a function of the correlation threshold ζ . However, an analytical (or closed-form) expression that characterizes the dependency between the complexity and the correlation threshold is hard to find, because we do not know exactly how many users are dropped after every selection

Fig. 1. Complexity comparison of the SVD-based method and Algorithm 1, $M = 2$, $K = K_{\max}$.

step for a given value of ζ . In what follows, we analyze the complexity when $\zeta = 0$, i.e., the set of remaining users is not further reduced.

In Algorithm 2, selecting the first user requires $4KMN$ flops, and computing the row basis for the rows of $\mathbf{H}_{u_1}$ via QRD needs $4M^2(3N - M)$ flops. For each $m \in \Phi_n$, $n \geq 2$, we need $8MN(N - (n-1)M)$ flops for $\tilde{\mathbf{H}}_m$; $8MN(N - (n-2)M) + 4M^2(3N - (n-1)M)$ flops for $\mathbf{G}$; and $8M(N - (n-1)M)(N - (n-2)M)$ for $\tilde{\mathbf{H}}_k^{(n)}$. The calculation of the objective function takes $4nM(N - (n-1)M)$ flops. Updating the null space $\mathbf{Q}$ takes $4M^2(3N - nM)$, and $\mathbf{B}_k^{(n)}$ takes $8N(N - (n-1)M)(N - (n-2)M)$ for each $k \in \Omega_n$. Therefore, the total number of flops of Algorithm 2 is

$$\begin{aligned} \psi_{\text{Algo. 2}} \approx & 4KMN + 4M^2(3N - M) + \sum_{n=2}^{K_{\max}} (K - n + 1) \{ 8MN \\ & (N - (n-1)M) + (n-1) [8MN(N - (n-2)M) \\ & + 4M^2(3N - (n-1)M) + 8M(N - (n-1)M)(N \\ & - (n-2)M)] + 4nM(N - (n-1)M) \} + 4M^2(3N \\ & - nM) + (n-1)8N(N - (n-1)M)(N - (n-2)M) \\ & = \mathcal{O}(K[K_{\max}]^2 MN^2). \end{aligned} \quad (17)$$

The number of flops of the c-algorithm, n-algorithm and Algorithm 2 is compared in Fig. 2. The complexity of all three scheduling algorithms grows linearly with the number of users. Algorithm 2 shows the lowest complexity.

IV. FAIR SCHEDULING

Algorithm 2 presented in the previous section aims at maximizing the total throughput. Thus, the algorithm will favor users with high SNR. In a cellular network, user fairness is of particular concern since users at the cell edge may not receive any throughput due to severe intercell interference. Fair scheduling attempts to provide a certain degree of fairness among users. In this section, we modify the proposed algorithm in the previous section to address this problem. The problem of user selection with fairness is usually obtained by

Fig. 2. Complexity comparison of various user scheduling algorithms, $N = 8$, $M = 2$.

optimizing a suitably weighted sum rate expression

$$R^{\text{BD}}(t) = \max_{\sum_k \text{tr}(\Omega_k) \leq P} \sum_{k=1}^{|S|} \mu(t) R_k(t), \quad (18)$$

where, for proportional fairness (PF), $\mu(t) = 1/\bar{R}_k(t)$ and $R_k(t)$ are the weighting factor and the instant rate of the k th user during the t th scheduling interval, respectively. Note that when $\mu(t) = 1$, the maximization problem in (18) is identical to maximizing the sum rate in (9). The PF algorithm serves the users with sufficiently high $R_k(t)/\bar{R}_k(t)$. Consequently, the PF algorithm is able to achieve high throughput and maintain proportional fairness among all users by giving priority to users with a high-quality channel rate $R_k(t)$ and a low current average service rate $\bar{R}_k(t)$. If a user receives less throughput in previous slots, its average service rate is low, the possibility of being scheduled in subsequent slots increases. The average service rate of the k th user at time t is recursively updated as

$$\bar{R}_k(t) = \begin{cases} \delta \bar{R}_k(t-1) + (1-\delta) R_k(t) & \text{if selected} \\ \delta \bar{R}_k(t-1) & \text{if not selected} \end{cases} \quad (19)$$

where δ is the forgetting factor. If the data rate of each user is averaged over a window of T slots, then $\delta = 1 - 1/T$ [18].

Since the selection metric of the c-algorithm in [10] is directly based on the data rate of each user of the BD scheme, it is straightforward to modify this algorithm to incorporate fairness among users. Specifically, at the n th selection step, we choose a user such that

$$u_n = \arg \max_{k \in \Phi_n} \left(\mu_k(t) R_k(t) + \sum_{i \in \Omega_{n-1}} \mu_i(t) R_i(t) \right), \quad (20)$$

where Φ_n , and Ω_{n-1} are the sets of remaining users and selected users at the n th and $(n-1)$ th selection step, respectively. The temporary data rates $R_k(t)$ and $R_i(t)$ can be attained by water-filling to the corresponding eigenvalues. This modification of the c-algorithm is referred to as ‘PF c-algorithm’ in this paper. The main problem with Algorithm 2 (and other existing selection algorithms, of which the selection criterion is not directly based on the sum rate) is that the exact

Algorithm 3 PF scheduling with $\log_2(1 + \gamma)$

Step 1: Initialization

- 1: All steps are analogous to those in Algorithm 2, except for
- 2: Select the first user as

$$u_1 = \arg \max_{k \in \Phi} \mu_k(t) \hat{R}_k(t). \quad (22)$$

Step 2: Main loop

- 3: All steps are analogous to those in Algorithm 2, except for
- 4: Select the user that maximizes

$$u_n = \arg \max_{m \in \Phi_n} \left(\mu_k(t) \hat{R}_k(t) + \sum_{i \in \Omega_{n-1}} \mu_i(t) \hat{R}_i(t) \right). \quad (23)$$

Step 3: Finalization

- 5: Apply the PF c-algorithm to find the final set and actual rate of selected users. Update the average data rate as (19).
-

data rate $R_k(t)$ in each selection step is unknown until the whole user selection process completes. To solve this problem, we approximate the data rate of the k th user at the n th iteration step as

$$\hat{R}_k(t) = \sum_{i=1}^{M_k} \log_2 \left(1 + \frac{P}{M^{(n)}} d_{k,i} \right), \quad (21)$$

where $M^{(n)} = \sum_{k=1}^n M_k$ is the total of data streams at the n th iteration, and $d_{k,i} = [\tilde{\mathbf{H}}_k^{(n)} \tilde{\mathbf{H}}_k^{(n)H}]_{i,i}$ is the i th squared row norm of $\tilde{\mathbf{H}}_k^{(n)}$. The rate approximation in (21) is inspired by the equal power allocation scheme with eigenvalues being replaced by the squared row norms. In fact, the true sum rate of the k th user is obtained from water-filling a total power P to $M^{(n)}$ eigenvalues of the corresponding effective channels. However, we cannot express the sum rate of the k th user with water-filling in a closed-form. If equal power allocation is used and M_k eigenvalues of $\tilde{\mathbf{H}}_k^{(n)} \tilde{\mathbf{H}}_k^{(n)H}$ are replaced by the squared row norms of $\tilde{\mathbf{H}}_k^{(n)}$, we obtain an approximation of data rate shown in (21). With this approximation of the data rate, the proposed PF scheduling is described in Algorithm 3. After steps 1 and 2, the true rates of selected users are still unknown. Thus, in step 3, we apply the PF c-algorithm to find the true sum rate which is needed to update the average data rate, and accordingly the weighting factor. In this way, the weighting factor is updated with the actual rate, while the instant rate is approximated by (21).

We can further simplify Algorithm 3 by directly employing fairness in the product of squared row norms of the effective channels, which is shown in Algorithm 4. Unlike Algorithm 3, Algorithm 4 provides fairness to the product of squared row norms of the effective channels, rather than to the actual rate or approximation of the actual rate. This follows from the approximation in (21), since

$$2^{R_k(t)} = \prod_{i=1}^{M_k} \left(1 + \frac{P}{M^{(n)}} d_{k,i} \right) \approx \left(\frac{P}{M^{(n)}} \right)^{M_k} \prod_{i=1}^{M_k} d_{k,i} \sim \gamma(\tilde{\mathbf{H}}_k^{(n)}). \quad (24)$$

Algorithm 4 Fair scheduling with γ

Step 1: Initialization

- 1: All steps are analogous to those in Algorithm 2, except for
- 2: Select the first user

$$u_1 = \arg \max_{k \in \Phi} \mu_k(t) \gamma_k(t), \quad (28)$$

where $\gamma_k(t) = \gamma(\mathbf{H}_k)$.

Step 2: Main loop

- 3: All steps are analogous to those in Algorithm 2, except for
- 4: Select the user that maximizes

$$u_n = \arg \max_{m \in \Phi_n} \mu_m(t) \gamma_m(t) \prod_{i \in \Omega_{n-1}} \mu_i(t) \gamma_i(t), \quad (29)$$

where $\gamma_k(t) = \gamma(\tilde{\mathbf{H}}_k^{(n)})$.

Step 3: Finalization

- 5: The average value of $\gamma_k(t)$ is updated as

$$\bar{\gamma}_k(t) = \begin{cases} \delta \bar{\gamma}_k(t-1) + (1-\delta) \gamma_k(t) & \text{if selected} \\ \delta \bar{\gamma}_k(t-1), & \text{if not selected} \end{cases} \quad (30)$$

and $\mu_k(t) = 1/\bar{\gamma}_k(t)$.

The selection criterion in (23) can be written as

$$\begin{aligned} u_n &= \arg \max_{k \in \Phi_n} 2^{\mu_k(t) R_k(t) + \sum_{i \in \Omega_{n-1}} \mu_i(t) R_i(t)} \\ &= \arg \max_{k \in \Phi_n} 2^{\mu_k(t) R_k(t)} \prod_{i \in \Omega_{n-1}} 2^{\mu_i(t) R_i(t)} \\ &= \arg \max_{k \in \Phi_n} \gamma(\tilde{\mathbf{H}}_k^{(n)})^{\mu_k(t)} \prod_{i \in \Omega_{n-1}} \gamma(\tilde{\mathbf{H}}_i^{(n)})^{\mu_i(t)}. \end{aligned} \quad (25)$$

We note that, in [11], a FS algorithm was proposed which maintains fairness among the Frobenius norm of the effective channel. Eq. (25) shows that we can impose the user fairness through the product of squared row norms of the effective channels. Herein, to yield a different degree of fairness, we let $\mu_k(t) = 1/\bar{\gamma}_k(t)$, i.e., $\mu_k(t)$ is the inverse of the average of $\gamma_k(t)$ in previous time slots, and consider the *weighted product* of the squared row norms given in (29) of Algorithm 4. In fact, Algorithm 4 is a greedy method to find a suboptimal set of users of the following scheduling problem

$$\max_{\substack{S \in \{1, 2, \dots, K\} \\ |S| \leq K_{\max}}} \sum_{k \in S} \log\left(\frac{\gamma_k}{\bar{\gamma}_k}\right). \quad (26)$$

Assuming $\gamma_k/\bar{\gamma}_k > 1$, we have inequality $\log(\gamma_k/\bar{\gamma}_k) \leq \gamma_k/\bar{\gamma}_k - 1$. Thus, (26) is bounded above by

$$\max_{\substack{S \in \{1, 2, \dots, K\} \\ |S| \leq K_{\max}}} \sum_{k \in S} \frac{\gamma_k}{\bar{\gamma}_k}. \quad (27)$$

The scheduling problem described in (27) is actually a scheduler that yields proportional fairness to γ_k . As a result, Algorithm 4 is less fair than Algorithm 3, as numerically shown in Fig. 7.

V. ASYMPTOTIC ANALYSIS

In this section, we analyze the asymptotic sum rate of Algorithm 2 presented in Section III when N and M are fixed and the number of users, K , is going to infinity. To simplify notations, we assume that $M_k = M$ for all k , and $N = M \times K_{\max}$, where $K_{\max}$ is consequently the supportable number of users. In the case of single-antenna receivers, significant research efforts have been made to prove the asymptotic optimality of ZFBF in terms of multiuser diversity gain. An asymptotic analysis of a greedy selection method for the BD scheme seems to be intractable due to several reasons. First, the data rate of the selected users is reduced after each selection step since the precoders of already selected users must be redesigned to deal with the interference induced when a new user is added. Second, the data rate of each selected user is determined by the vector of eigenvalues of the corresponding effective channel matrix, and thus the joint distribution of the eigenvalues is needed to study the sum rate. To simplify the analysis, current works often consider single-antenna receivers, or treat each receive antenna as an independent user, or assign at most one data stream to each user [18], [23]–[25]. In this paper, a user is assigned either all possible data streams or no data stream at all. To facilitate the analysis, we make the following assumptions.

- (i) The BS is always able to find a set of $K_{\max}$ users to communicate.
- (ii) Transmit power is equally allocated to all selected users. In fact, the optimal power allocation strategy can be found through the water-filling process which adapts allocated power according to the instantaneous channel gains. This will make the analysis intractable. It is well known that the gain difference between equal power and water-filling allocation schemes is minimal in the high SNR regime.

The following theorem presents the asymptotic result of the sum rate of Algorithm 2

Theorem 1. *In the limit of large K and high SNR regime, the sum rate of the proposed algorithm is bounded by*

$$R \leq N \log\left(\frac{P}{NM} \log K + P \log \log K + \mathcal{O}(\log \log \log K)\right) + \mathcal{O}(1). \quad (31)$$

Proof: Please refer to the Appendix. ■

For large K , Theorem 1 shows that the achievable sum rate of Algorithm 2 scales as $N \log \log K$, which is the scaling law of the sum rate of DPC for a MIMO broadcast channel with an N -antenna BS and K single-antenna receivers. If we treat each antenna of a receiver as if it were a separate user, the resulting scaling law is $N \log \log MK$. However, as mentioned earlier, a large amount of feedback information has to be exchanged to exploit the RAS, which is not considered in this paper. Theorem 1 indicates that, when N , P , and K are fixed, increasing the number of receive antennas M reduces the achievable sum rate, which is the penalty for not exploiting RAS. Further discussions are given with numerical results in Section VI.

VI. NUMERICAL RESULTS

In this section, the sum rate performance of the BD with the proposed scheduling algorithms is numerically evaluated and compared with that of the following greedy algorithms:

- The c-algorithm and n-algorithm proposed in [10]. The c-algorithm is claimed to obtain a sum rate very close to the exhaustive search method.
- The algorithm based on the capacity upper bound proposed in [12].

A quasi-static fading model is used in our simulation, where independent realizations of $\mathbf{H}_k$ are generated for each scheduling interval. We consider a MU-MIMO downlink channel with $N = 8$ antennas at the BS and $M = 2$ antennas at each user. The numerical study considers both uncorrelated and correlated channels. We note that the performance of most existing user scheduling algorithms for BD schemes have been numerically evaluated for uncorrelated channels [10]–[12], [14], [15], [23], [27]. Thus, how these algorithms work in the correlated channels is still a question. Here, the MIMO channel matrix is simply modeled as $\mathbf{H}_k = \mathbf{H}_{k,\omega} \mathbf{R}_{k,t}^{1/2}$ (i.e., the Kronecker model without receive correlation), where $\mathbf{H}_{k,\omega}$ is a matrix of i.i.d. complex Gaussian distribution with zero mean and unit variance and $\mathbf{R}_{k,t} = \mathbf{R}_{k,t}^{1/2} \mathbf{R}_{k,t}^{H/2}$ is an $N \times N$ transmit covariance matrix [28], [29]. In our simulation setup, the exponential correlation model is used, i.e., $[\mathbf{R}_{k,t}]_{i,j} = \rho_k^{|i-j|}$, where the correlation coefficient for the k th user $\rho_k = \rho e^{j\phi_k}$ with ϕ_k being i.i.d. uniformly distributed over $[0, 2\pi]$ [30]–[32].

Fig. 3 plots the average sum rate versus the number of users for $\rho = 0$ (uncorrelated channels) and $\rho = 0.95$ (highly correlated channels), and the total power is $P = 10$ dB. We can see that the c-algorithm outperforms the other methods, because the selection criterion is directly based on the sum rate of the BD scheme. Algorithm 2 achieves a higher sum rate than the n-algorithm and the upper bound-based algorithm. When the correlation is high, the n-algorithm and Algorithm 2 yields a similar throughput, while the method based on an upper bound on the channel capacity is dramatically inferior. In addition, the numerical results for the case $P = 20$ dB, which are not shown in Fig. 3, indicate that the sum rate of the n-algorithm is dramatically lower than that of Algorithm 2. The reason is that the Frobenius norm is not a reliable indicator of the channel capacity in the high SNR region.

In Fig. 4, we investigate the impact of transmit-side channel correlation on the multiuser diversity gain by varying the value of ρ . The sum rate of the proposed algorithm versus the number of users is drawn for three cases when $\rho = 0.25$, $\rho = 0.5$, and $\rho = 0.95$. Our numerical results indicate that antenna correlation may be beneficial in exploiting the multiuser diversity gain. Specifically, for a given total transmit power budget, we observe a crossing point between the sum rate curves for systems with and without transmit correlation. Above this point, channel correlation *improves* the sum rate performances and vice versa. This observation also applies to other greedy scheduling algorithms in comparison. In fact, as the channel correlation increases, the rate of a single user MIMO channel is mostly determined by the dominant

Fig. 3. Average sum rate versus number of users with and without transmit correlation, $N = 8$, $M = 2$.

eigenmodes. From a physical perspective, the mainlobe of the beam for the BS to a user becomes narrower. As a result, when the number of users is large, we can select a group of users whose dominant eigenmodes are nearly orthogonal. In the absence of channel correlation, the transmission strategy is to distribute power to all eigenmodes. Thus, the beams of selected users become more correlated, and thereby degrading the sum rate performance of BD approach. In the case of a small number of users, channel correlation reduces the sum rate since it reduces the spatial degrees of freedom, and multiuser diversity gain is not available.

Fig. 5 illustrates how the average sum rate and complexity reduction of Algorithm 2 depends on the choice of the correlation threshold ζ . If ζ is too large, the total throughput dramatically deteriorates due to the loss of multiuser diversity gain, i.e., a lot of potential users are dropped after every selection step (Fig 5(a)). In turn, the complexity is reduced when ζ is large (Fig 5(b)). Decreasing ζ improves the system throughput, at the expense of an increased computational complexity. It is of particular interest to find an optimal value for ζ such that the acquired sum rate is close to that without further dropping users after each step, i.e., when $\zeta = 0$.

Fig. 4. Average sum rate versus number of users for Algorithm 2 with different values of ρ , $N = 8$, $M = 2$.

Fig. 6. Average sum rate of Algorithm 2 with different number of receive antennas, $N = 8$, $P = 20$ dB.

(a) Average sum rate

(b) Computation load reduction

Fig. 5. Average sum rate and computational complexity reduction of Algorithm 2 versus correlation threshold ζ , $N = 8$, $M = 2$.

Unfortunately, this optimal choice of ζ is difficult to find analytically. Our numerical results reveal that the good choices for ζ for the spatially uncorrelated multiuser channels are from 0.87 to 0.92, depending on the number of users in the systems. For the case of $K = 40$, setting $\zeta = 0.9$ can reduce 80% the

computational complexity, and still achieve 96% of the sum rate for $\zeta = 0$. For a fixed ζ , too many users may be neglected after every iteration if the correlation among users' channels is high, e.g. when users lie in a narrow beam seen from the base station. This can lead to a severe degradation on the sum rate of Algorithm 2. Thus, in practice, a good setting of ζ would require some mechanism that adaptively adjusts the correlation threshold based on the channel conditions of the users that happen to be active and scheduled for the moment. However, this problem is beyond scope of the paper, and probably needs more extensive studies. Note that, the optimal range for ζ does not change with the total power transmit because ζ only measures the correlation between the subspaces, not the actual magnitudes of users' channel matrices. Same observations were also reported in [11] for the Frobenius norm based scheduling algorithm.

Fig. 6 plots the average sum rate of Algorithm 2 with different values of M , the number of receive antennas per user, for N and P fixed. For illustration purpose, we denote a MU-MIMO downlink channel with N transmit antennas at the BS, and K users, each with M receive antennas by the notation (N, K, M) , and its achievable sum rate by $R_{(N, K, M)}$. When K is small (e.g., $K = 4$), users are spatially separated with high probability. Thus, increasing M offers more spatial degrees of freedom, and we are still able to find a set of users whose channels are highly uncorrelated, leading to an improvement in the sum rate. The effect is opposite for large K , i.e., the sum rate is reduced as M increases. The asymptotic analysis in Section V indicates that this situation also holds true for the upper bounds on the sum rate. Moreover, the experimental results with other greedy scheduling algorithms such as the c-algorithm and n-algorithm in [10] reveal the same properties. This problem is due to the fact that RAS diversity gain is not exploited. Specifically, consider the $(8, 40, 2)$ downlink channel. Treating each antenna of a receiver as if it were a separate user results in the $(8, 80, 1)$ downlink channel. By the RAS diversity gain, it follows that $R_{(8, 80, 1)} > R_{(8, 40, 2)}$. Since the multiuser diversity gain scales as $\log \log K$, we can approximate $R_{(8, 40, 1)} \approx R_{(8, 80, 1)}$, and thus $R_{(8, 40, 1)} > R_{(8, 40, 2)}$. By the same arguments, we have

Fig. 7. Average rates of individual users with different FS algorithms for asymmetric SNRs, $N = 8$, $M = 2$.

$$R_{(8,40,2)} > R_{(8,40,4)}.$$

We plot the data rates for individual users for the case of asymmetric SNRs in Fig. 7. In this simulation, a set of $K = 40$ users is considered with the received SNRs varying according to the loglinear scale [11], [18]. Specifically, the channel matrix for the k th user is expressed as $\mathbf{H}_k = \sqrt{\alpha_k} \mathbf{H}_{k,\omega}$ where $\mathbf{H}_{k,\omega}$ is a matrix of i.i.d. complex Gaussian distribution with zero mean and unit variance. The weighing factor α_k of user k is on the loglinear scale from $\text{SNR}_{\min} = 0$ dB to $\text{SNR}_{\max} = 20$ dB, i.e., $\log_{10}(\alpha_{k+1}) = \frac{\text{SNR}_{\max} - \text{SNR}_{\min}}{K-1}k + \text{SNR}_{\min}$, for $k = 0, 1, \dots, K-1$. Furthermore, the values of α_k are normalized such that $\sum_{k=1}^K \alpha_k = K$. From Fig. 7, we can see that the maximum sum rate algorithm (i.e., Algorithm 2) favors the users with high SNR. More explicitly, users 1 through 20 do not receive any data at all, while users 30 through 40 occupy most of the system throughput. The PF c-algorithm and Algorithm 3 can maintain a similar degree of fairness among users since Algorithm 3 is a low-complexity approximation of the PF c-algorithm. In particular, Algorithm 4 achieves a higher throughput than Algorithm 3, but results in a worse fairness level. Recall that Algorithm 4 attempts to maintain the fairness index of the product of the row norms of users' effective channels. Accordingly, Algorithm 4 pays more attention to users with high instantaneous data rate than users with low average serviced rate, due to the nonlinear relation between the product of row norms and the true capacity of a MIMO channel.

VII. CONCLUSIONS

We are concerned with the precoder design and user scheduling problem for MU-MIMO downlink channels using BD. First, an iterative precoder design based on QR decompositions is proposed. The idea of this iterative approach is then used to solve the scheduling problem, where a low-complexity user selection algorithm for sum rate maximization is proposed. The decision metric at each selection step of the proposed scheduling algorithm is based on the product of squared row norms of the effective channel matrices. Our algorithm requires lower complexity than other greedy selection methods because the use of SVD is completely

avoided, and the precoders of previously selected users at each iteration are efficiently updated in a recursive way, using QR decompositions. The numerical results evaluated with and without channel correlation demonstrate that the proposed algorithm can achieve a better sum rate than other greedy scheduling algorithms of similar complexity. We conclude that product of squared row norms is probably a more reliable indicator of channel capacity for a MIMO channel than the Frobenius norm. We also consider the problem of maintaining fairness among users which plays a crucial role in many environments such as cellular networks, where users usually experience different channel conditions. In this regard, two FS algorithms are proposed, one based upon an approximation of the data rate, another imposing fairness into the product of squared row norms. The proposed methods show a good trade-off between performance and computational complexity to other techniques.

APPENDIX PROOF OF THEOREM 1

In this appendix, we prove the asymptotic result stated in Theorem 1. Let $\boldsymbol{\lambda}$ be the vector of eigenvalues of the effective channels resulting from the set of selected users for a channel realization. The average data rate is given by

$$E\{R_{\text{BD}}\} \approx E\left\{\sum_{k=1}^{K_{\max}} \sum_{i=1}^M \log\left(1 + \frac{P}{N} \lambda_{k,i}\right)\right\} \quad (32a)$$

$$\leq E\left\{\sum_{k=1}^{K_{\max}} \sum_{i=1}^M \log\left(1 + \frac{P}{N} d_{k,i}\right)\right\} \quad (32b)$$

$$\approx E\left\{\sum_{k=1}^{K_{\max}} \log\left(1 + \left(\frac{P}{N}\right)^M \underbrace{\prod_{i=1}^M d_{k,i}}_{\gamma_k}\right)\right\}, \quad (32c)$$

where $d_{k,i}$ is the i th squared row norm of the matrix $\tilde{\mathbf{H}}_k$. The approximation (32a) is due to the assumptions (i) and (ii). Inequality (32b) follows from the fact that $\log(x)$ is concave and $\mathbf{d}_k$ is majorized by $\boldsymbol{\lambda}_k$ [17]. The approximation in (32c) is due to the assumption of high SNR. In (32c), γ_k is the product of squared row norms of the effective channel of the selected users at the k th iteration of the proposed algorithm, which can be further bounded by

$$\gamma_k = \max_{1 \leq j \leq K-k+1} \gamma(\tilde{\mathbf{H}}_j) \quad (33a)$$

$$\leq \max_{1 \leq j \leq K-k+1} \gamma(\mathbf{H}_j) \quad (33b)$$

$$\leq \max_{1 \leq j \leq K-k+1} \left(\frac{\|\mathbf{H}_j\|_F^2}{M}\right)^M, \quad (33c)$$

where inequality (33b) is true by noting that $d_{k,i} = [\tilde{\mathbf{H}}_k \tilde{\mathbf{H}}_k^H]_{i,i} = [\mathbf{H}_k \mathbf{B}_k \mathbf{B}_k^H \mathbf{H}_k^H]_{i,i} \leq [\mathbf{H}_k \mathbf{H}_k^H]_{i,i}$, and inequality (33c) follows from the inequality between arithmetic and geometric mean values. From (33c), we can bound the average sum rate by

$$E\{R_{\text{BD}}\} \leq E\left\{\sum_{k=1}^{K_S} \log\left(1 + \left(\frac{P}{NM} \mu_k\right)^M\right)\right\}, \quad (34)$$

where μ_k is the maximum of $(K - k + 1)$ i.i.d Chi-squared random variables with $2NM$ degrees of freedom, according to (33c). Denote by $F(x) = \gamma(NM, x)/(NM - 1)!$ where $\gamma(k, x)$ is the lower incomplete Gamma function, and $f(x) = x^{NM-1}e^{-x}/(NM - 1)!$ the cdf and pdf of a Chi-squared random variable with $2NM$ degrees of freedom. From order statistics [33], we obtain the cdf and pdf of μ_k as

$$F_{\mu_k}(x) = F^{K-k+1}(x), \quad (35)$$

$$f_{\mu_k}(x) = (K - k + 1) \times F^{K-k}(x)f(x). \quad (36)$$

Applying an asymptotic result as $K \rightarrow \infty$ (see [34] and references therein), we have

$$\Pr\{|\mu_k - \theta| \leq \log \log K\} \approx 1 - \mathcal{O}(1/\log K), \quad (37)$$

where θ is found as

$$\theta = \log K + (NM - 1) \log \log K + \mathcal{O}(\log \log \log K). \quad (38)$$

From (34) and (37), we arrive at

$$\begin{aligned} E\{R_{BD}\} &\leq E\left\{\sum_{k=1}^{K_{\max}} \log\left(1 + \left(\frac{P}{NM}\mu_k\right)^M\right)\right\} \\ &\approx N \log \frac{P}{NM} + M \times E\left\{\sum_{k=1}^{K_{\max}} \log \mu_k\right\}. \end{aligned} \quad (39)$$

The key to bound (39) is to write

$$\begin{aligned} E\left\{\sum_{k=1}^{K_{\max}} \log \mu_k\right\} &= \Pr(\mu_k \leq \theta + \log \log K) \\ &\quad \times E\left\{\sum_{k=1}^{K_{\max}} \log \mu_k \middle| \mu_k \leq \theta + \log \log K\right\} \\ &\quad + \Pr(\mu_k \geq \theta + \log \log K) \\ &\quad \times E\left\{\sum_{k=1}^{K_{\max}} \log \mu_k \middle| \mu_k \geq \theta + \log \log K\right\} \\ &\leq K_{\max} \log(\theta + \log \log K) \\ &\quad + \mathcal{O}(1/\log K) \log(1 + \frac{P}{N}K) \\ &\leq K_{\max} \log(\theta + \log \log K) + \mathcal{O}(1). \end{aligned} \quad (40)$$

Inserting (40) into (39) results in

$$\begin{aligned} E\{R_{BD}\} &\leq N \log \frac{P}{NM} + N \log(\log K + NM \log \log K) \\ &\quad + \mathcal{O}(\log \log \log K) + \mathcal{O}(1) \\ &= N \log\left(\frac{P}{NM} \log K + P \log \log K\right) \\ &\quad + \mathcal{O}(\log \log \log K) + \mathcal{O}(1), \end{aligned} \quad (41)$$

which completes the proof.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers, whose careful consideration of the manuscript improved the presentation of the material.

REFERENCES

- [1] E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–598, Nov. 1999.
- [2] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311–335, Mar. 1998.
- [3] D. Gesbert, M. Kountouris, R. Heath, C.-B. Chae, and T. Salzer, "Shifting the MIMO paradigm," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 36–46, Sep. 2007.
- [4] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.
- [5] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2658–2668, 2003.
- [6] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.
- [7] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.
- [8] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part I: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.
- [9] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.
- [10] Z. Shen, R. Chen, J. Andrews, J. Heath, R.W., and B. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658–3663, Sep. 2006.
- [11] S. Sigdel and W. Krzymien, "Simplified fair scheduling and antenna selection algorithms for multiuser MIMO orthogonal space-division multiplexing downlink," *IEEE Trans. Veh. Technol.*, vol. 58, no. 3, pp. 1329–1344, Mar. 2009.
- [12] X. Zhang and J. Lee, "Low complexity MIMO scheduling with channel decomposition using capacity upperbound," *IEEE Commun. Lett.*, vol. 56, no. 6, pp. 871–876, Jun. 2008.
- [13] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.
- [14] B. H. Wang, H. T. Hui, and Y. T. Yu, "Maximum volume criterion for user selection of multiuser MIMO downlink with multiantenna users and block diagonalization beamforming," in *Proc. IEEE APSURS 2010*, Jul. 2010, pp. 1–4.
- [15] L. Jin, X. Gu, and Z. Hu, "A novel volume-based scheduling scheme for multiuser multiple-input multiple-output downlink system," in *Proc. IEEE RWS 2010*, Jan. 2010, pp. 448–451.
- [16] Z. Youtuan, T. Zhihua, and Z. Jinkang, "An improved norm-based user selection algorithm for multiuser MIMO systems with block diagonalization," in *In Proc. VTC 2007-Fall*, Oct. 2007, pp. 601–605.
- [17] E. Jorswieck and H. Boche, "Majorization and matrix-monotone functions in wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 6, pp. 553–701, 2007.
- [18] 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.
- [19] W. Yu, D. Varodayan, and J. Cioffi, "Trellis and convolutional precoding for transmitter-based interference presubtraction," *IEEE Trans. Commun.*, vol. 53, no. 7, Jul. 2005.
- [20] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The John Hopkins Univ. Press, 1996.
- [21] H. Sung, S.-R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3489–3499, Nov. 2009.
- [22] W. Li and M. Latva-aho, "An efficient channel block diagonalization method for generalized zero forcing assisted MIMO broadcasting systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, pp. 739–744, Mar. 2011.
- [23] L. Sun and M. McKay, "Eigen-based transceivers for the MIMO broadcast channel with semi-orthogonal user selection," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5246–5261, Oct. 2010.
- [24] A. Bayesteh and A. Khandani, "On the user selection for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1086–1107, Mar. 2008.

- [25] J. Wang, D. Love, and M. Zoltowski, "User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3713–3726, Aug. 2008.
- [26] Le-Nam Tran and Een-Kee Hong, "Multiuser diversity for successive zero-forcing dirty paper coding: Greedy scheduling algorithms and asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3411–3416, Jun. 2010.
- [27] S. Sigdel, R. Elliott, W. Krzymien, and M. Al-Shalash, "Greedy and genetic user scheduling algorithms for multiuser MIMO systems with block diagonalization," in *Proc. IEEE VTC 2009-Fall*, Sep. 2009, pp. 1–6.
- [28] C.-N. Chuah, D. Tse, J. Kahn, and R. Valenzuela, "Capacity scaling in MIMO wireless systems under correlated fading," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 637–650, Mar. 2002.
- [29] J. Kermoal, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 6, pp. 1211–1226, Aug. 2002.
- [30] S. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, Sep. 2001.
- [31] C. Martin and B. Ottersten, "Asymptotic eigenvalue distributions and capacity for MIMO channels under correlated fading," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1350–1359, Jul. 2004.
- [32] K. Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach, "Second order statistics of NLOS indoor MIMO channels based on 5.2 GHz measurements," in *Proc. IEEE GLOBECOM 2001*, vol. 1, Nov. 2001, pp. 156–160.
- [33] H. A. David, *Order Statistics*, 3rd ed. Wiley, 2003.
- [34] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.
- [14] B. H. Wang, H. T. Hui, and Y. T. Yu, "Maximum volume criterion for user selection of multiuser MIMO downlink with multiantenna users and block diagonalization beamforming," in *Proc. IEEE APSURSI*, Jul. 2010, pp. 1–4.
- [15] L. Jin, X. Gu, and Z. Hu, "A novel volume-based scheduling scheme for multiuser multiple-input multiple-output downlink system," in *Proc. IEEE RWS*, Jan. 2010, pp. 448–451.
- [16] Z. Youtuan, T. Zhihua, and Z. Jinkang, "An improved norm-based user selection algorithm for multiuser MIMO systems with block diagonalization," in *In Proc. VTC 2007-Fall*, Oct. 2007, pp. 601–605.
- [17] E. Jorswieck and H. Boche, "Majorization and matrix-monotone functions in wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 6, pp. 553–701, 2007.
- [18] 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.
- [19] W. Yu, D. Varodayan, and J. Cioffi, "Trellis and convolutional precoding for transmitter-based interference presubtraction," *IEEE Trans. Commun.*, vol. 53, no. 7, Jul. 2005.
- [20] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The John Hopkins Univ. Press, 1996.
- [21] H. Sung, S.-R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3489–3499, Nov. 2009.
- [22] W. Li and M. Latva-aho, "An efficient channel block diagonalization method for generalized zero forcing assisted mimo broadcasting systems," vol. 10, no. 3, pp. 739–744, Mar. 2011.
- [23] L. Sun and M. McKay, "Eigen-based transceivers for the MIMO broadcast channel with semi-orthogonal user selection," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5246–5261, Oct. 2010.
- [24] A. Bayesteh and A. Khandani, "On the user selection for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1086–1107, Mar. 2008.
- [25] J. Wang, D. Love, and M. Zoltowski, "User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3713–3726, Aug. 2008.
- [26] Le-Nam Tran and Een-Kee Hong, "Multiuser diversity for successive zero-forcing dirty paper coding: Greedy scheduling algorithms and asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3411–3416, Jun. 2010.
- [27] S. Sigdel, R. Elliott, W. Krzymien, and M. Al-Shalash, "Greedy and genetic user scheduling algorithms for multiuser MIMO systems with block diagonalization," in *Proc. IEEE VTC 2009-Fall*, Sep. 2009, pp. 1–6.
- [28] C.-N. Chuah, D. Tse, J. Kahn, and R. Valenzuela, "Capacity scaling in MIMO wireless systems under correlated fading," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 637–650, Mar. 2002.
- [29] J. Kermoal, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 6, pp. 1211–1226, Aug. 2002.
- [30] S. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, Sep. 2001.
- [31] C. Martin and B. Ottersten, "Asymptotic eigenvalue distributions and capacity for MIMO channels under correlated fading," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1350–1359, Jul. 2004.
- [32] K. Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach, "Second order statistics of NLOS indoor MIMO channels based on 5.2 GHz measurements," in *IEEE GLOBECOM*, vol. 1, Nov. 2001, pp. 156–160.
- [33] H. A. David, *Order Statistics*, 3rd ed. Wiley, 2003.
- [34] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

REFERENCES

PLACE
PHOTO
HERE

Le-Nam Tran (M'09) received the B.S. degree in Electrical Engineering from Ho Chi Minh National University of Technology, Vietnam in 2003, and M.S and PhD in Radio Engineering from Kyung Hee University, Republic of Korea, in 2006 and 2009, respectively. His PhD study was supported by Korean Government IT Scholarship Program.

In 2009, he joined the Department of Electrical Engineering, Kyung Hee University, Republic of Korea, as a lecturer. From September 2010 to July 2011, he was a postdoc fellow at the Signal Processing Laboratory, ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Sweden. Since August 2011, he has been with Centre for Wireless Communications and Department of Communications Engineering, University of Oulu, Finland. His current research interests include mobile communication systems, equalization techniques, multiuser MIMO systems. He received the Best Paper Award from IITA in August 2005.

PLACE
PHOTO
HERE

Mats Bengtsson (M'00-SM'06) received the M.S. degree in computer science from Linköping University, Linköping, Sweden, in 1991 and the Tech. Lic. and Ph.D. degrees in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1997 and 2000, respectively.

From 1991 to 1995, he was with Ericsson Telecom AB Karlstad. He currently holds a position as Associate Professor at the Signal Processing Laboratory, School of Electrical Engineering, KTH. His research interests include statistical signal processing and its applications to communications, multi-antenna processing, radio resource management, and propagation channel modelling. Dr. Bengtsson served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING 2007-2009 and is a member of the IEEE SPCOM Technical Committee.

PLACE
PHOTO
HERE

Björn Ottersten (S'87-M'89-SM'99-F'04) was born in Stockholm, Sweden, 1961. He received the M.S. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986. In 1990 he received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has held research positions at the Department of Electrical Engineering, Linköping University, the Information Systems Laboratory, Stanford University, and the Katholieke Universiteit Leuven, Leuven. During 96/97 Dr. Ottersten was Director of Research at ArrayComm Inc, San Jose, California, a start-up company based on Ottersten's patented technology. He has co-authored papers that received an IEEE Signal Processing Society Best Paper Award in 1993, 2001, and 2006, and 3 IEEE conference papers receiving Best Paper Awards. In 1991 he was appointed Professor of Signal Processing at the Royal Institute of Technology (KTH), Stockholm. From 2004 to 2008 he was dean of the School of Electrical Engineering at KTH and from 1992 to 2004 he was head of the department for Signals, Sensors, and Systems at KTH. Currently, Dr. Ottersten is Director for the Interdisciplinary Centre for Security, Reliability and Trust at the University of Luxembourg.

Dr. Ottersten has served as Associate Editor for the IEEE Transactions on Signal Processing and on the editorial board of IEEE Signal Processing Magazine. He is currently editor in chief of EURASIP Signal Processing Journal and a member of the editorial board of EURASIP Journal of Advances Signal Processing. Dr. Ottersten is a Fellow of the IEEE and EURASIP. In 2011 he received the IEEE Signal Processing Society Technical Achievement Award. He is a first recipient of the European Research Council advanced research grant. His research interests include wireless communications, stochastic signal processing, sensor array processing, and time series analysis.