A New Approach to User Scheduling in Massive Multi-User MIMO Broadcast Channels

Gilwon Lee, *Member, IEEE*, and Youngchul Sung, *Senior Member, IEEE*

***Abstract*— In this paper, a new two-phase-feedback-based user-scheduling-and-beamforming method is proposed for multi-user multiple-input multiple-output downlink in the context of two-stage beamforming. The key ideas of the proposed method are: 1) to use a set of *orthogonal reference* beams and construct a *cone* around each reference beam to select "nearly-optimal" semi-orthogonal users based only on channel quality indicator feedback; and 2) to apply *post-user-selection beam design* with zero-forcing beamforming (ZFBF) based on channel state information (CSI) feedback only from the selected users. It is proven that the proposed scheduling-and-beamforming method is asymptotically optimal as the number of users increase. Furthermore, the proposed scheduling-and-beamforming method almost achieves the performance of the existing semi-orthogonal user selection with ZFBF that requires full CSI for all users, with a significantly reduced amount of required channel information which is even less than that required by random beamforming.**

***Index Terms*— Multi-user MIMO, user scheduling, two-phase feedback, two-stage beamforming.**

I. Introduction

User scheduling is one of the crucial issues in MU-MIMO which supports multiple users simultaneously in the same frequency and time based on spatial-division multiplexing [2]. Since it was revealed that the performance degradation of linear beamforming schemes compared to optimal dirty-paper coding (DPC) [3]–[5] for a Gaussian MIMO broadcast channel is negligible with proper user scheduling as the number of users in the cell becomes large [6], [7], extensive research has been conducted on user scheduling for MU-MIMO for the last decade (see [2] and references therein). In this paper, we revisit the user-scheduling-and-beamforming problem for MU-MIMO downlink in the context of frequency-division duplexing (FDD) massive MU-MIMO downlink with two-stage beamforming [8], although the proposed user scheduling method can readily be applied to conventional single-stage multiple-input single-output (MISO) downlink.

In FDD massive MU-MIMO, downlink channel estimation is not easy since the number of channel parameters to estimate is very large compared to the number of available training symbols limited by the channel coherence time [9], [10]. In addition, CSI feedback for downlink user scheduling induces heavy system overhead. To circumvent these difficulties in FDD massive MU-MIMO, two-stage beamforming was proposed and studied for FDD massive MU-MIMO [8], [11]–[14]. In the two-stage beamforming strategy, users in the cell are divided into multiple spatial groups each having approximately the same channel covariance matrix through user grouping [15], and downlink beamforming at the base station (BS) is performed in two steps: first by an outer-beamformer and second by an inner-beamformer, as shown in Fig. 1, where the outer-beamformer between the actual channel and the inner-beamformer suppresses spatial inter-group interference and the inner-beamformer provides multiplexing among served in-group users based on ZFBF or minimum mean-square-error (MMSE) beamforming [8]. The key point in two-stage beamforming is that the outer-beamformer is designed based only on channel's statistical information such as group coverage angle not on CSI. Thus, the required CSI is only for the MU-MIMO inner-beamformer and is given by the effective channel (defined as the product of the actual channel and the outer-beamformer) of which dimension is much reduced compared to the actual channel dimension.

In this paper, we consider the problem of user scheduling for FDD massive MU-MISO downlink based on the aforementioned two-stage beamforming. The contributions of this paper are summarized as follows.

- A new two-phase-feedback-based user-scheduling-and-beamforming method for FDD massive MU-MISO with two-stage beamforming is proposed. The proposed method eliminates the necessity of full or coarse CSI feedback from all users but realizes the essence of user selection summarized as *selection of semi-orthogonal users with large channel norms* in a distributed manner by applying two-phase feedback.
- The asymptotic optimality is proved under the multi-group two-stage beamforming setting with inter-group interference.
- The proposed method suggests how the training beams for two-phase feedback scheduling are selected and what kind of CQI is required for the multi-group two-stage beamforming.
- When the new method is applied to the single-cell single-group case, it outperforms the major previous

Manuscript received May 25, 2017; revised October 17, 2017; accepted December 18, 2017. Date of publication December 25, 2017; date of current version April 16, 2018. This research was supported in part by the Cross-Ministry Giga KOREA Project of The Ministry of Science, ICT and Future Planning, Korea (GK14N0100, 5G mobile communication system development based on mmWave) and in part by "Global Leader Education Program for Future Electronics and Communications" of the BK21 Plus. This paper was presented in part at SPAWC 2014 [1]. The associate editor coordinating the review of this paper and approving it for publication was Y. J. Zhang. *(Corresponding author: Youngchul Sung.)*

G. Lee is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: gilwonlee@utexas.edu).

Y. Sung is with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea (e-mail: ysung@ee.kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2017.2786670

two-phase-feedback-based scheduling methods. Hence, there is an improvement over the existing methods in the single-cell single-group case.

- Based on the new idea in this paper, a new proportionally fair (PF) scheduling method with distributed individual control of the semi-orthogonality is proposed. This new method enhances user fairness over the simple PF scheme.

Related Works: The key reference for our invention is widely-considered SUS-ZFBF [7]. SUS-ZFBF sequentially selects semi-orthogonal users with large channel norms from all users in the cell based on full CSI at the BS, and yields good performance [7]. However, it requires CSI from all users at the BS and this induces heavy feedback overhead. There were some works to reduce feedback overhead based on two-phase feedback [16]–[19] in which coarse CSI is fed back to the BS from all users or a subset of users in the first-phase feedback, users for service are selected based on the first-phase feedback, and refined CSI is fed back to the BS only from the selected users in the second-phase feedback. However, such schemes still requires heavy feedback overhead since the first-phase feedback requires CSI-type information from users although it is coarse. There have been corrections of CQI-based RBF [6] to improve its performance by using two-phase feedback, e.g., [20], [21]. However, the proposed method here has the distinct feature that it defines semi-orthogonality-enforcing cones around the orthogonal reference beams and the quantity feedbacked in the first phase is the channel norm (or its modified version) and the cone index, and the first and second phases are well interweaved for the proposed method to yield better performance than the previous corrections [20], [21], as shown in Section VI.

Notations and Organization: Vectors and matrices are written in boldface with matrices in capitals. All vectors are column vectors. For matrix $\mathbf{A}$, $\mathbf{A}^T$, $\mathbf{A}^H$, $\text{tr}(\mathbf{A})$, and $[\mathbf{A}]_{i,j}$ indicate the transpose, conjugate transpose, trace, and entry at the i-th row and j-th column of $\mathbf{A}$, respectively. $\mathbf{A}(:, c_1 : c_2)$ is the submatrix of $\mathbf{A}$ consisting of the columns from c_1 to c_2. $\text{diag}(\mathbf{A}_1, \cdots, \mathbf{A}_n)$ denotes the diagonal matrix with diagonal elements $\mathbf{A}_1$, ... ,$\mathbf{A}_n$. $\|\cdot\|$ and $\|\cdot\|_F$ represent the 2-norm and the Frobenius norm, respectively. $\mathbf{I}_K$ is the $K \times K$ identity matrix. $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}, \Sigma)$ means that random vector $\mathbf{x}$ is complex Gaussian distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. $\mathbb{E}[\cdot]$ denotes statistical expectation. For a subspace $\mathcal{A}$, $\mathcal{A}^\perp$ denotes the orthogonal complement of $\mathcal{A}$. $\mathbb{R}$, $\mathbb{R}^+$, and $\mathbb{C}$ are the sets of real, non-negative real, and complex numbers, respectively. $\imath := \sqrt{-1}$.

The remainder of this paper is organized as follows. The system model is described in Section II. The proposed user-scheduling-and-beamforming method is presented in detail in Section III and its asymptotic optimality is proved in Section IV. Fairness issues are discussed in Section V. Numerical results are provided in Section VI, followed by conclusions in Section VII.

II. System Model

In this paper, we consider a single-cell MU-MISO downlink based on two-stage beamforming, where a single BS with M transmit antennas serves to K single-antenna users, as shown in Fig. 1. We assume that the users in the cell are partitioned into G spatial groups such that (s.t.) $K = \sum_{g=1}^{G} K_g$, where K_g is the number of users in group g. As in [8], we assume that each group has a different channel covariance matrix and every user in group g has the same channel covariance matrix $\mathbf{R}_g$, $g = 1, \cdots, G$. Thus, the $M \times 1$ channel vector $\mathbf{h}_{g_k}$ of user k in group g (or simply user g_k) is assumed to be given by

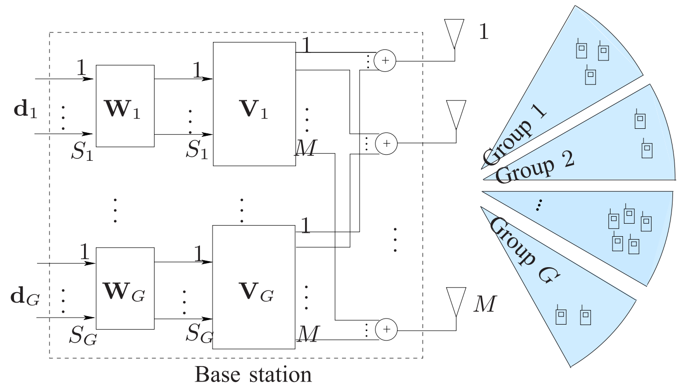


Fig. 1. Multi-group MU-MIMO downlink with two-stage beamforming.

$$\mathbf{h}_{g_k} = \mathbf{U}_g \Lambda_g^{1/2} \boldsymbol{\eta}_{g_k}, \tag{1}$$

where $\mathbf{R}_g = \mathbf{U}_g \Lambda_g \mathbf{U}_g^H$ is the eigendecomposition (ED) of $\mathbf{R}_g$; $\mathbf{U}_g$ is the $M \times r_g$ matrix composed of the orthonormal eigenvectors corresponding to the r_g non-zero eigenvalues of $\mathbf{R}_g$; Λ_g is the $r_g \times r_g$ diagonal matrix composed of the non-zero eigenvalues of $\mathbf{R}_g$; and $\boldsymbol{\eta}_{g_k} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{r_g})$. Let the elements of $\mathbf{U}_g$ and Λ_g be

$$\begin{aligned} \mathbf{U}_g &= [\mathbf{u}_{g,1}, \mathbf{u}_{g,2}, \cdots, \mathbf{u}_{g,r_g}], \\ \Lambda_g &= \text{diag}(\lambda_{g,1}, \cdots, \lambda_{g,r_g}), \ \ \lambda_{g,1} \geq \cdots \geq \lambda_{g,r_g} > 0. \end{aligned} \tag{2}$$

Denoting the $K_g \times M$ channel matrix for the users in group g by $\mathbf{H}_g = [\mathbf{h}_{g_1}, \cdots, \mathbf{h}_{g_{K_g}}]^H$ and stacking $\mathbf{H}_g, g = 1, \cdots, G$, we have the overall $K \times M$ channel matrix $\mathbf{H} = [\mathbf{H}_1^H, \cdots, \mathbf{H}_G^H]^H$. The received signal vector containing all user signals in the cell is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{3}$$

where $\mathbf{x}$ is the $M \times 1$ transmitted signal vector at the BS, $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ is the Gaussian noise vector, and the BS has a total average power constraint $\mathbb{E}[\|\mathbf{x}\|^2] \leq P$.

We assume that the BS selects S users with $S \leq M$ among the K users within the cell with S_g users selected from K_g users in group g s.t. $\sum_{g=1}^{G} S_g = S$, and broadcasts independent data streams to the selected users with one data stream for each selected user. In the considered two-stage beamforming, precoding of the data vector $\mathbf{d}$ is done in two steps: first, by an $S \times S$ MU-MIMO inner-beamformer $\mathbf{W}$ and then by an $M \times S$ outer-beamformer $\mathbf{V}$, i.e., $\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{d}$, where we assume $\mathbf{d} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_S)$. As mentioned earlier, the outer-beamforming matrix $\mathbf{V} = [\mathbf{V}_1, \cdots, \mathbf{V}_G]$ is designed based not on the instantaneous CSI but on channel's statistical information $\{\mathbf{U}_g, \Lambda_g\}$, where the $M \times S_g$ submatrix $\mathbf{V}_g$ is the outer-beamforming matrix for group g. Then, the received signal in (3) can be rewritten as [8]

$$\mathbf{y} = \mathbf{HVWd} + \mathbf{n} = \mathbf{GWd} + \mathbf{n}, \tag{4}$$

where $\mathbf{G} := \mathbf{HV}$. Although the MU-MIMO inner-beamformer $\mathbf{W}$ can be designed with full freedom, $\mathbf{W}$ is designed in a block-diagonal form as $\mathbf{W} = \text{diag}(\mathbf{W}_1, \cdots, \mathbf{W}_G)$ under the assumption that inter-group interference is reasonably suppressed by the outer-beamformer $\mathbf{V}$, where $\mathbf{W}_g = [\mathbf{w}_{g_1}, \cdots, \mathbf{w}_{g_{S_g}}]$ is the $S_g \times S_g$ MU-MIMO precoder for group g designed based on the selected CSI from the *effective channel* $\mathbf{G}_g := \mathbf{H}_g\mathbf{V}_g = [\mathbf{g}_{g_1}, \cdots, \mathbf{g}_{g_{K_g}}]^H$ for group g. Hence, in the considered two-stage beamforming scheme, the received signal vector for all users in group g can be expressed from (4) as

$$\mathbf{y}_g = \mathbf{G}_g\mathbf{W}_g\mathbf{d}_g + \sum_{g' \neq g} \mathbf{H}_g\mathbf{V}_{g'}\mathbf{W}_{g'}\mathbf{d}_{g'} + \mathbf{n}_g, \tag{5}$$

where $\mathbf{d}_g$ and $\mathbf{n}_g$ are the data and noise vectors for group g, respectively. For a scheduled user g_k in group g we have the received signal model given by

$$y_{g_k} = \mathbf{g}_{g_k}^H\mathbf{w}_{g_k}d_{g_k} + \sum_{k' \neq k} \mathbf{g}_{g_k}^H\mathbf{w}_{g_{k'}}d_{g_{k'}} + \sum_{g' \neq g} \mathbf{h}_{g_k}^H\mathbf{V}_{g'}\mathbf{W}_{g'}\mathbf{d}_{g'} + n_{g_k}, \tag{6}$$

where $\mathbf{g}_{g_k}$, $\mathbf{w}_{g_k}$, d_{g_k} and n_{g_k} are the $S_g \times 1$ effective channel, $S_g \times 1$ inner-beamforming vector, data and noise symbols of user g_k, respectively. Note that the dimension of the effective channel $\mathbf{g}_{g_k}$ for user g_k is reduced to S_g $(< M)$ and the second and third terms in the right-hand side (RHS) of (6) are the intra-group and inter-group interference, respectively. Concerning the inter-group interference, we assume that at least the approximate block diagonalization (BD) condition in the below holds:

Condition 1 (Inter-group interference condition [8]):

- *Exact BD*: Each group has a sufficient signal space to transmit S_g data streams, that does not interfere with the signal spaces of other groups, i.e.,

$$\dim\left(\text{span}(\mathbf{U}_g) \cap \text{span}^{\perp}(\{\mathbf{U}_{g'} : g' \neq g\})\right) \geq S_g. \tag{7}$$

- *Approximate BD*: When exact BD is impossible, approximate BD can be attained by selecting a matrix $\mathbf{U}_g^*$ consisting of the r_g^* $(< r_g)$ dominant eigenvectors of $\mathbf{R}_g$ such that

$$\dim\left(\text{span}(\mathbf{U}_g^*) \cap \text{span}^{\perp}(\{\mathbf{U}_{g'}^* : g' \neq g\})\right) \geq S_g. \tag{8}$$

Although several sophisticated design algorithms are available for outer-beamformer design for two-stage beamforming [8], [13], [14], we assume $\mathbf{V}_g = \mathbf{U}_g^*$ for the outer-beamformer for analytical tractability for the rest of this paper. In this case, the average transmit power for a scheduled user g_k is given by

$$P_{g_k}^{TX} = \text{tr}(\mathbf{V}_g\mathbf{w}_{g_k}\mathbf{w}_{g_k}^H\mathbf{V}_g^H) = \|\mathbf{w}_{g_k}\|^2 \tag{9}$$

since $d_k \sim \mathcal{CN}(0,1)$ by construction. Note that in the case of approximate BD, r_g^* is a control parameter and the inter-group interference still remains in (6) through the weakest $r_g - r_g^*$ eigenvectors of $\mathbf{R}_g$ not included in $\mathbf{U}_g^*$.

III. The Proposed User Scheduling Method

In this section, we propose a new two-phase-feedback-based user-scheduling-and-beamforming algorithm for the considered two-stage beamforming scheme with given outer-beamformer $\mathbf{V} = [\mathbf{V}_1, \cdots, \mathbf{V}_G]$ and ZFBF for the inner-beamformer $\mathbf{W}_g$. For the sake of simplicity, we assume $S_g = r_g^*$ for all g. We assume that each receiving user g_k (not the BS) estimates its *effective* CSI $\mathbf{g}_{g_k} = \mathbf{V}_g^H\mathbf{h}_{g_k}$.

A. Background: SUS-ZFBF Revisited

To gain some insight into well-performing user selection, we here briefly examine well-known SUS-ZFBF [7] devised under the linear beamforming framework for MU-MIMO downlink. For simplicity, let us just consider one group g only. Under SUS-ZFBF, the BS collects CSI $\mathbf{h}_{g_k}$ from every user g_k in group g in the beginning, and sequentially selects r_g^* users with large channel norms by enforcing semi-orthogonality among the selected users. That is, the BS first selects the user that has the largest channel norm. Let the firstly-selected user's index be $g_{\hat{k}}$. Then, based on the CSI $\mathbf{h}_{g_{\hat{k}}}$, SUS-ZFBF constructs a user-selection hyperslab defined as [7]

$$\mathcal{H}_{g,1} = \left\{\mathbf{f} \in \mathbb{C}^M : \frac{|\mathbf{h}_{g_{\hat{k}}}^H\mathbf{f}|}{||\mathbf{h}_{g_{\hat{k}}}|| \cdot ||\mathbf{f}||} \leq \gamma\right\}. \tag{10}$$

Note that if a vector $\mathbf{f}$ is contained in $\mathcal{H}_{g,1}$, $\mathbf{f}$ is semi-orthogonal to $\mathbf{h}_{g_{\hat{k}}}$. Then, SUS-ZFBF selects the user who has the maximum-length channel vector contained in the hyperslab $\mathcal{H}_{g,1}$. After the second user is selected, another hyperslab is constructed based on the secondly-selected user's channel vector and SUS-ZFBF selects as the third user for service the user that has the maximum-length channel vector contained in the intersection of the first and second hyperslabs. Thus, the third user's channel vector is semi-orthogonal to both the firstly and secondly selected users' channel vectors. In this way, at each step the user with the largest channel norm is selected while semi-orthogonality is maintained among the selected users. Once user selection is done, ZFBF is applied to eliminate inter-user interference due to semi-orthogonality not perfect-orthogonality among the selected users' channels. The effective channel gain loss[1] associated with ZFBF can be made small by making the thickness of the hyperslab (the parameter γ in (10)) small when the number of users is large. (For detail, refer to [7].)

Although SUS-ZFBF yields good performance, it requires CSI from all users in the cell at the BS and this induces

[1] The effective channel gain loss of ZFBF means the loss caused by the ZF constraint: the beamforming vector for the desired user should be contained in the orthogonal complement of the linear space spanned by the interfering user channel vectors. Hence, for the desired user the effective channel is given by the projection of the desired user's channel onto the orthogonal complement of the linear space spanned by the interfering user channel vectors, and the norm of the effective channel is smaller than the original channel norm unless the desired channel is already in the orthogonal complement. The effective channel gain in our case will be given in (22) in the next section.

heavy feedback overhead if SUS-ZFBF is implemented.[2] One way to avoid full CSI feedback for user scheduling is RBF based on beam training and partial CSI (CQI) feedback [6]. In this method, the BS just randomly determines a set of orthonormal beam vectors and then transmits each beam sequentially in time during the training period. During the training period, each user computes the SINR of each beam direction, assuming that the training beam vectors will be used again as the data-transmitting beam vectors later. Then, each user feeds back its maximum SINR value and the corresponding beam index. After the feedback period is finished, the BS selects one user for each beam such that the selected user for a beam has the maximum SINR for the considered beam among the users who reported that the considered beam is its best beam. Thus, RBF requires only CQI of a SINR value and a beam index from each user and reduces feedback overhead significantly compared to SUS-ZFBF. However, the performance RBF falls quite short of that of SUS-ZFBF in the finite-user case. There have been several corrections of RBF to improve its performance by using two-phase feedback [20], [21]. In the next subsection, we propose a new two-phase-feedback-based scheduling-and-beamforming method almost achieving the performance of SUS-ZFBF in the context of two-stage beamforming.

B. The Proposed User Selection and Beamforming Method

The key aspect of the newly-proposed two-phase-feedback-based scheduling-and-beamforming algorithm is to implement the essence of user scheduling, i.e., selection of semi-orthogonal users with large channel norms in a distributed manner without full CSI feedback from all users. Like RBF, we use a set of orthogonal reference beams and use $\mathbf{u}_{g,1}, \mathbf{u}_{g,2}, \cdots, \mathbf{u}_{g,r_g^*}$ as the orthogonal reference beam directions for each group g in the considered two-stage beamforming. To enforce semi-orthogonality among the selected users, we construct a *cone* $\mathcal{C}_{g,i}$ around each reference beam i, as shown in Fig. 2, given by

$$\mathcal{C}_{g,i} = \left\{ \mathbf{h} : \frac{|\mathbf{h}^H \mathbf{u}_{g,i}|}{||\mathbf{h}||} \geq \alpha' \right\}, \quad i = 1, 2, \cdots, r_g^*. \tag{11}$$

During the first-phase training period, each user g_k checks if its channel vector $\mathbf{h}_{g_k}$ is contained in $\mathcal{C}_{g,i}$ for each i. Note that this checking is distributed and done at users not at the BS. To check $\mathbf{h}_{g_k} \in \mathcal{C}_{g,i}$, each user g_k needs the original CSI $\mathbf{h}_{g_k}$. However, we are assuming that only the equivalent CSI $\mathbf{g}_{g_k}$ is available at user g_k under two-stage beamforming. From (4), (5), and (6), we have

$$\mathbf{g}_{g_k}^H = \mathbf{h}_{g_k}^H \mathbf{V}_g = \mathbf{h}_{g_k}^H \mathbf{U}_g^* = [\mathbf{h}_{g_k}^H \mathbf{u}_{g,1}, \cdots, \mathbf{h}_{g_k}^H \mathbf{u}_{g,r_g^*}]. \tag{12}$$

Hence, the cone-containment checking can be done simply by computing the normalized vector $\frac{\mathbf{g}_{g_k}^H}{||\mathbf{g}_{g_k}||}$ and checking if the absolute value of each element of $\frac{\mathbf{g}_{g_k}^H}{||\mathbf{g}_{g_k}||}$ is larger than or equal to a new threshold $\alpha \overset{\Delta}{=} \alpha' \frac{||\mathbf{h}_{g_k}||}{||\mathbf{g}_{g_k}||}$, since $\frac{\mathbf{g}_{g_k}^H}{||\mathbf{g}_{g_k}||} = \frac{||\mathbf{h}_{g_k}||}{||\mathbf{g}_{g_k}||} \left[\frac{\mathbf{h}_{g_k}^H \mathbf{u}_{g,1}}{||\mathbf{h}_{g_k}||}, \cdots, \frac{\mathbf{h}_{g_k}^H \mathbf{u}_{g,r_g^*}}{||\mathbf{h}_{g_k}||} \right]$. Note that the new threshold is bounded as $0 \leq \alpha \leq 1$ since each element of the normalized vector $\mathbf{g}_{g_k}^H / ||\mathbf{g}_{g_k}||$ is compared with α, and α is a system design parameter that controls the size of each cone and thus the semi-orthogonality of the selected user channels. If the channel $\mathbf{h}_{g_k}$ of user g_k is contained in cone i, user g_k's channel is well aligned with the reference beam i and user g_k belongs to the i-th candidate set. Two different users contained in two different candidate sets have semi-orthogonal channel vectors since the two corresponding cones are constructed around two different orthogonal reference beam vectors.

[2]For implementation of SUS, SUS with limited feedback was studied in [22].

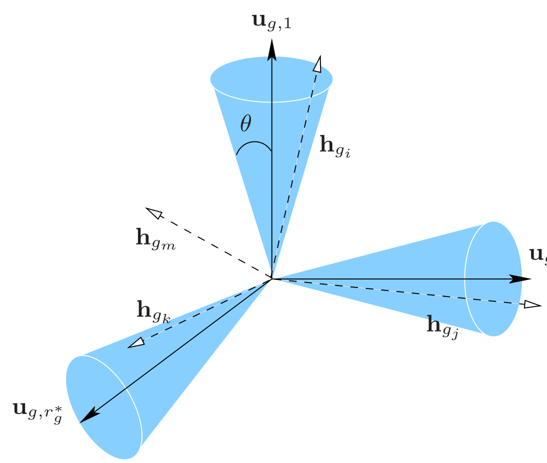


Fig. 2. User selection cones.

Now, the question is "which channel vector in $\mathcal{C}_{g,i}$ should be selected?" Since the semi-orthogonality of users to be selected is already guaranteed by the user-selection cones, we choose the user in $\mathcal{C}_{g,i}$ that has the maximum channel norm, following the spirit of SUS-ZFBF. This answers what the CQI for feedback should be. In the two-stage beamforming setting with the assumption of the availability of the effective channel $\mathbf{g}_{g_k}$, we just use $||\mathbf{g}_{g_k}||^2$ since

$$||\mathbf{g}_{g_k}||^2 = ||\mathbf{h}_{g_k}^H \mathbf{U}_g^*||^2 = ||(\sum_{j=1}^{r_g} c_{g_k}^j \mathbf{u}_{g,j})^H \mathbf{U}_g^*||^2$$
$$= \sum_{j=1}^{r_g^*} |c_{g_k}^j|^2 \overset{(a)}{\approx} \sum_{j=1}^{r_g} |c_{g_k}^j|^2 = ||\mathbf{h}_{g_k}||^2,$$

where $\mathbf{h}_{g_k} = \sum_{j=1}^{r_g} c_{g_k}^j \mathbf{u}_{g,j}$ with some coefficients $c_{g_k}^j$ by (1), and step (a) is valid because the most dominant r_g^* eigenvectors are included. In this way, we can select a set of semi-orthogonal users with large channel norms. Once the user selection is done, we apply ZFBF with water-filling power allocation based on the effective CSI obtained from the selected users in the second-phase feedback to completely remove inter-user interference resulting from the semi-orthgonality of the selected users' channels. In general, the performance of ZFBF degrades due to the effective channel gain loss. However, this effective channel gain loss is managed by the semi-orthogonality of the selected users controlled by the parameter α, as we shall see in Section IV.

For further improvement in the two-stage beamforming setting, we can take the inter-group interference into consideration. With $\mathbf{V}_g = \mathbf{U}_g^*$ for all g and equal power $\rho = \frac{P}{\sum_{g=1}^{G} r_g^*}$ for every scheduled user, the norm of every column of the second-stage beamformer $\mathbf{W}_g$ is ρ from (9). From (6), the average power of the inter-group interference-plus-noise is upper bounded by

$$1 + r_g^* \rho \sum_{g' \neq g} \|\mathbf{h}_{g_k}^H \mathbf{V}_{g'}\|^2 \tag{13}$$

due to norm's submultiplicativity, $\|\cdot\| \leq \|\cdot\|_F$, and $\|\mathbf{W}_{g'}\|_F = r_g^* \rho$. We define a *quasi-SINR* of user g_k as

$$Q(g_k) := \frac{\|\mathbf{g}_{g_k}\|^2}{\frac{1}{\rho} + r_g^* \sum_{g' \neq g} \|\mathbf{h}_{g_k}^H \mathbf{V}_{g'}\|^2}. \tag{14}$$

Note that the quasi-SINR is not a true SINR for any reference beam $\mathbf{u}_{g,i}$. In the definition, the numerator is simply the effective channel norm square, and the intra-group interference is not included in the denominator because ZFBF is used for $\mathbf{W}_g$. Without the inter-group interference, the quasi-SINR is simply a scaled version of the effective channel norm square.

Remark 1: The reason for the chosen definition of the quasi-SINR will become clear in Section IV. This metric guarantees the asymptotic optimality of the proposed method under the assumption of the approximated BD in Condition 1 for inter-group interference, i.e., the proposed method using the quasi-SINR has the same asymptotic behavior as DPC as $K \to \infty$, as shown in Theorem 2.

Remark 2: (Effective Channel Estimation and Interference-Plus-Noise Power Estimation): The effective CSI and the average inter-group-interference-plus-noise power can easily be estimated at users during the downlink training period based on (12). From (5) the received signal model at user g_k can be rewritten as

$$y_{g_k} = \underbrace{\mathbf{h}_{g_k}^H \mathbf{V}_g}_{\mathbf{g}_{g_k}^H} \mathbf{W}_g \mathbf{d}_g + \sum_{g' \neq g} \mathbf{h}_{g_k}^H \mathbf{V}_{g'} \mathbf{W}_{g'} \mathbf{d}_{g'} + n_{g_k}. \tag{15}$$

First, the effective CSI $\mathbf{g}_{g_k}$ can easily be estimated at users based on (12) by sequentially transmitting $\mathbf{u}_{g,1}, \cdots, \mathbf{u}_{g,r_g^*}$ for each group g during the downlink training period.[3] Second, during the downlink training period, the average inter-group interference-plus-noise power $1 + \rho \sum_{g' \neq g} \|\mathbf{h}_{g_k}^H \mathbf{V}_{g'}\|^2$ can also be estimated easily based on (15). That is, $\mathbf{W}_g$ and $\mathbf{d}_g$ used for training are known and stored in all users in group g. Once an estimate $\hat{\mathbf{g}}_{g_k}$ for the effective channel is obtained at user g_k, user g_k constructs $\hat{\mathbf{g}}_{g_k}^H \mathbf{W}_g \mathbf{d}_g$, computes $y_{g_k} - \hat{\mathbf{g}}_{g_k}^H \mathbf{W}_g \mathbf{d}_g \approx \sum_{g' \neq g} \mathbf{h}_{g_k}^H \mathbf{V}_{g'} \mathbf{W}_{g'} \mathbf{d}_{g'} + n_{g_k}$, squares $y_{g_k} - \hat{\mathbf{g}}_{g_k}^H \mathbf{W}_g \mathbf{d}_g$, and averages the result over a few training symbol times to obtain the desired value. If the training $\mathbf{W}_{g'}$ and the actual data-transmitting $\mathbf{W}_{g'}$ have similar norms, the estimated average inter-group interference-plus-noise power will be valid for the data-transmission period. Thus, users can easily compute the proposed quasi-SINR during the training period.

[3]For more detail on channel estimation, please see [9].

Summarizing the above, we present our proposed two-phase feedback-based user-scheduling-and-beamforming algorithm named *REference-based Distributed (semi-)Orthogonal user Selection with Post-selection Beam Refinement (ReDOS-PBR)*:

Algorithm 1 (ReDOS-PBR):

0) $\alpha \in (0, 1)$ is a pre-determined parameter and is shared by the BS and all users. The BS initializes the candidate set $\mathcal{W}_{g,i}$ associated with cone i for group g and the selected user set $\mathcal{S}_g$ for group g: $\mathcal{W}_{g,i} = \emptyset$, for $i = 1, \cdots, r_g^*$ and $\mathcal{S}_g = \emptyset$. Every user g_k estimates $\mathbf{g}_{g_k}$ and $1 + \rho \sum_{g' \neq g} \|\mathbf{h}_{g_k}^H \mathbf{V}_{g'}\|^2$ during the training period in which $\mathbf{u}_{g,1}, \cdots, \mathbf{u}_{g,r_g^*}$ are sequentially transmitted for each group g.
1) Each user g_k independently computes the following set:
$$\mathcal{I}_{g_k} := \left\{ i : \left| (\mathbf{e}_i^{(g)})^T \frac{\mathbf{g}_{g_k}}{\|\mathbf{g}_{g_k}\|} \right| \geq \alpha, \ i = 1, \cdots, r_g^* \right\}, \tag{16}$$
where $\mathbf{e}_i^{(g)}$ is the i-th column of $\mathbf{I}_{r_g^*}$. ($(\mathbf{e}_i^{(g)})^T \frac{\mathbf{g}_{g_k}}{\|\mathbf{g}_{g_k}\|}$ is simply the i-th element of $\frac{\mathbf{g}_{g_k}^H}{\|\mathbf{g}_{g_k}\|} = \frac{\|\mathbf{h}_{g_k}\|}{\|\mathbf{g}_{g_k}\|} \left[\frac{\mathbf{h}_{g_k}^H \mathbf{u}_{g,1}}{\|\mathbf{h}_{g_k}\|}, \cdots, \frac{\mathbf{h}_{g_k}^H \mathbf{u}_{g,r_g^*}}{\|\mathbf{h}_{g_k}\|} \right]$.)
If user g_k has $\mathcal{I}_{g_k} \neq \emptyset$, then user g_k finds $i_{g_k}^* = \arg\max_{i \in \mathcal{I}_{g_k}} \left| (\mathbf{e}_i^{(g)})^T \frac{\mathbf{g}_{g_k}}{\|\mathbf{g}_{g_k}\|} \right|$ and feedbacks the CQI pair $(i_{g_k}^*, Q(g_k))$ to the BS in the first-phase feedback. If $\mathcal{I}_{g_k} = \emptyset$, user g_k does not feedback. After the feedback, the BS updates $\mathcal{W}_{g,i_{g_k}^*} \leftarrow \mathcal{W}_{g,i_{g_k}^*} \cup \{g_k\}$ and stores $Q(g_k)$.
2) For $i = 1, \cdots, r_g^*$, the BS finds
$$\kappa_{g,i} = \arg\max_{g_k \in \mathcal{W}_{g,i}} Q(g_k), \tag{17}$$
and updates $\mathcal{S}_g \leftarrow \mathcal{S}_g \cup \{\kappa_{g,i}\}$.
3) The BS transmits a paging signal to notify that the users in $\mathcal{S}_g$ are scheduled and then, only the corresponding scheduled users feedback their effective CSI to the BS in the second-phase feedback. Finally, the BS constructs the MU-MIMO ZFBF inner-beamformer with water-filling power allocation for each group based on the signal model (5) and the acquired effective CSI from the scheduled users, and transmits data to the scheduled users.

In step 1), each user checks if its channel vector is contained in each of the user-selection cones. If user g_k has a non-empty set $\mathcal{I}_{g_k}$, then user g_k finds the reference beam that has the largest channel component and feedbacks the corresponding reference cone index $i_{g_k}^*$ and the quasi-SINR $Q(g_k)$ to the BS. If $\mathcal{I}_{g_k} = \emptyset$, then user g_k does not feedback any information to the BS. After the first-phase feedback period is over, the BS makes r_g^* candidate sets $\mathcal{W}_{g,1}, \cdots, \mathcal{W}_{g,r_g^*}$ associated with r_g^* cones for group g, based on the feedbacked CQI information. Here, $\mathcal{W}_{g,i}$ represents the set of users whose channels are contained in the user-selection cone around the i-th reference beam. In step 2), the BS chooses the user $\kappa_{g,i}$ having the largest quasi-SINR in each candidate set $\mathcal{W}_{g,i}$, $i = 1, \cdots, r_g^*$, to construct the set $\mathcal{S}_g$ of scheduled users for each group

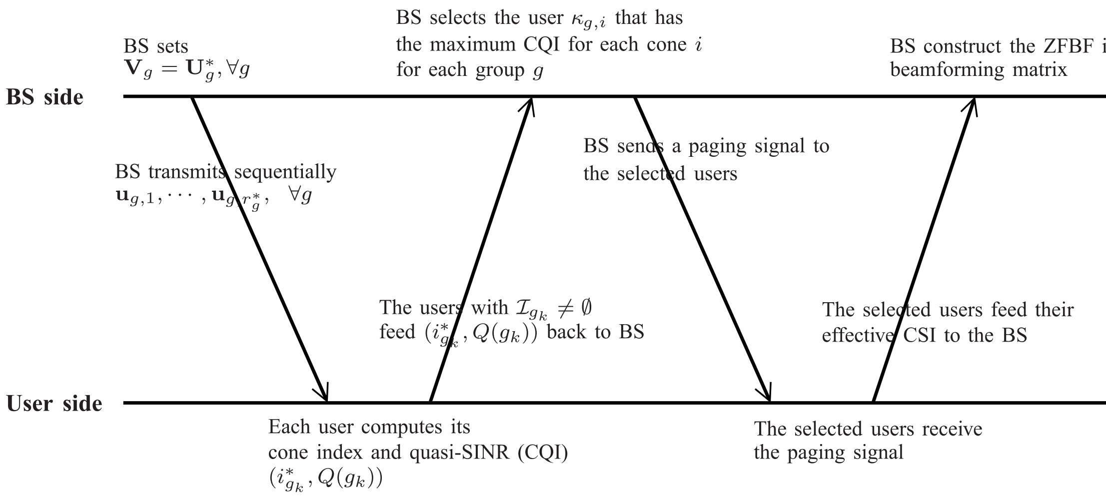


Fig. 3. The flow of Algorithm 1.

g. In step 3), ZFBF is used for the scheduled users. Here, more sophisticated MU-MIMO BF like MMSE BF can also be used for the post-user-selection beam refinement to yield better performance, if additional information of inter-group interference and noise variance is available at the BS for the signal model (5). However, since the semi-orthogonality among the selected users for each group and the approximated BD condition for inter-group interference are satisfied, ZFBF is a reasonable choice for beamforming. In this case, the effective channel gain loss by ZFBF is not significant since the selected user channels are semi-orthogonal. Algorithm 1 is summarized in Fig. 3.

Remark 3 (Amount of Feedback): First note that in ReDOS-PBR, user selection is done based on only CQI feedback from possibly all users in the first-phase feedback and post-user-selection beam refinement is done based on the CSI feedback from only the scheduled users in the second-phase feedback. The feedback difference in CQI and CSI is significant in MIMO systems. The amount of feedback required for the proposed method for group g for one scheduling interval is $\sum_{i=1}^{r_g^*} |\mathcal{W}_{g,i}|$ integers for user beam index feedback, $\sum_{i=1}^{r_g^*} |\mathcal{W}_{g,i}|$ real numbers for quasi-SINR feedback, and $2(r_g^*)^2$ real numbers for later effective CSI feedback because only r_g^* users per group need to feedback their effective CSI $\mathbf{g}_{\kappa_{g,i}}$ of complex dimension r_g^* for $\mathbf{V}_g = \mathbf{U}_g^*$. As shown in Lemma 1 in the below, when $\alpha \leq 1/\sqrt{r_g^*}$, $\mathcal{I}_{g_k}$ is a non-empty set for all g_k and thus, every user feedbacks its quasi-SINR to the BS. Hence, in this case, $\sum_{i=1}^{r_g^*} |\mathcal{W}_{g,i}|$ reduces to K_g. When $\alpha > 1/\sqrt{r_g^*}$, on the other hand, $\mathcal{I}_{g_k} = \emptyset$ for some users[4] and thus in this case, $\sum_{i=1}^{r_g^*} |\mathcal{W}_{g,i}|$ can be less than K_g. In Section VI, numerical results show that many users do not feedback even CQI to the BS for optimally chosen α as in RBF with thresholding [6] and the feedback overhead is reduced drastically.

[4]In the case of $\alpha > 1/\sqrt{r_g^*}$. the probability of $\mathcal{I}_{g_k} = \emptyset$ is given by $\Pr\{\mathcal{I}_{g_k} = \emptyset\} = 1 - r_g^* \left[1 - F_{2,2(r_g^*-1)}\left(\frac{(r_g^*-1)\alpha^2}{(1-\alpha^2)}\right)\right]$ under the assumption $\mathbf{g}_{g_k} \sim \mathcal{CN}(0, \mathbf{I}_{r_g^*})$, where $F_{n,m}(x)$ is the cumulative distribution function (CDF) of the F distribution with parameters n and m.

Lemma 1: When $\alpha \leq \alpha_{min} := 1/\sqrt{r_g^*}$, $\mathcal{I}_{g_k}$ is a non-empty set for all g_k.

Proof: Suppose that there is a user g_k such that $\mathcal{I}_{g_k} = \emptyset$. Then, $\left|(\mathbf{e}_i^{(g)})^T \frac{\mathbf{g}_{g_k}}{\|\mathbf{g}_{g_k}\|}\right| < 1/\sqrt{r_g^*}$ for all $i = 1, \cdots, r_g^*$. Therefore, we have $1 = \left\|\frac{\mathbf{g}_{g_k}}{\|\mathbf{g}_{g_k}\|}\right\|^2 = \sum_{i=1}^{r_g^*} \left|(\mathbf{e}_i^{(g)})^T \frac{\mathbf{g}_{g_k}}{\|\mathbf{g}_{g_k}\|}\right|^2 < 1$, which is contradiction. Hence, the claim follows. ■

Remark 4: Lemma 1 implies that $\bigcup_{i=1}^{r_g^*} \mathcal{C}_{g,i} \supset \mathbb{C}^{r_g*}$ for $\alpha \leq \alpha_{min} = 1/\sqrt{r_g^*}$. On the other hand, when $\alpha > \frac{1}{\sqrt{2}}$, $\mathcal{C}_{g,i} \cap \mathcal{C}_{g,j} = \emptyset$ for $i \neq j$, because the angle θ in Fig. 2 is $\pi/4$ when $\alpha = 1/\sqrt{2}$.

IV. Optimality of the Proposed Method

In this section, we prove the asymptotical optimality of the proposed method as $K \to \infty$. The sum capacity scaling law of a K-user MIMO broadcast channel consisting of multiple groups with each group's having the same channel covariance matrix is provided in [8]:

Theorem 1 [8]: In a MU-MIMO downlink system composed of a BS with M transmit antennas and total power constraint P and K single-antenna users divided into G groups of equal size $K' = K/G = K_g$, where the channel vector of each user in group g is independent and identically distributed (i.i.d.) from $\mathcal{CN}(\mathbf{0}, \mathbf{R}_g)$ for $g = 1, \cdots, G$, the sum capacity (which is achieved by DPC) scales as $R_{DPC} = \beta \log\log(K') + \beta \log \frac{P}{\beta} + O(1)$, where $\beta = \min\{M, \sum_{g=1}^{G} r_g\}$ and $O(1)$ denotes a bounded constant independent of K', as $K' \to \infty$.

The same scaling law is achieved by ReDOS-PBR under the approximate BD condition in Condition 1, as shown in the following theorem.

Theorem 2: Under the condition described in Theorem 1, the sum rate of the scheduled user sets $\{\mathcal{S}_g\}$ by ReDOS-PBR scales as $\mathbb{E}\left[\sum_{g=1}^{G} R_{ZF,g}(\mathcal{S}_g)\right] \sim R_{DPC}$, where R_{DPC} is given in Theorem 1 and $x \sim y$ indicates that

$\lim_{K' \to \infty} x/y = 1$. Here, $R_{ZF,g}(\mathcal{S}_g)$ is the sum rate of the users in $\mathcal{S}_g$ determined by ReDOS-PBR.

Proof: Our proof of the asymptotic optimality of ReDOS-PBR is by first showing that the effective channel gain associated with ReDOS-PBR is bounded below away from zero for some *fixed* α strictly less than one and then showing that the multi-user diversity gain reduction associated with ReDOS-PBR for that fixed α become negligible as $K' \to \infty$. (We borrowed the flow of our proof from [7]. However, different techniques and ideas are used for our proof of the asymptotic optimality of ReDOS-PBR.)

From (5), we have the received signal model for the scheduled users in $\mathcal{S}_g$ as

$$\mathbf{y}_g(\mathcal{S}_g) = \mathbf{G}_g(\mathcal{S}_g)\mathbf{W}_g\mathbf{d}_g + \sum_{g' \neq g} \mathbf{H}_g(\mathcal{S}_g)\mathbf{V}_{g'}\mathbf{W}_{g'}\mathbf{d}_{g'} + \mathbf{n}_g(\mathcal{S}_g), \tag{18}$$

where $\mathbf{d}_g = [d_{\kappa_{g,1}}, \cdots, d_{\kappa_{g,r_g^*}}]$, $\mathbf{W}_g = [\mathbf{w}_{\kappa_{g,1}}, \cdots, \mathbf{w}_{\kappa_{g,r_g^*}}]$, $\mathbf{G}_g(\mathcal{S}_g) = [\{\mathbf{g}_{gk}\}_{k \in \mathcal{S}_g}]^H = [\mathbf{g}_{\kappa_{g,1}}, \cdots, \mathbf{g}_{\kappa_{g,r_g^*}}]^H$ is a submatrix of $\mathbf{G}_g$ composed of the rows corresponding to $\mathcal{S}_g$, and $\mathbf{H}_g(\mathcal{S}_g) = [\{\mathbf{h}_{gk}\}_{k \in \mathcal{S}_g}]^H$.

i) Lower bound on the effective channel gain: Since ZFBF is assumed for the inner beamforming with the signal model (18), we have

$$\begin{aligned} \mathbf{W}_g &= [\mathbf{w}_{\kappa_{g,1}}, \cdots, \mathbf{w}_{\kappa_{g,r_g^*}}] \\ &= \mathbf{G}_g^H(\mathcal{S}_g)\left[\mathbf{G}_g(\mathcal{S}_g)\mathbf{G}_g^H(\mathcal{S}_g)\right]^{-1}\mathbf{P}_g \\ &=: \tilde{\mathbf{W}}_g\mathbf{P}_g = [\tilde{\mathbf{w}}_{\kappa_{g,1}}, \cdots, \tilde{\mathbf{w}}_{\kappa_{g,r_g^*}}]\mathbf{P}_g, \end{aligned} \tag{19}$$

where $\mathbf{P}_g = \mathrm{diag}(\sqrt{P_{\kappa_{g,1}}}, \cdots, \sqrt{P_{\kappa_{g,r_g^*}}})$, and $P_{\kappa_{g,i}}$ is the transmit power scaling factor for the scheduled user $\kappa_{g,i} \in \mathcal{S}_g$. Substituting the above ZF $\mathbf{w}_{\kappa_{g,1}}, \cdots, \mathbf{w}_{\kappa_{g,r_g^*}}$ into the received signal model (6) of user $\kappa_{g,i}$ yields

$$y_{\kappa_{g,i}} = \sqrt{P_{\kappa_{g,i}}}d_{\kappa_{g,i}} + \sum_{g' \neq g} \mathbf{h}_{\kappa_{g,i}}^H\mathbf{V}_{g'}\mathbf{W}_{g'}\mathbf{d}_{g'} + n_{\kappa_{g,i}} \tag{20}$$

for $i = 1, \cdots, r_g^*$, since $\mathbf{G}_g(\mathcal{S}_g)\mathbf{W}_g = \mathbf{P}_g = \mathrm{diag}(\sqrt{P_{\kappa_{g,1}}}, \cdots, \sqrt{P_{\kappa_{g,r_g^*}}})$. From (20), the sum rate of the ZF MU-MIMO broadcast channel consisting of users $\{\kappa_{g,1}, \cdots, \kappa_{g,r_g^*}\}$ with power scaling $\{P_{\kappa_{g,1}}, \cdots, P_{\kappa_{g,r_g^*}}\}$ is given by [23]

$$\begin{aligned} &R_{ZF,g}(\mathcal{S}_g) \\ &= \max_{\{P_{\kappa_{g,i}}\}} \sum_{i=1}^{r_g^*} \log\left(1 + \frac{P_{\kappa_{g,i}}}{1 + \sum_{g' \neq g} \|\mathbf{h}_{\kappa_{g,i}}^H\mathbf{V}_{g'}\mathbf{W}_{g'}\|^2}\right) \\ &\quad \text{s.t.} \sum_{i=1}^{r_g^*} \gamma_{\kappa_{g,i}}^{-1}P_{\kappa_{g,i}} \leq r_g^*\rho, \end{aligned} \tag{21}$$

where the effective channel gain $\gamma_{\kappa_{g,i}}$ for user $\kappa_{g,i}$ is given by [7], [24]

$$\gamma_{\kappa_{g,i}} = \frac{1}{[(\mathbf{G}_g(\mathcal{S}_g)\mathbf{G}_g(\mathcal{S}_g)^H)^{-1}]_{i,i}}. \tag{22}$$

(Here, we assume that the total transmit power assigned to group g is proportional to the number of the scheduled users in group g, and hence is given by $r_g^*\rho$. (Recall that $\rho = P/\sum_{g=1}^G r_g^*$.) By (9) the actual power assigned to user $\kappa_{g,i}$ is given by $P_{\kappa_{g,i}}^{TX} = \|\mathbf{w}_{\kappa_{g,i}}\|^2 = \gamma_{\kappa_{g,i}}^{-1}P_{\kappa_{g,i}}$ because $\|\mathbf{w}_{\kappa_{g,i}}\|^2 = \|\tilde{\mathbf{w}}_{\kappa_{g,i}}\|^2P_{\kappa_{g,i}}$ and $\|\tilde{\mathbf{w}}_{\kappa_{g,i}}\|^2 = [\tilde{\mathbf{W}}_g^H\tilde{\mathbf{W}}_g]_{i,i} = [(\mathbf{G}_g(\mathcal{S}_g)\mathbf{G}_g(\mathcal{S}_g)^H)^{-1}]_{i,i}$ from (19). Hence, the power constraint in (21) follows. Note that the ZF loss appears as the shrinkage of the feasible region of $(P_{\kappa_{g,1}}, \cdots, P_{\kappa_{g,1}})$ in (21).)

Now consider the denominator term in the RHS of (22). Since $[\mathbf{G}_g(\mathcal{S}_g)\mathbf{G}_g(\mathcal{S}_g)^H]_{i,j} = \mathbf{g}_{\kappa_{g,i}}^H\mathbf{g}_{\kappa_{g,j}}$, $\forall i, j$, it can be decomposed as

$$\mathbf{G}_g(\mathcal{S}_g)\mathbf{G}_g(\mathcal{S}_g)^H = \mathbf{D}\tilde{\mathbf{G}}\mathbf{D}, \tag{23}$$

where $\mathbf{D} = \mathrm{diag}(\|\mathbf{g}_{\kappa_{g,1}}\|, \cdots, \|\mathbf{g}_{\kappa_{g,r_g^*}}\|)$ and $\tilde{\mathbf{G}} =$

$$\begin{bmatrix} 1 & \tilde{\mathbf{g}}_{\kappa_{g,1}}^H\tilde{\mathbf{g}}_{\kappa_{g,2}} & \cdots & \tilde{\mathbf{g}}_{\kappa_{g,1}}^H\tilde{\mathbf{g}}_{\kappa_{g,r_g^*}} \\ \tilde{\mathbf{g}}_{\kappa_{g,2}}^H\tilde{\mathbf{g}}_{\kappa_{g,1}} & 1 & & \vdots \\ \vdots & & \ddots & \tilde{\mathbf{g}}_{\kappa_{g,r_g^*-1}}^H\tilde{\mathbf{g}}_{\kappa_{g,r_g^*}} \\ \tilde{\mathbf{g}}_{\kappa_{g,r_g^*}}^H\tilde{\mathbf{g}}_{\kappa_{g,1}} & \cdots & \tilde{\mathbf{g}}_{\kappa_{g,r_g^*}}^H\tilde{\mathbf{g}}_{\kappa_{g,r_g^*-1}} & 1 \end{bmatrix} \tag{24}$$

with $\tilde{\mathbf{g}}_{\kappa_{g,i}} = \frac{\mathbf{g}_{\kappa_{g,i}}}{\|\mathbf{g}_{\kappa_{g,i}}\|}$, $\forall i$. Substituting (23) into (22), we have

$$\begin{aligned} \gamma_{\kappa_{g,i}} &= \frac{1}{[(\mathbf{G}_g(\mathcal{S}_g)\mathbf{G}_g(\mathcal{S}_g)^H)^{-1}]_{i,i}} = \frac{1}{[\mathbf{D}^{-1}\tilde{\mathbf{G}}^{-1}\mathbf{D}^{-1}]_{i,i}} \\ &= \frac{\|\mathbf{g}_{\kappa_{g,i}}\|^2}{[\tilde{\mathbf{G}}^{-1}]_{i,i}}. \end{aligned} \tag{25}$$

Consider the term $[\tilde{\mathbf{G}}^{-1}]_{i,i}$ in (25). By Lemma 2 in Appendix, we have

$$|\tilde{\mathbf{g}}_{\kappa_{g,i}}^H\tilde{\mathbf{g}}_{\kappa_{g,j}}| \leq 2\alpha\sqrt{1-\alpha^2} \quad \text{for } i \neq j \tag{26}$$

when $\alpha \geq 1/\sqrt{2}$. By the Gershgorin circle theorem [25] and (26), every eigenvalue of the Hermitian matrix $\tilde{\mathbf{G}}$ is in a Gershgorin disk,[5] i.e.,

$$\begin{aligned} \lambda(\tilde{\mathbf{G}}) \subset &\{z \in \mathbb{R}^+ : |z-1| \leq (r_g^*-1)2\alpha\sqrt{1-\alpha^2}\}, \\ = &\{z \in \mathbb{R}^+ : 1-(r_g^*-1)2\alpha\sqrt{1-\alpha^2} \leq z \\ &\leq 1+(r_g^*-1)2\alpha\sqrt{1-\alpha^2}\}, \end{aligned} \tag{27}$$

where $\lambda(\tilde{\mathbf{G}})$ is the set of eigenvalues of $\tilde{\mathbf{G}}$. When $(r_g^*-1)2\alpha\sqrt{1-\alpha^2} < 1$, equivalently, $\alpha > \sqrt{\frac{1+\sqrt{\frac{r_g^*-2}{r_g^*-1}}}{2}}$, we have a non-trivial lower bound on the minimum eigenvalue $\lambda_{min}(\tilde{\mathbf{G}})$ of $\tilde{\mathbf{G}}$ and

$$[\tilde{\mathbf{G}}^{-1}]_{i,i} \overset{(a)}{\leq} [\lambda_{min}(\tilde{\mathbf{G}})]^{-1} \overset{(b)}{\leq} \frac{1}{1-(r_g^*-1)2\alpha\sqrt{1-\alpha^2}}, \tag{28}$$

where (a) follows since $\tilde{\mathbf{G}}$ is self-adjoint and (b) follows from (27). Thus, from (25) and (28), the effective channel

[5] All Gershgorin disks of $\tilde{\mathbf{G}}$ have the same center of one and the same radius upper bound. So, we can use any of the Gershgorin disks of $\tilde{\mathbf{G}}$.

gain $\gamma_{\kappa_{g,i}}$ is lower bounded by

$$\gamma_{\kappa_{g,i}} \geq \frac{\|\mathbf{g}_{\kappa_{g,i}}\|^2}{\frac{1}{1-(r_g^*-1)2\alpha\sqrt{1-\alpha^2}}} = \|\mathbf{g}_{\kappa_{g,i}}\|^2(1-(r_g^*-1)2\alpha\sqrt{1-\alpha^2}). \tag{29}$$

Note that the derived lower bound (29) on the effective channel gain is valid for any *fixed* α satisfying $\alpha > \sqrt{\frac{1+\sqrt{\frac{r_g^*-2}{r_g^*-1}}}{2}} \overset{(a)}{\geq} \frac{1}{\sqrt{2}}$, where (a) for the validity of (26) is valid for any $r_g^* \geq 2$. By making $\alpha \uparrow 1$, we can completely eliminate the ZFBF loss. However, $\alpha \uparrow 1$ will lose the multiuser diversity gain. So, we fix α to an arbitrary number $\bar{\alpha}$ strictly less than one, independent of K' such that

$$\bar{\alpha} \in \left(\sqrt{\frac{1+\sqrt{\frac{r_g^*-2}{r_g^*-1}}}{2}}, \quad 1\right). \tag{30}$$

ii) Multi-user diversity gain: There are several difficult points in handling the multi-user diversity gain of ReDOS-PBR in the multi-group setting of two-stage beamforming. The first point is that only users whose channel vectors are contained in one of the user-selection cones report quasi-SINR and the second point is that we should handle the inter-group interference properly. Despite such difficulty we were able to show that the multi-user diversity gain is still preserved for ReDOS-PBR under the approximate BD condition.

As in [7], the first difficulty mentioned above can be handled by defining

$$\phi_{g_k}^i = \begin{cases} Q(g_k), & g_k \in \mathcal{W}_{g,i}, \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

for all users g_k, $k = 1\cdots, K_g = K'$ in group g. Then, for a given i, the random variable $\phi_{g_k}^i$ is i.i.d. across k in the same group g since $\mathbf{h}_{g_k} \overset{i.i.d.}{\sim} \mathcal{CN}(\mathbf{0}, \mathbf{R}_g)$. Note for the step (17) that

$$\kappa_{g,i} = \underset{g_k \in \mathcal{W}_{g,i}}{\arg\max}\, Q(g_k) = \underset{g_k, k\in\{1,\cdots,K_g=K'\}}{\arg\max}\, \phi_{g_k}^i. \tag{32}$$

The multi-user diversity gain results from choosing the best user among all users with i.i.d. channel realizations. However, with ReDOS-PBR, for each data stream, the best user within $\mathcal{W}_{g,i}$ is chosen, and thus there exists some loss in the multi-user diversity gain. However, based on extreme value theory, it can be shown that for each i

$$\Pr\{\phi_{\kappa_{g,i}}^i > u_g^i\} \geq 1 - O(1/K'), \tag{33}$$

for ReDOS-PBR under the approximate BD condition in Condition 1, where

$$u_g^i = (\lambda_{g,1}\log K' - \lambda_{g,1}\log\log K' + a_i)/(1/\rho + \epsilon). \tag{34}$$

Here, $\lambda_{g,1}$ is the maximum eigenvalue of $\mathbf{R}_g$ (see (2)) and a_i and ϵ are constants independent of K'. (For detailed proof of (33) and (34), please see Appendices B and C in [26].)

iii) Finally, we show the asymptotic optimality of ReDOS-PBR based on *i)* and *ii)*. Fix α as $\bar{\alpha}$ in (30). Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{g=1}^{G} R_{ZF,g}(\mathcal{S}_g)\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[\sum_{g=1}^{G}\sum_{i=1}^{r_g^*}\log(1+\frac{\rho\gamma_{\kappa_{g,i}}}{1+\sum_{g'\neq g}\|\mathbf{h}_{\kappa_{g,i}}^H\mathbf{V}_{g'}\mathbf{W}_{g'}\|^2})\right] \\
&\overset{(b)}{\geq} \mathbb{E}\left[\sum_{g=1}^{G}\sum_{i=1}^{r_g^\star}\log(1+\frac{\|\mathbf{g}_{\kappa_{g,i}}\|^2[1-(r_g^\star-1)2\alpha\sqrt{1-\alpha^2}]}{\frac{1}{\rho}+r_g^*\sum_{g'\neq g}\|\mathbf{h}_{\kappa_{g,i}}^H\mathbf{V}_{g'}\|^2})\right] \\
&\overset{(c)}{\geq} \sum_{g=1}^{G}\sum_{i=1}^{r_g^\star}\Pr\{\phi_{\kappa_{g,i}}^i > u_g^i\}\log(1+ \\
&\qquad\qquad u_g^i[1-(r_g^\star-1)2\alpha\sqrt{1-\alpha^2}]) \\
&\overset{(d)}{\geq} \sum_{g=1}^{G}\sum_{i=1}^{r_g^\star}\left[1-O\left(\frac{1}{K'}\right)\right]\log(1+ \\
&\qquad\qquad u_g^i[1-(r_g^\star-1)2\alpha\sqrt{1-\alpha^2}]) \\
&\overset{(e)}{\sim} \sum_{g=1}^{G}\sum_{i=1}^{r_g^*}\log(1+(\frac{1-(r_g^*-1)2\alpha\sqrt{1-\alpha^2}}{1/\rho+\epsilon})\lambda_{g,1}\log K')
\end{aligned} \tag{35}$$

$$\overset{(f)}{\sim} \sum_{g=1}^{G} r_g^*\log(1+\rho\lambda_{g,1}\log K') \tag{36}$$

$$\sim \left(\sum_{g=1}^{G} r_g^*\right)\log\rho + \sum_{g=1}^{G} r_g^*\log\lambda_{g,1} + \left(\sum_{g=1}^{G} r_g^*\right)\log\log K' \tag{37}$$

where (a) follows from the suboptimal equal power allocation $\rho = \frac{P}{\sum_{g=1}^{G} r_g^*} = \|\mathbf{w}_{\kappa_{g,i}}\|^2 = \gamma_{\kappa_{g,i}}^{-1}P_{\kappa_{g,i}}$, $\forall g, i$; (b) is obtained by (29) and (13) valid for $\bar{\alpha}$; (c) holds due to the definition (14) of quasi-SINR $Q(g_k)$, the definition (31) of $\phi_{g_k}^i$, and $\mathbb{E}f(X) = \int_0^\infty f(x)p(x)dx \geq \Pr(X \geq u)f(u)$ for a monotone increasing function f (here, $f = \log$) (this step is the reason for the definition (14) of quasi-SINR); (d) holds by (33); (e) follows from $(1 - O(1/K')) \sim 1$ and $u_g^i \sim (\lambda_{g,1}\log K')/(1/\rho+\epsilon)$ from (34); and (f) follows since the difference between the two logarithmic terms in (35) and (36) converges to a constant $\sum_{g=1}^{G} r_g^*\log\left(\frac{1+\rho\epsilon}{1-(r_g^*-1)2\alpha\sqrt{1-\alpha^2}}\right)$ independent of K'. Finally, consider (37). In both cases of $\sum_{g=1}^{G} r_g < M$ and $\sum_{g=1}^{G} r_g \geq M$, we can choose r_g^* such that $\sum_{g=1}^{G} r_g^* = \min\{M, \sum_{g=1}^{G} r_g\} = \beta$. Then, (37) is the same as R_{DPC} in Theorem 1 since $P/\beta = \rho$. ■

Note that fixed α in the range of (30) *guarantees* the asymptotic optimality of ReDOS-PBR. We do not know whether α outside this range yields asymptotic optimality or not. (This depends on the tightness of the bound given by the Gershgorin circle theorem used in (27).) For proof of asymptotic optimality, the existence of one α value, i.e., $\bar{\alpha}$, is sufficient. In the practical case of *finite* users in the cell, optimal α may be smaller than $\sqrt{\frac{1+\sqrt{\frac{r_g^*-2}{r_g^*-1}}}{2}}$. Numerical results in Section VI show that the performance of ReDOS-PBR in the finite-user case is quite insensitive to α.

V. Extension

In the previous section, we only discussed user selection and beamforming for maximizing the sum rate. Now, consider fairness among users. If the channel statistics are the same across users and the channel realizations are i.i.d. across scheduling intervals, the fairness issue will be resolved automatically [27]. However, in slow-fading environments or in practical downlink systems with different large-scale fading for users at different locations, some measures should be applied to impose fairness among users. Among several well-known fairness-imposing schemes [7], [27], [28], we here consider the widely-used proportional fairness (PF) scheme, and modify ReDOS-PBR for PF. During this modification, we exploit the degree-of-freedom associated with the parameter α of ReDOS-PBR (i.e., cone-containment checking is done at users and α can be adapted properly) and the fact that every user reports CQI when $\alpha \leq \alpha_{min}$ by Lemma 1.

The PF scheduling algorithm exploits multiuser diversity gain with consideration of fairness [27]. In the single-input single-output (SISO) PF algorithm, the BS keeps track of the average past served rate μ_{g_k} for each user g_k and selects the user that has the maximum of the current supportable rate $R_{g_k}(t) = \log(1 + |h_{g_k}(t)|^2)$ (determined by the user's current channel state) divided by the user's past average served rate μ_{g_k}. That is, the selection criterion is $\frac{R_{g_k}(t)}{\mu_{g_k}}$ and the average served rate is updated by a simple first-order autoregressive (AR) filter as

$$\mu_{g_k}(t+1) = (1-\delta)\,\mu_{g_k}(t) + \delta R_{g_k}(t) I_{\{g_k \in \mathcal{S}_g(t)\}}, \quad (38)$$

where I_A is the indicator function of event A, and $\mathcal{S}_g(t)$ is the set of scheduled users at scheduling interval t. In [7], the PF algorithm was extended to incorporate MIMO situation and was applied to SUS-ZFBF. The main difference between the SISO and MIMO cases is that the supportable rate $R(g_k, \mathcal{S}_g(t))$ of each user g_k cannot be computed before user selection, because the rate itself depends on the user selection in the MIMO case. However, this difficulty was intelligently circumvented in [7], based on the semi-orthogonality of the selected users. Since ReDOS-PBR also possesses the semi-orthogonality among the selected users, we can apply the same idea as that in [7] here. Since the selected users are semi-orthogonal, we approximate the supportable rate simply by

$$R(g_k, \mathcal{S}_g(t)) \approx \log(1 + Q(g_k)) =: \hat{R}(g_k)(t). \quad (39)$$

Thus, in the modified ReDOS-PBR for PF (ReDOS-PBR-PF), for each reference direction, after the first-phase CQI feedback, we select

$$\kappa_{g,i} = \arg\max_{k \in \mathcal{W}_{g,i}} \frac{\hat{R}(g_k)(t)}{\mu_{g_k}(t)} \quad \text{for } i = 1, \cdots, r_g^*. \quad (40)$$

Then, the BS collects CSI from the selected users in the second-phase feedback, transmits data with ZFBF, computes the exact served rate for the scheduled users, and update μ_{g_k} by (38).

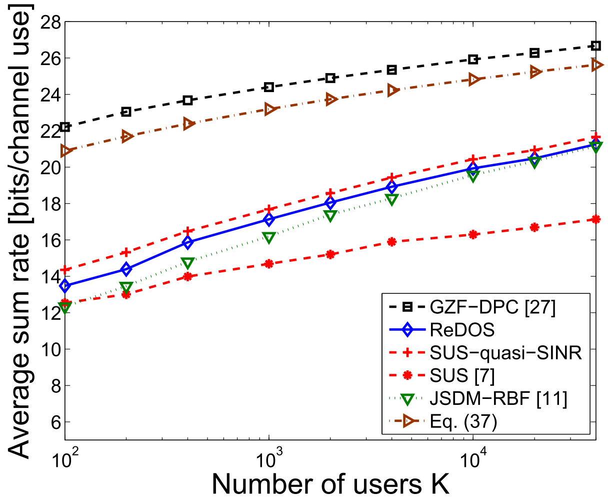
One requirement for ReDOS-PBR-PF to compute (40) for all users at each scheduling interval t is that all users should report their CQI (the cone index and quasi-SINR) to the BS at every interval t. This can be done simply by setting $\alpha = \alpha_{min}$ for all users by Lemma 1. However, CQI feedback can be reduced by exploiting the property of PF itself and by distributed and individual control α at each user. Note that once a user g_k is served, μ_{g_k} increases suddenly and the selection criterion in (40) decreases suddenly. Hence, user g_k will not be selected in the next scheduling interval unless user g_k's channel vector at the next scheduling interval is highly aligned with some reference beam with a large magnitude. Therefore, the served user can increase its own α denoted by $\alpha_{g_k}(t)$ by some step $\Delta_{\alpha,up}$, targeting a bigger chance for good channel realization. When the user is not served, α_{g_k} is reduced by $\Delta_{\alpha,down}$ (say, $\Delta_{\alpha,down} = \Delta_{\alpha,up}/T$ with $T > 1$). Then, $\alpha_{g_k}(t)$ comes back to α_{min} in some time and user g_k surely reports CQI again. Here, $\Delta_{\alpha,up}$ and $\Delta_{\alpha,down}$ are system design parameters which should be determined properly. Note that $\Delta_{\alpha,up}$ and $\Delta_{\alpha,down}$ can be used not only for feedback reduction but also for fairness enhancement, since it is highly likely that a served user will not be served again successively. Such an efficient semi-orthogonality and feedback control is possible for ReDOS-PBR because cone-containment checking for semi-orthogonality is done individually at users. Summarizing the above-mentioned ideas, we now present the proposed ReDOS-PBR-PF:

Algorithm 2 (ReDOS-PBR-PF):

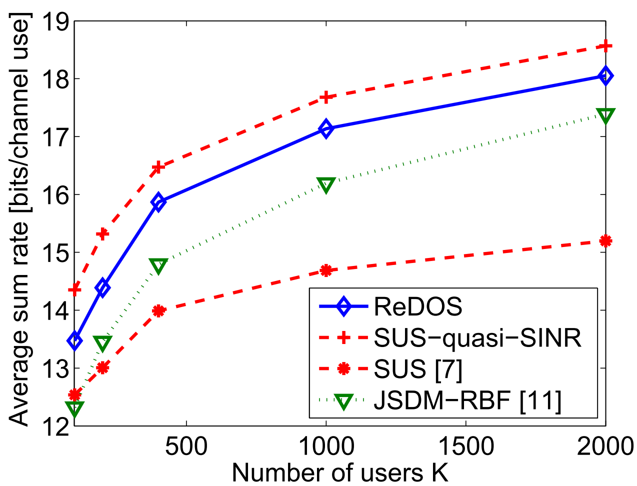
0) Initialize $\alpha_{g_k}(1) = \alpha_{min}$, $\mu_{g_k}(1) = \mu > 0$, $\forall g, k$, and $t = 1$, and $\Delta_{\alpha,up} > \Delta_{\alpha,down} > 0$. α_{min} is defined in Lemma 1. (Now each user has its own $\alpha_{g_k}(t)$.)
1) At time t, each user g_k computes $\mathcal{I}_{g_k}$ in (16) based on its own $\alpha_{g_k}(t)$. Then, follow the remaining sub-steps in step 1) in Algorithm 1 of original ReDOS-PBR.
2) The BS chooses the set of users $\mathcal{S}_g(t)$ by computing (40) after the first-phase CQI feedback.
3) After the second-phase CSI feedback, the BS serves the scheduled users in $\mathcal{S}_g(t)$ with ZFBF. Then, the BS updates $\mu_{g_k}(t)$ according to (38) with the actually served rate $R(g_k, \mathcal{S}_g(t))$.
4) The users in $\mathcal{S}_g(t)$ update $\alpha_{g_k}(t+1) \leftarrow \alpha_{g_k}(t) + \Delta_{\alpha,up}$ and other unserved users update $\alpha_{g_k}(t+1) \leftarrow \alpha_{g_k}(t) - \Delta_{\alpha,down}$. (Users know whether they are served or not during the scheduled user paging period.) When $\alpha_{g_k}(t+1) \notin [\alpha_{min}, 1)$, $\alpha_{g_k}(t+1) \leftarrow \alpha_{g_k}(t)$.
5) Update $t \leftarrow t+1$ and go to step 1).

VI. Numerical Results

In this section, we provide some numerical results regarding the proposed user-scheduling-and-beamforming method. First, we verified the asymptotic analysis in Section IV. To verify the asymptotic analysis, we considered a small MISO downlink system (with two groups ($G = 2$) and inter-group interference) to which DPC-based beamforming [29]



(a)



(b)

Fig. 4. Multi-group performance - number of groups $G = 2$, number of antennas $M = 4$, total transmit power over unit noise variance $P/1 = 15dB$: (a) average sum rate versus K and (b) the same figure as (a) with the range from $K = 100$ to $K = 2000$.

can be applied. The system consisted of a BS with four transmit antennas ($M = 4$) and $P/1 = 15$ dB and K single-antenna users with zero-mean unit noise variance (for all simulations in this section, each receiver noise had unit variance), and the channel vectors were independently generated according to the model (1) with $\mathbf{R}_1 = \mathbf{U}_1\Lambda_1\mathbf{U}_1^H$ and $\mathbf{R}_2 = \mathbf{U}_2\Lambda_2\mathbf{U}_2^H$, where $\mathbf{U}_1 = \mathbf{F}_{DFT}^{(4)}(:, 1 : 3)$, $\mathbf{U}_2 = \mathbf{F}_{DFT}^{(4)}(:, 3 : 4)$, $\Lambda_1 = \text{diag}(1, r, r^2)$, $\Lambda_2 = \text{diag}(1, r)$, $r = 0.7$, and $\mathbf{F}_{DFT}^{(N)}$ is the N-point discrete Fourier transform (DFT) matrix. The outer-beamforming matrices were chosen as $\mathbf{V}_1 = \mathbf{U}_1^* = \mathbf{U}_1(:, 1 : 2)$ and $\mathbf{V}_2 = \mathbf{U}_2^* = \mathbf{U}_2$ to satisfy the approximate BD condition in Condition 1. Fig. 4(a) shows the result.[6] In Fig. 4(a), the performance of the DPC-based beamforming in [29] is shown as the performance reference, and the theoretical analysis result eq. (37) of ReDOS-PBR is also included. It is seen that the actual asymptotic scaling behavior of ReDOS-PBR has a converging slope to the DPC-based user selection method in [29] at large K. The finite-user sum-rate behavior is shown in Fig. 4(b), where ReDOS-PBR, RBF modified for the multi-group case proposed in [11], original SUS-ZFBF in [7], and modified SUS-ZFBF using quasi-SINR in (14) are included. (Since original SUS-ZFBF with the channel norm criterion was proposed for the single-group case, we considered SUS-ZFBF with quasi-SINR for the multi-group case for fair comparison.) It is seen that SUS-ZFBF with quasi-SINR, ReDOS-PBR and RBF all follow the same slope as expected. It is also seen that SUS-ZFBF with the norm criterion does not handle inter-group interference properly. As expected, SUS-ZFBF with quasi-SINR performs best, RBF performs worst, and ReDOS-PBR is in-between. In the considered small system case, the performance difference between the three algorithms is not so significant.

With the asymptotic scaling behavior with respect to (w.r.t.) K verified, we considered a more realistic scenario. We considered the case of the BS with $P/1 = 15$ dB and $M = 32$ antennas. The users were grouped into eight groups ($G = 8$), and the BS served four users ($r_g^* = 4$) simultaneously for each group. The channel covariance matrix $\mathbf{R}_g = \mathbf{U}_g\Lambda_g\mathbf{U}_g^H$ and the outer-beamforming matrix $\mathbf{V}_g$ of each group were chosen as (41), as shown in bottom of this page, with $\Lambda_i = \text{diag}(1, r, r^2, r^3, r^4)$ with $r = 0.6$ for $i = 1, \cdots, 7$ and $\Lambda_8 = \text{diag}(1, r, r^2, r^3)$. This setting of channel covariance matrices and outer-beamforming matrices satisfies the approximate BD condition (8). Fig. 5(a) shows the sum-rate performance of the three schemes: SUS-ZFBF,

[6]The user-selection hyperslab thickness for SUS-ZFBF and the user-selection cone angle for ReDOS-PBR were optimally set for each case.

$$
\begin{aligned}
&\mathbf{U}_1 = \mathbf{F}_{DFT}^{(32)}[:, 1 : 5], \quad \mathbf{V}_1 = \mathbf{U}_1^* = \mathbf{U}_1[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 1 : 4])\\
&\mathbf{U}_2 = \mathbf{F}_{DFT}^{(32)}[:, 5 : 9], \quad \mathbf{V}_2 = \mathbf{U}_2^* = \mathbf{U}_2[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 5 : 8])\\
&\mathbf{U}_3 = \mathbf{F}_{DFT}^{(32)}[:, 9 : 13], \quad \mathbf{V}_3 = \mathbf{U}_3^* = \mathbf{U}_3[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 9 : 12])\\
&\mathbf{U}_4 = \mathbf{F}_{DFT}^{(32)}[:, 13 : 17], \quad \mathbf{V}_4 = \mathbf{U}_4^* = \mathbf{U}_4[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 13 : 16])\\
&\mathbf{U}_5 = \mathbf{F}_{DFT}^{(32)}[:, 17 : 21], \quad \mathbf{V}_5 = \mathbf{U}_5^* = \mathbf{U}_5[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 17 : 20])\\
&\mathbf{U}_6 = \mathbf{F}_{DFT}^{(32)}[:, 21 : 25], \quad \mathbf{V}_6 = \mathbf{U}_6^* = \mathbf{U}_6[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 21 : 24])\\
&\mathbf{U}_7 = \mathbf{F}_{DFT}^{(32)}[:, 25 : 29], \quad \mathbf{V}_7 = \mathbf{U}_7^* = \mathbf{U}_7[:, 1 : 4](= \mathbf{F}_{DFT}^{(32)}[:, 25 : 28])\\
&\mathbf{U}_8 = \mathbf{F}_{DFT}^{(32)}[:, 29 : 32], \quad \mathbf{V}_8 = \mathbf{U}_8^* = \mathbf{U}_8(= \mathbf{F}_{DFT}^{(32)}[:, 29 : 32])
\end{aligned}
\tag{41}
$$

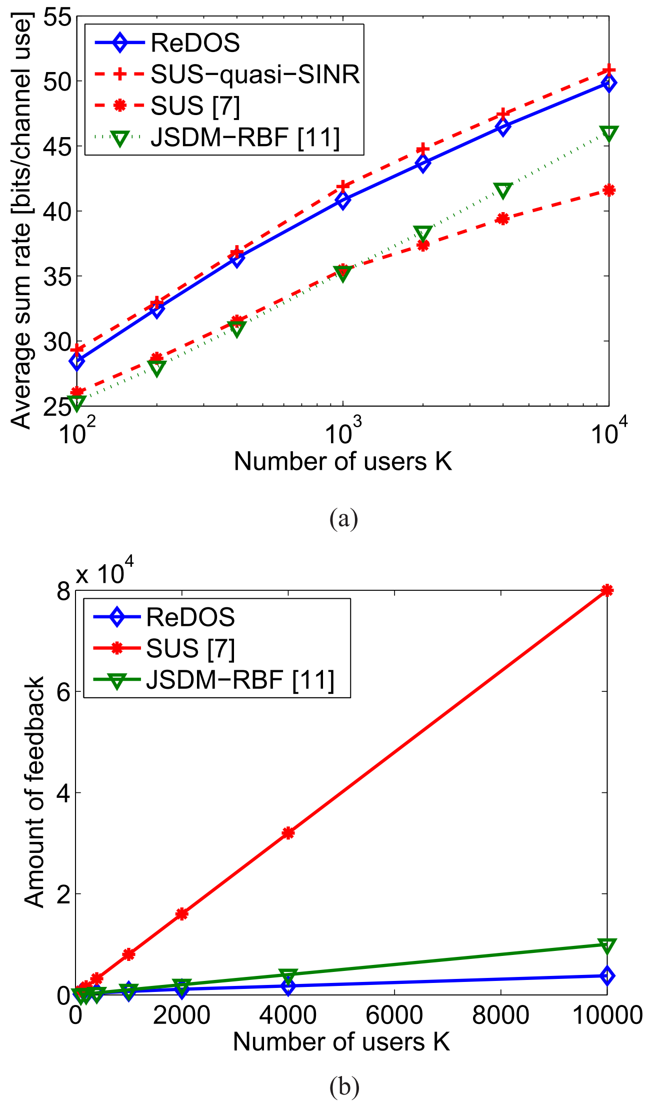


Fig. 5. Multi-group performance - number of groups $G = 8$, number of antennas $M = 32$, total transmit power over unit noise variance $P/1 = 15dB$: (a) average sum rate performance and (b) amount of feedback (number of required real numbers).

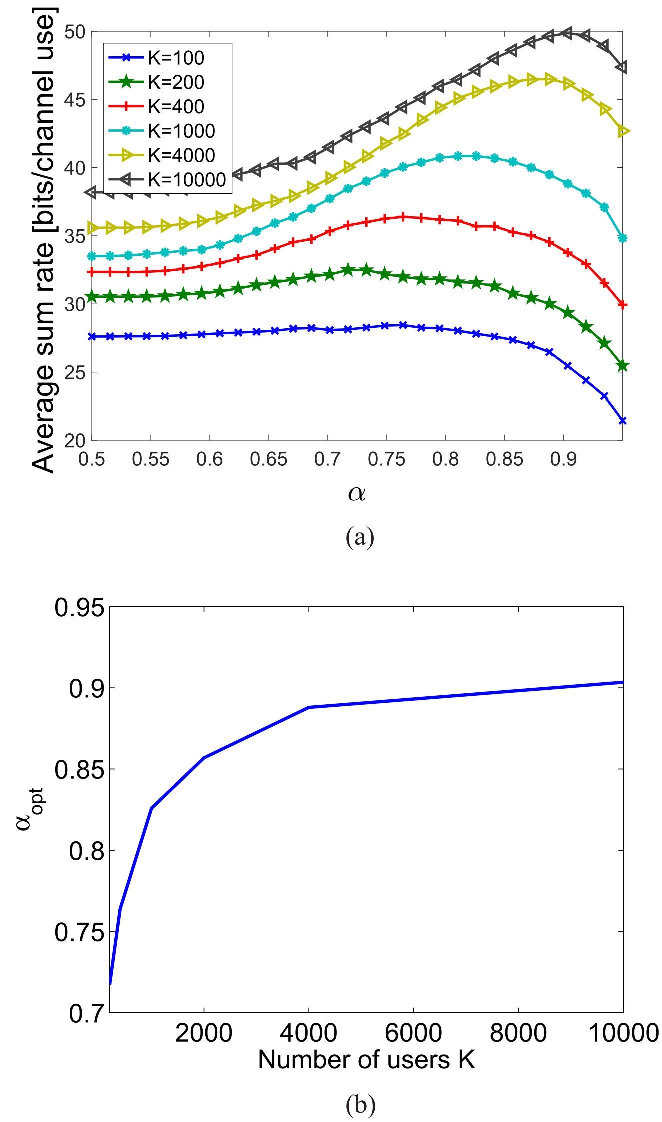


Fig. 6. Impact of the cone angle parameter α - number of groups $G = 8$, number of antennas $M = 32$, total transmit power over unit noise variance $P/1 = 15dB$: (a) sum rate w.r.t. α and (b) optimal α.

RBF, and ReDOS-PBR. 200 independent channel realizations according to the model (1) were used for each K and the average sum rate is the average over the 200 channel realizations. Now it is seen that the performance gap between SUS-ZFBF with quasi-SINR and RBF is significant but ReDOS-PBR closely follows SUS-ZFBF with quasi-SINR. Fig. 5(b) shows the amount of feedback for the same setting as in Fig. 5(a). As expected, SUS-ZFBF requires the largest amount of feedback. Note that the amount of feedback required for ReDOS-PBR is even less than RBF (without thresholding).

We then investigated the performance sensitivity of ReDOS-PBR w.r.t. α in the same setting as in Fig. 5, and the result is shown in Fig. 6. It is seen that optimal α increases as K increases. An observation of practical importance is that the performance of ReDOS-PBR is quite insensitive w.r.t. α for the practical range of the number of users.

We then investigated the impact of the outer beamformer design. For simplicity of analysis, we set $\mathbf{V}_g = \mathbf{U}_g^*$ in this paper. However, we can adopt a different outer beamformer. Currently, there exist several advanced algorithms to design the outer beamformer $\mathbf{V}_g$ based only on $\mathbf{R}_g, g = 1, \cdots, G$ [8], [13], [14]. These methods return an outer beamformer $\mathbf{V}_g$ with orthonormal columns. Hence, the orthonormal columns of $\mathbf{V}_g$ from the advanced design algorithms can be used as the training beams for the proposed scheduling method. We adopted the outer beamformer design method in [14] and applied the obtained outer beamformer to the proposed scheduling method. The result is shown in Fig. 7, where $M = 128$ with ULA, $G = 8$, $r_g^* = 2$ for all g, SNR=15 dB, and $\mathbf{R}_g$ is generated from the one-ring model with 15° two-sided angle spread and the group center angles −52.5°, −37.5°, −22.5°, −7.5°, 7.5°, 22.5°, 37.5°, and 52.5°. It is seen that with the advanced outer beamformer design the performance of the proposed algorithm is slightly improved as compared to the simple setting $\mathbf{V}_g = \mathbf{U}_g^*$.

Next, we considered a single-group case for which SUS-ZFBF [7] and RBF [6] were originally proposed.

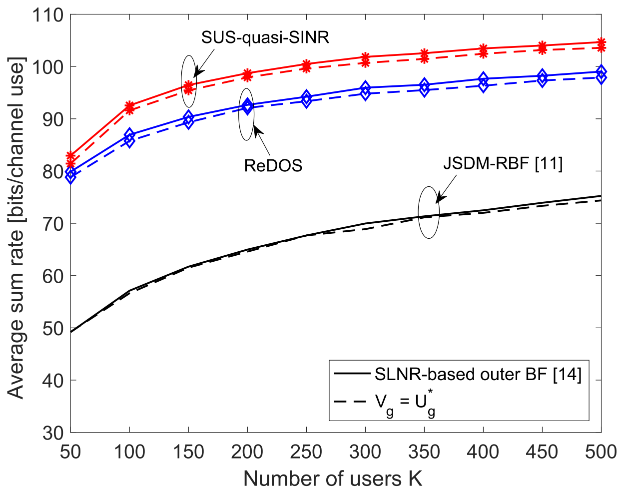
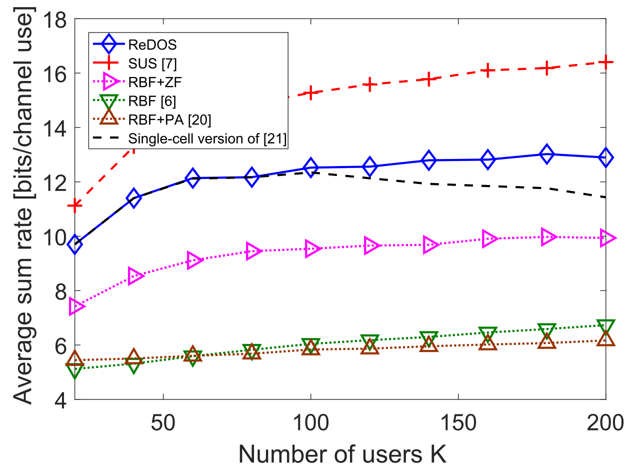
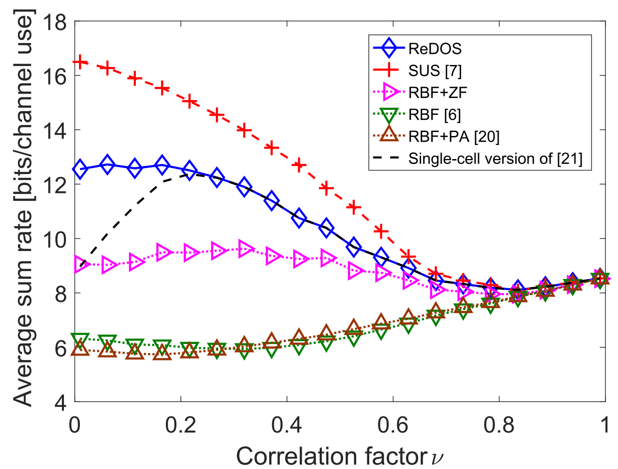


Fig. 7. Impact of the outer beamformer.



(a)



(b)

Fig. 8. Single-group performance - $G = 1$, $M = 8$, $P/1 = 10dB$: (a) average sum rate w.r.t. K and (b) average sum rate w.r.t. the channel correlation factor.

The considered system consists of a BS with $M = r_g^* = 8$, $P/1 = 10dB$, and K single-antenna users with unit noise variance. The channel vector for each user was generated i.i.d. according to the model (1), where for the channel covariance matrix $\mathbf{R}_1$, the exponential correlation model is used, i.e., [30]

$$[\mathbf{R}_1]_{i,j} = \nu^{|i-j|}, \quad 0 \le \nu \le 1. \tag{42}$$

In the case of single group, we compare the performance of two existing corrections of RBF based on two-phase feedback [20], [21] as well as the three schemes SUS-ZFBF [7], RBF [6], ReDOS-PBR. In the method in [20] denoted by 'RBF-PA', the BS schedules users based on RBF and then reallocates transmit power with the same beam directions of RBF by using beam power gain information provided by the selected users in the second-phase feedback. In the single-cell version of [21], the BS selects for each beam the user with the maximum magnitude for the projection of user's channel vector onto the beam direction in the first-phase feedback and transmits data streams via ZFBF based on CSI obtained from the selected users in the second-phase feedback. In addition, we considered the scheme denoted by 'RBF-ZF' in which the BS selects users based on RBF and transmits data streams to the selected users with ZFBF using CSI obtained from the selected users in the second-phase feedback. Fig. 8(a) shows the average sum-rate performance of the considered methods w.r.t. the number K of users for $\nu = 0.2$. It is seen that the ReDOS-PBR outperforms the other three two-phase-feedback-based corrections of RBF: [20], [21] and 'RBF+ZF'. It is seen that the method in [20] has gain over RBF only for small K (up to approximately $K = 60 \sim 70$ in this case) which is the target region of K of the method in [20] as the title of [20] implies. It is seen that the performance of the single-cell version of [21] degrades as K increases or ν decreases. This performance degradation results from the loss in the ZF effective gain caused by the fact that the method in [21] does not directly exploit the angle between user channel vectors for user selection. That is, when the lengths of projections of two user channel vectors onto a reference beam direction are the same, the user with a smaller angle to the reference beam is preferred when considering ZFBF effective gain loss. However, this explicit angle information is not exploited in [21]. It is also seen that RBF+ZF improves the performance over RBF but its performance falls quite short of ReDOS-PBR and [21]. This is because ZFBF based on the second-phase CSI feedback from the selected users removes inter-user interference but the first-phase user selection based on RBF does not well fit into the second-phase ZFBF. Hence, the first-phase user selection should be adapted well to the second-phase beam refinement for good performance of a two-phase-feedback-based scheme. Fig. 8(b) shows the performance of the considered methods w.r.t. the channel correlation factor ν with K fixed to 100 for the same setting as in Fig. 8(a). As expected, when $\nu = 0$, i.e., the channel is isotropic, SUS-ZFBF performs best. On the other hand, when $\nu = 1$, i.e., the channel matrix has rank one and only one beam can be supported, all the six algorithms perform equally. It is seen that the noticeable gap between SUS-ZFBF and ReDOS-PBR at $\nu = 0$ decreases as ν increases towards one. This is because when the channel becomes more correlated, there start to exist dominant eigen-directions of the channel and hence it is enough to make user-selection cones around these dominant eigen-directions to capture most of users.

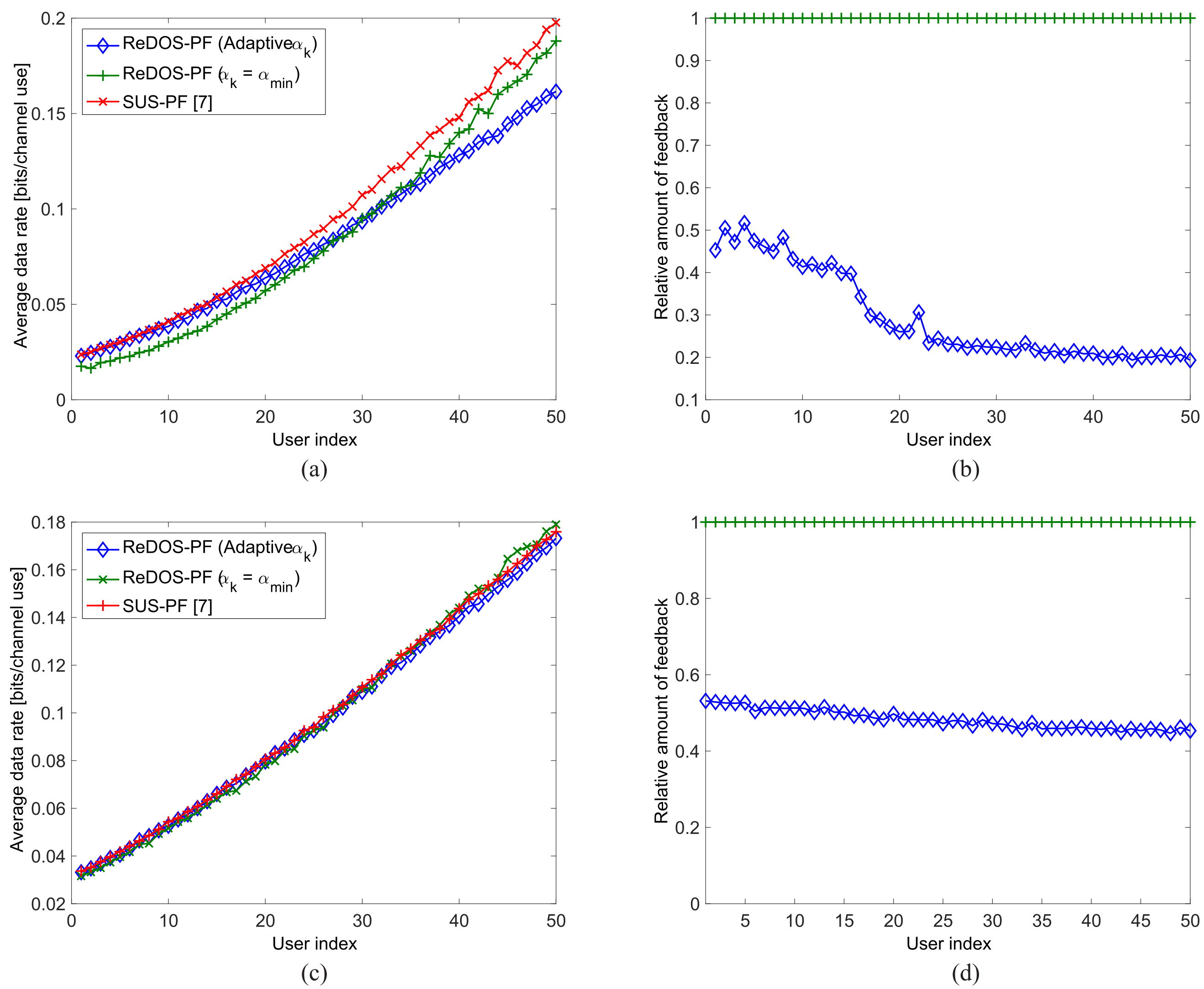


Fig. 9. Performance of ReDOS-PBR-PF - $G = 1$, $M = 4$: (a) each user's served rate, (b) relative amount of feedback between two ReDOS-PBR-PF algorithms: one with fixed α and the other with adaptive α, (c) each user's served rate, and (d) relative amount of feedback between two ReDOS-PBR-PF algorithms ((a) and (b): $\nu = 0.1$ and $\Delta_{\alpha,up} = 0.1$ and $\Delta_{\alpha,down} = \Delta_{\alpha,up}/50$, and (c) and (d): $\nu = 0.3$ and $\Delta_{\alpha,up} = 0.2$ and $\Delta_{\alpha,down} = \Delta_{\alpha,up}/100$).

Finally, we examined the performance of ReDOS-PBR-PF. We considered two ReDOS-PBR-PF algorithms: One with fixed $\alpha = \alpha_{min}$ for all users and the other with adaptive α for each user with steps $\Delta_{\alpha,up}$ and $\Delta_{\alpha,down}$ described in Algorithm 2. To simplify simulation, we just considered a single-group case in which we fixed $K = 50$, BS transmit power $P/1 = 0dB$, and $M = 4$. The channel vector for each user k was generated as $\mathbf{h}_k \sim \sqrt{l_k}\mathbf{R}_1^{1/2}\boldsymbol{\eta}_k$, where $\mathbf{R}_1$ is given in (42); $\boldsymbol{\eta}_k \overset{i.i.d.}{\sim} \mathcal{CN}(\mathbf{0}, \mathbf{I}_4)$; and the large-scale fading effect is captured in l_k. The large-scale fading factor l_k for 50 users were designed such that the lowest power user has $l_k = 1$ and the highest power user has $l_k = 100$ (20 dB difference), and other users' power is equally spaced in dB scale in the 20 dB power range. We ran 10,000 scheduling intervals. For each interval, the channel vector for each user was generated independently as described in the above. Fig. 9(a) shows the average served rate for 50 users (users are ordered in an ascending order of their l_k values) over 10,000 scheduling intervals, when the channel is almost isotropic, i.e., $\nu = 0.1$. It is seen in this case that there is some loss of ReDOS-PBR-PF compared to SUS-ZFBF-PF. Note that ReDOS-PBR-PF with fixed $\alpha = \alpha_{min}$ tracks SUS-ZFBF-PF for all users with equal gap, but ReDOS-PBR-PF with adaptive α sacrifices high-SNR users and gives more chances to low-SNR users. (This is evident in Fig. 9(b).) This is because high-SNR users have more chances to be selected and thus increase their α_k to reduce this increased chance. Fig. 9(b) shows the relative amount of feedback for ReDOS-PBR-PF with adaptive α to that of ReDOS-PBR-PF with fixed α. It is seen that the amount of feedback is significantly reduced by adapting α. Figs. 9(c) and (d) show the performance and the relative amount of feedback in the case of $\nu = 0.3$. It is seen that when the channel correlation increases, the performance difference between ReDOS-PBR-PF and SUS-ZFBF-PF is negligible.

VII. Conclusion

In this paper, we have proposed a new efficient two-phase-feedback-based user-scheduling-and-beamforming method for MU-MIMO downlink in the context of two-stage beamforming. The proposed method realizes the essence of SUS-ZFBF without full CSI feedback based on two-phase feedback. The proposed scheduling-and-beamforming method is asymptotically optimal as the number of users increases, and yields better performance than previous two-phase-feedback-based corrections of RBF. Furthermore, when combined with PF, the proposed method has easy distributed adaptive control of the user selection cone angle for practical implementation. Hence, the proposed method provides an attractive alternative to user selection and beamforming for MU-MIMO downlink.

APPENDIX

Lemma 2: Inner Product Between Two Vectors in Two Different Cones) For two channel vectors contained in two different user-selection cones, i.e., $\mathbf{h}_{\kappa_{g,i}} \in \mathcal{C}_{g,i}$ and $\mathbf{h}_{\kappa_{g,j}} \in \mathcal{C}_{g,j}$, $i \neq j$, the inner product between the corresponding normalized effective channel vectors $\tilde{\mathbf{g}}_{\kappa_{g,i}}$ and $\tilde{\mathbf{g}}_{\kappa_{g,j}}$ with norm one is bounded by

$$|\tilde{\mathbf{g}}_{\kappa_{g,i}}^H \tilde{\mathbf{g}}_{\kappa_{g,j}}| \leq 2\alpha\sqrt{1-\alpha^2} \quad \text{for } i \neq j, \tag{43}$$

when $\alpha \geq 1/\sqrt{2}$ (i.e., the angle $\theta \leq \pi/4$ in Fig. 2).

Proof: Let $\tilde{\mathbf{g}}_{\kappa_{g,i}} = \sum_{m=1}^{r_g^*} c_{\kappa_{g,i}}^m \mathbf{e}_m^{(g)}$ and $\tilde{\mathbf{g}}_{\kappa_{g,j}} = \sum_{m=1}^{r_g^*} c_{\kappa_{g,j}}^m \mathbf{e}_m^{(g)}$, where $\mathbf{e}_m^{(g)}$ is the m-th column of $\mathbf{I}_{r_g^*}$. Then, we have $\sum_m |c_{\kappa_{g,i}}^m|^2 = \sum_m |c_{\kappa_{g,j}}^m|^2 = 1$ and

$$\begin{aligned} |\tilde{\mathbf{g}}_{\kappa_{g,i}}^H \tilde{\mathbf{g}}_{\kappa_{g,j}}| &= \left| \sum_{m=1}^{r_g^*} \bar{c}_{\kappa_{g,i}}^m c_{\kappa_{g,j}}^m \right| \leq \sum_{m=1}^{r_g^*} |\bar{c}_{\kappa_{g,i}}^m| \cdot |c_{\kappa_{g,j}}^m| \\ &= |\bar{c}_{\kappa_{g,i}}^i| \cdot |c_{\kappa_{g,j}}^i| + |\bar{c}_{\kappa_{g,i}}^j| \cdot |c_{\kappa_{g,j}}^j| \\ &\quad + \sum_{m=1,m\neq i,j}^{r_g^*} |\bar{c}_{\kappa_{g,i}}^m||c_{\kappa_{g,j}}^m| \\ &\leq |\bar{c}_{\kappa_{g,i}}^i| \cdot |c_{\kappa_{g,j}}^i| + |\bar{c}_{\kappa_{g,i}}^j| \cdot |c_{\kappa_{g,j}}^j| \\ &\quad + \sqrt{\sum_{m=1,m\neq i,j}^{r_g^*} |\bar{c}_{\kappa_{g,i}}^m|^2} \sqrt{\sum_{m=1,m\neq i,j}^{r_g^*} |c_{\kappa_{g,j}}^m|^2} \end{aligned} \tag{44}$$

where $\bar{c}$ is the complex conjugate of c, and the last step follows from the Cauchy-Schwarz inequality.

Now consider the RHS in (44). First, fix $|\bar{c}_{\kappa_{g,i}}^i|$ and $\{c_{\kappa_{g,j}}^m\}_{m=1}^{r_g^*}$, and view the RHS in (44) as a function of unfixed $\{c_{\kappa_{g,i}}^m, m = 1, \cdots, r_g^* \text{ and } m \neq i \mid \sum_{m=1,m\neq i}^{r_g^*} |c_{\kappa_{g,i}}^m|^2 = 1 - |c_{\kappa_{g,i}}^i|^2\}$. Then, the RHS in (44) is in the form of $a+bx+cy$, where the constants $a, b, c \geq 0$ are given by $a = |\bar{c}_{\kappa_{g,i}}^i| \cdot |c_{\kappa_{g,j}}^i|$, $b = |c_{\kappa_{g,j}}^j|$, and $c = \left(\sum_{m=1,m\neq i,j}^{r_g^*} |c_{\kappa_{g,j}}^m|^2\right)^{\frac{1}{2}}$, and the variables $x, y \geq 0$ are given by $x = |\bar{c}_{\kappa_{g,i}}^j|$ and $y = \sqrt{\sum_{m=1,m\neq i,j}^{r_g^*} |\bar{c}_{\kappa_{g,i}}^m|^2}$, with a constraint $x^2 + y^2 = 1 - |\bar{c}_{\kappa_{g,i}}^i|^2$. The convex optimization of maximizing $a + bx + cy$ under the constraint $x^2 + y^2 = 1 - |\bar{c}_{\kappa_{g,i}}^i|^2$ is solved by using the Karush-Kuhn-Tucker conditions [31], and the solution is given by

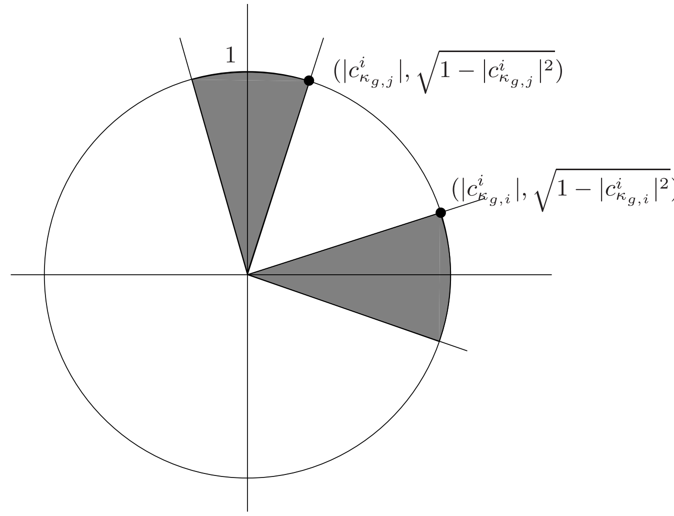


Fig. 10. Maximum inner product between two cones.

$$x = b\sqrt{\frac{1-|c_{\kappa_{g,i}}^i|^2}{1-|c_{\kappa_{g,j}}^i|^2}} \quad \text{and} \quad y = c\sqrt{\frac{1-|c_{\kappa_{g,i}}^i|^2}{1-|c_{\kappa_{g,j}}^i|^2}}. \tag{45}$$

Substituting this x, y into the RHS of (44), we have

$$\begin{aligned} |\tilde{\mathbf{g}}_{\kappa_{g,i}}^H \tilde{\mathbf{g}}_{\kappa_{g,j}}| &\leq a + bx + cy = a + (b^2+c^2)\sqrt{\frac{1-|c_{\kappa_{g,i}}^i|^2}{1-|c_{\kappa_{g,j}}^i|^2}} \\ &\overset{(a)}{=} |c_{\kappa_{g,i}}^i| \cdot |c_{\kappa_{g,j}}^i| + \sqrt{1-|c_{\kappa_{g,i}}^i|^2} \cdot \sqrt{1-|c_{\kappa_{g,j}}^i|^2} \end{aligned} \tag{46}$$

where (a) follows from $b^2 + c^2 = 1 - |c_{\kappa_{g,j}}^i|^2$. Now, the RHS in (46) is expressed in terms of $|c_{\kappa_{g,i}}^i|$ and $|c_{\kappa_{g,j}}^i|$ only. Here, we have the following conditions for the terms in the RHS in (46):

$$|c_{\kappa_{g,i}}^i| \geq \alpha \tag{47}$$

$$\sqrt{1-|c_{\kappa_{g,i}}^i|^2} \leq \sqrt{1-\alpha^2} \tag{48}$$

$$\begin{aligned} |c_{\kappa_{g,j}}^i| &= \sqrt{1 - \sum_{m=1,m\neq i}^{r_g^*} |c_{\kappa_{g,j}}^m|^2} \leq \sqrt{1-|c_{\kappa_{g,j}}^j|^2} \\ &\leq \sqrt{1-\alpha^2} \end{aligned} \tag{49}$$

$$\sqrt{1-|c_{\kappa_{g,j}}^i|^2} \geq \alpha, \tag{50}$$

where (47) and (49) are valid by the cone-containment condition, i.e., the magnitude of the i-component of the normalized effective channel vector $\tilde{\mathbf{g}}_{\kappa_{g,i}}$ is larger than or equal to α (see (16)). The RHS in (46) is the inner product between two points $(|c_{\kappa_{g,i}}^i|, \sqrt{1-|c_{\kappa_{g,i}}^i|^2})$ and $(|c_{\kappa_{g,j}}^i|, \sqrt{1-|c_{\kappa_{g,j}}^i|^2})$ with constraints (47) to (50). The situation is depicted in Fig. 10. The maximum inner product occurs between $(\alpha, \sqrt{1-\alpha^2})$ and $(\sqrt{1-\alpha^2}, \alpha)$ and is given by $2\alpha\sqrt{1-\alpha^2}$. Therefore, we have

$$|\tilde{\mathbf{g}}_{\kappa_{g,i}}^H \tilde{\mathbf{g}}_{\kappa_{g,j}}| \leq 2\alpha\sqrt{1-\alpha^2} \quad \text{for } i \neq j. \tag{51}$$

Without the condition $\alpha \geq 1/\sqrt{2}$, the two shaded regions in Fig. 10 overlap, and we have a trivial upper bound of one. ■

REFERENCES

[1] G. Lee and Y. Sung, "Asymptotically optimal simple user scheduling for massive MIMO downlink with two-stage beamforming," in *Proc. SPAWC*, Toronto, ON, Canada, Jun. 2014, pp. 60–64.

[2] E. Castañeda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.

[3] M. Costa, "Writing on dirty paper (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[4] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[5] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. ISIT*, Chicago, IL, USA, Jun./Jul. 2004, p. 174.

[6] 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.

[7] 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.

[8] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[9] S. Noh, M. D. Zoltowski, Y. Sung, and D. J. Love, "Pilot beam pattern design for channel estimation in massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 787–801, Oct. 2014.

[10] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 802–814, Oct. 2014.

[11] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct. 2014.

[12] J. Nam, G. Caire, and J. Ha, "On the role of transmit correlation diversity in multiuser MIMO systems," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 336–354, Jan. 2017.

[13] J. Chen and V. K. N. Lau, "Two-tier precoding for FDD multi-cell massive MIMO time-varying interference networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1230–1238, Jun. 2014.

[14] D. Kim, G. Lee, and Y. Sung, "Two-stage beamformer design for massive MIMO downlink by trace quotient formulation," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2200–2211, Jun. 2015.

[15] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access*, vol. 2, pp. 947–959, 2014.

[16] R. Zakhour and D. Gesbert, "A two-stage approach to feedback design in multi-user MIMO channels with limited channel state information," in *Proc. IEEE PIMRC*, Sep. 2007, pp. 1–5.

[17] W. Xu and C. Zhao, "Two-phase multiuser scheduling for multiantenna downlinks exploiting reduced finite-rate feedback," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1367–1380, Mar. 2010.

[18] M. Min, D. Kim, H.-M. Kim, and G.-H. Im, "Opportunistic two-stage feedback and scheduling for MIMO downlink systems," *IEEE Trans. Commun.*, vol. 61, no. 1, pp. 312–324, Jan. 2013.

[19] I. Sohn, C. S. Park, and K. B. Lee, "Downlink multiuser MIMO systems with adaptive feedback rate," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 1445–1451, Mar. 2012.

[20] M. Kountouris, D. Gesbert, and T. Sälzer, "Enhanced multiuser random beamforming: Dealing with the not so large number of users case," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1536–1545, Oct. 2008.

[21] B. Khoshnevis, W. Yu, and Y. Lostanlen, "Two-stage channel quantization for scheduling and beamforming in network mimo systems: Feedback design and scaling laws," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2028–2042, Oct. 2013.

[22] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sep. 2007.

[23] T. M. Cover and J. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[24] C. B. Peel, B. M. Hochwald, and A. L. 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.

[25] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[26] G. Lee and Y. Sung. (Mar. 2014). "A new approach to user scheduling in massive multi-user MIMO broadcast channels." [Online]. Available: https://arxiv.org/abs/1403.6931

[27] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[28] Y. Huang and B. D. Rao, "Random beamforming with heterogeneous users and selective feedback: Individual sum rate and individual scaling laws," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2080–2090, May 2013.

[29] G. Dimic and N. D. 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.

[30] T. Al-Naffouri, M. Sharif, and B. Hassibi, "How much does transmit correlation affect the sum-rate scaling of MIMO Gaussian broadcast channels?" *IEEE Trans. Commun.*, vol. 57, no. 2, pp. 562–572, Feb. 2009.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

Gilwon Lee (S'10–M'16) received the B.S., M.S., and Ph.D. degrees in electrical engineering from KAIST, Daejeon, South Korea, in 2010, 2012, and 2016, respectively.

He is currently with Prof. R. Heath as a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, The University of Texas at Austin. His research interests are on signal processing for large-scale MIMO, millimeter wave MIMO, and vehicular-to-everything communication systems and on applying machine learning to future wireless networks.

Youngchul Sung (S'92–M'93–SM'09) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1993 and 1995, respectively, the Ph.D. degree in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2005. He was with LG Electronics, Ltd., Seoul, from 1995 to 2000. From 2005 to 2007, he was a Senior Engineer with the Corporate Research and Development Center of Qualcomm, Inc., San Diego, CA, USA, where he was involved in the design of WCDMA base station modem. Since 2007, he has been on the Faculty of the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. His research interests include signal processing for communications, statistical signal processing, asymptotic statistics, information geometry, and learning.

Dr. Sung was the Chair of the IEEE Communications Society, the Asia Pacific Board (IEEE ComSoc), the Membership Development Committee from 2016 to 2017. He served in various technical committees, including the IEEE Communications Society, Signal Processing and Communications Electronics Technical Committee, the IEEE Signal Processing Society, the Asia-Pacific Signal and Information Processing Association, and the Signal and Information Processing Theory and Methods Technical Committee. He served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2012 to 2014. He has been an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING since 2017. He served as a Guest Editor for the 2012 IEEE JSAC special issue theories and methods for advanced wireless relays.