On the Optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming

Taesang Yoo, *Student Member, IEEE,* and Andrea Goldsmith, *Fellow, IEEE*

***Abstract*—Although the capacity of multiple-input/multiple-output (MIMO) broadcast channels (BCs) can be achieved by dirty paper coding (DPC), it is difficult to implement in practical systems. This paper investigates if, for a large number of users, simpler schemes can achieve the same performance. Specifically, we show that a zero-forcing beamforming (ZFBF) strategy, while generally suboptimal, can achieve the same asymptotic sum capacity as that of DPC, as the number of users goes to infinity. In proving this asymptotic result, we provide an algorithm for determining which users should be active under ZFBF. These users are semiorthogonal to one another and can be grouped for simultaneous transmission to enhance the throughput of scheduling algorithms. Based on the user grouping, we propose and compare two fair scheduling schemes in round-robin ZFBF and proportional-fair ZFBF. We provide numerical results to confirm the optimality of ZFBF and to compare the performance of ZFBF and proposed fair scheduling schemes with that of various MIMO BC strategies.**

***Index Terms*—Broadcast channel, dirty paper coding (DPC), downlink scheduling, fair scheduling, imperfect channel state information (CSI), multiple-input/multiple-output (MIMO), multiple-input/multiple-output capacity, multiuser diversity, proportional fair, zero-forcing beamforming (ZFBF).**

I. Introduction

MULTIPLE-INPUT/multiple-output (MIMO) systems have great potential to achieve high throughput in wireless systems [1], [2]. With M transmit antennas at the base station and N receive antennas at the user terminal, it is well known that in a single-user case the capacity gain is roughly $\min\{M, N\}$ times that of single-input/single-output (SISO) systems [1], [2]. In cellular systems, multiple antennas can be easily deployed at the base station to achieve this benefit. In many cases, however, mobile terminals have a smaller number of antennas than the base station due to size and cost constraints. In this case, since $\min\{M, N\} = N$, it may appear that we do not obtain significant capacity benefit from the multiple transmit antennas. Indeed, this is true with the transmit strategy of time-division multiple access (TDMA), where the base station serves one user at a time. Thus, with a limited number of receive antennas at each user, TDMA cannot achieve a linear increase of sum capacity (sum rate or system throughput, i.e., the aggregate data rate of users) in the number of transmit antennas [3], [4].

The solution to this problem is to serve multiple users simultaneously. One way to accomplish this is to use a coding scheme called dirty paper coding (DPC), which is a multi-user encoding strategy based on interference presubtraction [5]. In particular, when the number of users K exceeds the number of transmit antennas M, regardless of N, a linear increase of capacity in M can be achieved by using DPC. In fact, DPC is the optimal (capacity achieving) strategy in MIMO broadcast channels (BCs or downlink, i.e., channels from the base station to mobile users) [6]. However, DPC is difficult to implement in practical systems due to the high computational burden of successive encodings and decodings, especially when the number of users K is large.

Beamforming (BF) [7] is a suboptimal strategy that can also serve multiple users at a time, but with reduced complexity relative to DPC. In BF, each user stream is coded independently and multiplied by a beamforming weight vector for transmission through multiple antennas. Careful selection of weight vectors can reduce (or eliminate) mutual interference among different streams by taking advantage of spatial separation between users and thereby support multiple users simultaneously. This type of multiuser communication scheme is called space-division multiple access (SDMA). Despite its reduced complexity, BF has been shown to achieve a fairly large fraction of DPC capacity when the base station has multiple antennas and each user has a single antenna [7]–[10]. Moreover, it has been shown that if the beamforming vectors are chosen optimally, the sum rate of BF approaches that of DPC as the number of users K goes to infinity [4].

Finding the optimal beamforming weight vectors, however, is still a difficult nonconvex optimization problem [7]. In this paper, we seek a very simple transmit strategy that can easily be implemented in practice but whose performance is comparable to that of DPC. In particular, we consider a suboptimal beamforming strategy, *zero-forcing beamforming* (ZFBF), where the weight vectors are chosen to avoid interference among user streams. Such beamforming weights can be easily found by inverting the composite channel matrix of the users. ZFBF is generally power inefficient because beamforming weights are not matched to user channels. However, we will see that when the number of users K is sufficiently large, its sum-rate performance comes close to that of DPC. This is due to a *multiuser diversity* effect [11], [12].

Multiuser diversity is a form of selection diversity among users; when the number of users K is large, the base station can schedule its transmission to those users with favorable channel fading conditions to improve the system throughput. In MIMO channels with independently fading coefficients, the benefit of multiuser diversity comes from two different factors. First, multiuser diversity provides increased channel *magnitudes*. For example, assuming homogeneous [equal average signal-to-noise

Manuscript received January 14, 2005; revised October 5, 2005. This work was supported in part by the National Science Foundation under Contract CCR-0325639-001.

The authors are with Stanford University, Stanford, CA 94305 USA (e-mail: yoots@wsl.stanford.edu; andrea@wsl.stanford.edu).

Digital Object Identifier 10.1109/JSAC.2005.862421

0733-8716/$20.00 © 2006 IEEE

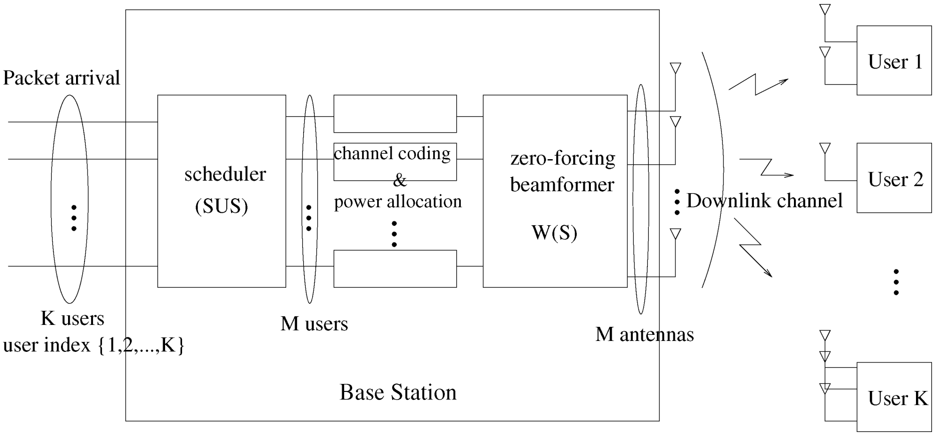

Fig. 1. MIMO downlink system with M transmit antennas and $K \geq M$ users each with N_k receive antennas. Plot shows proposed downlink strategy of ZFBF combined with SUS algorithm (see Section IV-B).

ratio (SNR)] users, the channel gain (SNR) of the best user is roughly $\log K$ times higher than the average channel gains, so multiuser diversity increases SNR by the same factor. Second, multiuser diversity offers abundant channel *directions*. This enables the base station to choose a user group with good spatial separations, which is why certain simple suboptimal schemes often exhibit fairly good performance under large K. For example, in [13] the authors demonstrate that a suboptimal zero-forcing receiver approaches the performance of the optimal receiver under large K. In [14], the authors propose an orthogonal random beamforming (RBF) scheme and show that it asymptotically achieves the optimal sum rate of DPC. This is made possible because multiuser diversity enables even a random beam direction to be nearly matched to certain users. One advantage of RBF is that it can be implemented with partial channel state information (CSI) at the base station. However, it has slow convergence in K, which yields poor performance for practical values of K, e.g., $K < 100$ for $M = 4$ and $N = 1$ (see numerical results in Section V). When full CSI is available, a better choice of beamforming directions can be made that has fairly good performance under the zero-forcing strategy. Namely, the transmitter can choose a group of users with high channel *magnitudes* and for which their channel *directions* are matched to zero-forcing beam directions. We propose an algorithm (called the semiorthogonal user selection (SUS) algorithm) for such a user group selection and prove that ZFBF with the chosen user set achieves the same asymptotic sum rate as DPC. Thus, even a very simple strategy like ZFBF with a heuristic user selection becomes asymptotically optimal in the sum-rate sense at large K. This asymptotic optimality results directly from multiuser diversity. The SUS algorithm we propose is based on a semiorthogonal user selection. We note that in a parallel work [15], ZFBF was combined with a similar heuristic criterion for a low-complexity scheduling problem with queueing.

Note that we use the term "large K" qualitatively. The question of how large K should be will become clearer as we go through Section IV. In that section, in proving the optimality of ZFBF, we make use of the law of large numbers and extreme value theory, which implies that we need K to approach infinity to achieve the optimal performance of DPC. However, numerical results in Section V will show that our ZFBF-SUS scheme performs reasonably well under practical values of K, e.g., $K < 100$. Thus, in the context of asymptotically optimal results, we shall take on the order of $K \approx 10^4$ or above to be large. However, in the context of practical values that give performance reasonably close to that of DPC, K in the range of several multiples of M will suffice to be considered large.

Although we mainly use sum rate as a performance metric for convenience, it does not capture fairness among users, which is an important issue in practical designs. We extend our SUS together with ZFBF to a fair scheduling scheme that essentially combines TDMA and SDMA. We propose two different fair scheduling methods: round-robin ZFBF (RR-ZFBF) and proportional-fair ZFBF (PF-ZFBF), of which PF-ZFBF makes use of multiuser diversity and hence has a higher average throughput. We also analyze the performance loss of ZFBF due to inaccurate CSI at the transmitter. Simulations show that the performance of ZFBF degrades rapidly if CSI at the transmitter is inaccurate. The benefit of ZFBF over TDMA is most conspicuous when the users have only a single receive antenna. However, we will see that the sum-rate asymptotic optimality of ZFBF is valid for any number of receive antennas at each user.

II. System Model

Consider a single-cell MIMO BC with a single base station supporting data traffic to K user terminals, as depicted in Fig. 1. The base station is equipped with M transmit antennas and the kth user terminal with N_k receive antennas. We assume that $K \geq M$. We use a simple channel model where the channel gain from a transmit antenna to a user is described by a zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variable, which is an appropriate model for narrowband systems operating in a nonline-of-sight rich scattering environment [16]. For simplicity, we assume that all the users are homogeneous and experience independent fading. The signal received by a user k may be written as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \mathbf{z}_k, \qquad k = 1, \dots, K \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is the transmitted symbol from the base station antennas, $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$ is the channel gain matrix to the kth user, $\mathbf{z}_k \in \mathbb{C}^{N_k \times 1}$ is the additive white Gaussian noise

(AWGN) at the kth user,[1] and $\mathbf{y}_k$ is the received signal vector by user k. The entries of $\mathbf{H}_k$ are assumed to be independent. We normalize the channel and noise such that the entries of $\mathbf{H}_k$ and $\mathbf{z}_k$ have unit variance, and the transmitter has an average power constraint $\mathcal{E}\{\mathbf{x}\mathbf{x}^*\} \leq P$. Unless otherwise stated, perfect CSI at the base station is assumed. For ease of explanation, we assume $N_k = 1$, $k = 1, \ldots, K$ in most parts of the paper. Extension to multiple receive antennas is straightforward and will be discussed in Section VIII. We will focus on the ergodic sum rate of this system, i.e., the long term sum rate averaged over channel realizations. Note that this metric is most suitable for applications without a stringent delay constraint [17].

A. Notation

We use uppercase boldface letters for matrices and lowercase boldface for vectors. Consistent with this rule, we use $\mathbf{h}_k$ in place of $\mathbf{H}_k$ for the kth user's channel when $N_k = 1$. $\mathcal{E}(\bullet)$ stands for the expectation operator, and $P\{\bullet\}$ is the probability of the given event. $\mathbf{X}^*$ ($\mathbf{x}^*$) stands for the conjugate transpose of a matrix $\mathbf{X}$ (vector $\mathbf{x}$). $|\mathcal{A}|$ denotes the size of a set $\mathcal{A}$.

III. Review of MIMO BC Strategies

In this section, we briefly go over three exemplary MIMO BC transmission schemes.

A. TDMA

In the conventional TDMA scheme, the base station transmits to only a single user at a time. In this case, the maximum sum rate, achieved by sending to the user with the largest channel gain, is given by

$$R_{\mathrm{TDMA}} = \max_{k \in \{1,\ldots,K\}} \log\left(1 + P\|\mathbf{h}_k\|^2\right). \qquad (2)$$

In [4], the scaling law of this TDMA scheme is shown to be

$$\mathcal{E}\left\{R_{\mathrm{TDMA}}\right\} \sim \log\left(1 + P \log K\right) \qquad (3)$$

where[2] $x \sim y$ indicates that $\lim_{K\to\infty} x/y = 1$. Compared to the single-user capacity of $\log(1+P)$, we observe that the sum rate increases double logarithmically in K. The effective SNR, being the maximum of K, i.i.d. χ^2_{2M} distributed random variables, benefits by a factor of $\log K$ asymptotically for large K. Thus, the *multiuser diversity gain* increases SNR by a factor of $\log K$.

B. DPC

DPC is the capacity-achieving strategy in MIMO BCs [6]. Specifically, it achieves the following sum-rate capacity [18]–[20]:

$$R_{\mathrm{DPC}} = \max_{P_k \geq 0, \sum_{k=1}^{K} \leq P} \log\left(\left|\mathbf{I} + \sum_{k=1}^{K} P_k \mathbf{h}_k^* \mathbf{h}_k\right|\right) \qquad (4)$$

where P_k is the transmit power allocated to user k. Furthermore, it is shown in [4] that in the limit of large K, R_{DPC} satisfies

$$\mathcal{E}\left\{R_{\mathrm{DPC}}\right\} \sim M \log\left(1 + \frac{P}{M}\log K\right). \qquad (5)$$

In addition to the multiuser diversity gain of $\log K$, DPC achieves a full spatial multiplexing gain, i.e., a linear increase of the sum rate in M. DPC is of great theoretical interest, but implementing DPC is a challenging task due to its complexity, as will be discussed in detail in Section IV-D.

C. BF

In BF, user streams are separated by different beamforming directions. Let s_k, $\mathbf{w}_k$, and P_k be, respectively, the data symbol, beamforming weight vector, and transmit power scaling factor for user k. Define $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]$, $\mathbf{s} = [s_1 \cdots s_K]^T$, and $\mathbf{P} = \mathrm{diag}\{P_1, \ldots, P_k\}$ so that the transmitted signal is $\mathbf{x} = \sum_{k=1}^{K} \sqrt{P_k}\mathbf{w}_k s_k$. For user k, we have the following:

$$y_k = (\sqrt{P_k}\mathbf{h}_k\mathbf{w}_k)s_k + \sum_{j \neq k} \sqrt{P_j}\mathbf{h}_k\mathbf{w}_j s_j + z_k. \qquad (6)$$

The receiver detects the transmitted symbol s_k by simply treating the interference terms as an additive Gaussian noise. The sum rate achieved by this scheme is [9]

$$R_{\mathrm{BF}} = \max_{\mathbf{w}_k, P_k} \sum_{k=1}^{K} \log\left(\frac{1 + \sum\limits_{j=1}^{K} P_j|\mathbf{h}_k\mathbf{w}_j|^2}{1 + \sum\limits_{j=1, j\neq k}^{K} P_j|\mathbf{h}_k\mathbf{w}_j|^2}\right)$$

$$\text{subject to } \sum_{k=1}^{K} \|\mathbf{w}_k\|^2 P_k \leq P. \qquad (7)$$

In [4], the authors show that the *optimal* BF in (7) has the same growth rate as DPC, i.e.,

$$\mathcal{E}\left\{R_{\mathrm{BF}}\right\} \sim M \log\left(1 + \frac{P}{M}\log K\right). \qquad (8)$$

The proof is based on the observation that R_{BF} is lower bounded by the sum rate of the orthogonal random beamforming (RBF) strategy proposed in [14], which achieves the same asymptotic throughput in (5) as DPC. Since RBF is a suboptimal BF strategy, the optimal BF in (7) should also be asymptotically optimal in K.

Determining the optimal $\mathbf{w}_k$s and P_ks of BF is difficult in practice, especially for large K [7]. In this paper, we are instead interested in a practical scheme that has very low computational complexity but still achieves the full spatial multiplexing and multiuser diversity gains as DPC. Moreover, the scheme should perform reasonably well under practical values of K, i.e., for a small number of users. In the next section, we investigate a simple but suboptimal BF strategy, the *zero-forcing beamforming (ZFBF)* [8], [9], [16]. We will see that ZFBF is also asymptotically optimal in the sum-rate sense when K is large.

[1] $\mathbf{z}_k$ includes out-of-cell interference, which we assume to be Gaussian for simplicity.

[2] Since the constant terms inside the outer logarithm are insignificant, we could equivalently write $\mathcal{E}\{R_{\mathrm{TDMA}}\} \sim \log\log K$. However, to explicitly state the effect of K on the SNR, we use formulas in the form of $\log(1 + \mathrm{SNR})$ throughout this paper.

IV. Optimality of ZFBF With SUS Algorithm

A. ZFBF

In ZFBF [8], [9], [16], beamforming vectors are selected such that they satisfy the zero-interference condition $\mathbf{h}_k\mathbf{w}_j = 0$ for $j \neq k$. Fig. 1 shows a block diagram of the ZFBF system. Let $\mathcal{S} \subset \{1, \ldots, K\}, |\mathcal{S}| \leq M$ be the scheduler output, i.e., a subset of user indexes that the base station intends to transmit to, and $\mathbf{H}(\mathcal{S})$ and $\mathbf{W}(\mathcal{S})$ are the corresponding submatrices of $\mathbf{H} = [\mathbf{h}_1^T \cdots \mathbf{h}_K^T]^T$ and $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]$, respectively. One easy choice of $\mathbf{W}(\mathcal{S})$ that gives zero-interference is the pseudoinverse of $\mathbf{H}(\mathcal{S})$

$$\mathbf{W}(\mathcal{S}) = \mathbf{H}(\mathcal{S})^\dagger = \mathbf{H}(\mathcal{S})^*(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^*)^{-1}. \tag{9}$$

Then, (7) becomes

$$R_{\mathrm{ZFBF}}(\mathcal{S}) = \max_{P_i:\sum_{i\in\mathcal{S}} \gamma_i^{-1}P_i \leq P} \sum_{i\in\mathcal{S}} \log(1 + P_i) \tag{10}$$

where

$$\gamma_i = \frac{1}{\|\mathbf{w}_i\|^2} = \frac{1}{[(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^*)^{-1}]_{i,i}} \tag{11}$$

can be interpreted as the effective channel gain[3] to the ith user. The optimal P_i in (10) is easily found by waterfilling

$$P_i = (\mu\gamma_i - 1)^+ \tag{13}$$

where $\{x\}^+$ denotes $\max\{x, 0\}$, and the water level μ is chosen to satisfy

$$\sum_{i\in\mathcal{S}} \left(\mu - \frac{1}{\gamma_i}\right)^+ = P. \tag{14}$$

Finally, the achievable sum rate of ZFBF is found by considering every possible choice of user groups $\mathcal{S}$

$$R_{\mathrm{ZFBF}} = \max_{\mathcal{S}\subset\{1,\ldots,K\}:|\mathcal{S}|\leq M} R_{\mathrm{ZFBF}}(\mathcal{S}). \tag{15}$$

Our objective is to show the following main theorem.

Theorem 1: In the limit of large K, the ZFBF transmit strategy can achieve an expected sum rate equal to that of DPC

$$\boxed{\mathcal{E}\{R_{\mathrm{ZFBF}}\} \sim M\log\left(1 + \tfrac{P}{M}\log K\right) \sim \mathcal{E}\{R_{\mathrm{DPC}}\}}. \tag{16}$$

Thus, ZFBF fully achieves both the multiplexing gain and the multiuser diversity gain, i.e., it is asymptotically optimal at large K.

The proof is given in the following two sections. The intuition behind the asymptotic optimality of ZFBF is that the randomness in users' channel gains reduces the loss coming from inverting the channel. In particular, when $\mathbf{H}(\mathcal{S})$ is poorly conditioned, the effective channel gain (11) is greatly reduced. However, with a large number of users, the transmitter can almost surely choose a group of M users that are nearly orthogonal to one another. Then, inverting the channel becomes merely a rotation operation, hence there is no loss in channel gains so we can achieve a linear increase in capacity of M. The $\log\log K$ term in (16) implies that imposing near orthogonality among selected users does not reduce the multiuser diversity gain, i.e., in selecting near-orthogonal users we still have choices among a user set with cardinality of order K.

From (9)–(15), we see that the implementation of ZFBF consists of two stages: the optimal user group $\mathcal{S}$ selection (the scheduler in Fig. 1) in (15) and the beamforming weight vector calculation and the optimal power allocation within the selected user group (the zero-forcing beamformer in Fig. 1) as in (9) and (10). While the latter has a very low implementation cost[4] regardless of K, the former stage, determining the optimal $\mathcal{S}$ in (15), requires an exhaustive search over the entire user set. In [8] and [9], the authors only consider a relatively small number of users ($K \leq M$ in [8] and K up to 16 in the numerical examples in [9]), and therefore such a brute-force search may be feasible. However, when K is large, such a method cannot be used any longer, since the size of its search space, $\sum_{i=1}^{M}\binom{K}{i}$, becomes prohibitively large. For example, with $K = 100$ users and $M = 4$ transmit antennas, we have $\sum_{i=1}^{M}\binom{K}{i} \approx 4 \times 10^6$. This makes a low-complexity implementation of the optimal ZFBF challenging. However, our proof of Theorem 1 will be constructive in the sense that we propose a low-complexity suboptimal user selection scheme called the SUS algorithm (Section IV-B) and show that its performance is still asymptotically optimal (Section IV-C). The complexity of the proposed algorithm is discussed in Section IV-D.

B. Construction of Semiorthogonal User Group (SUS Algorithm)

In this section, we construct a suboptimal user group $\mathcal{S}_0$ using a semiorthogonal user selection (SUS) algorithm as follows.

Step 1) Initialization:

$$\mathcal{T}_1 = \{1, \ldots, K\} \tag{17}$$

$$i = 1 \tag{18}$$

$$\mathcal{S}_0 = \phi \ (\text{empty set}). \tag{19}$$

Step 2) For each user $k \in \mathcal{T}_i$, calculate $\mathbf{g}_k$, the component of $\mathbf{h}_k$ orthogonal to the subspace spanned by $\{\mathbf{g}_{(1)}, \ldots, \mathbf{g}_{(i-1)}\}$

$$\mathbf{g}_k = \mathbf{h}_k - \sum_{j=1}^{i-1} \frac{\mathbf{h}_k\mathbf{g}_{(j)}^*}{\|\mathbf{g}_{(j)}\|^2}\mathbf{g}_{(j)} \tag{20}$$

$$= \mathbf{h}_k\left(\mathbf{I} - \sum_{j=1}^{i-1}\frac{\mathbf{g}_{(j)}^*\mathbf{g}_{(j)}}{\|\mathbf{g}_{(j)}\|^2}\right). \tag{21}$$

When $i = 1$, this implies $\mathbf{g}_k = \mathbf{h}_k$.

[3] An equivalent expression to (10) is

$$R_{\mathrm{ZFBF}}(\mathcal{S}) = \max_{\tilde{P}_i:\sum_{i\in\mathcal{S}} \tilde{P}_i \leq P} \sum_{i\in\mathcal{S}} \log(1 + \gamma_i\tilde{P}_i) \tag{12}$$

where $\tilde{P}_i = \gamma_i^{-1}P_i = \|\mathbf{w}_i\|^2 P_i$ is the transmit power allocated to the ith user, and $\gamma_i\tilde{P}_i$ is the received SNR of the user. In this regard, γ_i can be interpreted as the effective channel gain to the ith user.

[4] This will be elaborated upon in Section IV-D.

Step 3) Select the ith user as follows:

$$\pi(i) = \arg\max_{k \in \mathcal{T}_i} \|\mathbf{g}_k\| \tag{22}$$

$$\mathcal{S}_0 \leftarrow \mathcal{S}_0 \cup \{\pi(i)\} \tag{23}$$

$$\mathbf{h}_{(i)} = \mathbf{h}_{\pi(i)} \tag{24}$$

$$\mathbf{g}_{(i)} = \mathbf{g}_{\pi(i)}. \tag{25}$$

Step 4) If $|\mathcal{S}_0| < M$, then calculate $\mathcal{T}_{i+1}$, the set of users semiorthogonal to $\mathbf{g}_{(i)}$

$$\mathcal{T}_{i+1} = \left\{ k \in \mathcal{T}_i,\ k \neq \pi(i) \mid \frac{|\mathbf{h}_k \mathbf{g}^*_{(i)}|}{\|\mathbf{h}_k\| \|\mathbf{g}_{(i)}\|} < \alpha \right\} \tag{26}$$

$$i \leftarrow i + 1 \tag{27}$$

where α is a small positive constant.[5] If $\mathcal{T}_{i+1}$ is nonempty and $|\mathcal{S}_0|$, the number of elements in the set $\mathcal{S}_0$, satisfies $|\mathcal{S}_0| < M$, then go to Step 2). Otherwise, the algorithm is finished.

By construction, it is easily seen that $\{\mathbf{g}_{(i)}, 1 \leq i \leq |\mathcal{S}_0|\}$ is a set of orthogonal vectors in $\mathbb{C}^{1\times M}$. The algorithm works as follows: in Step 2) we project user channels in $\mathcal{T}_i$ to the orthogonal complement of $\mathrm{span}\{\mathbf{g}_{(1)}, \ldots, \mathbf{g}_{(i-1)}\}$. Note that those user channels in $\mathcal{T}_i$ are already semiorthogonal to $\mathbf{g}_{(1)}, \ldots, \mathbf{g}(i-1)$, because those users whose channels are not semiorthogonal to one of the $\mathbf{g}_{(1)}, \ldots, \mathbf{g}_{(i-1)}$ would have been dropped off $\mathcal{T}_i$ in Step 4) of the previous iterations. Thus, $\mathbf{g}_k \approx \mathbf{h}_k$ for $k \in \mathcal{T}_i$. Then, in Step 3), we select the best user $\pi(i)$ (the one with the largest projected norm), its channel $\mathbf{h}_{(i)}$, and the next basis vector $\mathbf{g}_{(i)}$. Since $\mathbf{g}_{(i)} \approx \mathbf{h}_{(i)}$, the selected user channels $\{\mathbf{h}_{(1)}, \ldots, \mathbf{h}_{(M)}\}$ become semiorthogonal to one another with relatively large gains.

A scheduling algorithm based on semiorthogonality has been proposed in [21] and [15]. A similar algorithm to the SUS has been proposed in [22] and [23] for user subset selection in *zero-forcing DPC*, which is a suboptimal DPC scheme based on a QR decomposition of the channel [8]. This algorithm is similar to Steps 1)–3) of our proposed SUS algorithm, but the algorithm in [22] and [23] does not force semiorthogonality among users as the SUS does through Step 4). Although this step may not be necessary in zero-forcing DPC, where any remaining interference terms can be cancelled by dirty paper precoding, in ZFBF, however, selecting a nonorthogonal user degrades the effective channel gains (11) of the other users, as will be seen in Lemma 2. Therefore, forcing semiorthogonality among users is useful in ZFBF. In fact, we will see in the next section that the assumption of semiorthogonality among users plays a pivotal role in proving Theorem 1. Another advantage of having Step 4) is that it reduces the complexity (running time) of the algorithm by eliminating those users not semiorthogonal to the selected user from further consideration in subsequent iterations. Without Step 4, we would need all the users to go through M iterations of Steps 2)–4). With Step 4, however, a large fraction of the users are dropped off at intermediate iterations, greatly reducing the running time of the algorithm.[6]

C. *Performance Lower Bound of ZFBF-SUS*

In this section, we provide a proof of Theorem 1. Rather than directly finding $\mathcal{E}\{R_{\mathrm{ZFBF}}\}$, we derive an asymptotic lower bound to $\mathcal{E}\{R_{\mathrm{ZFBF}}(\mathcal{S}_0)\}$ and show that $\mathcal{E}\{R_{\mathrm{ZFBF}}(\mathcal{S}_0)\} \sim \mathcal{E}\{R_{\mathrm{DPC}}\}$. Since $R_{\mathrm{ZFBF}} \geq R_{\mathrm{ZFBF}}(\mathcal{S}_0)$ for every channel realization, Theorem 1 readily follows. Throughout the proof, we will assume that $|\mathcal{S}_0| = M$. This is almost surely true if K is very large and α is not very small; more precise conditions on K and α will be discussed through this section.

Let $\mathbf{H}(\mathcal{S}_0) = \left[\mathbf{h}^T_{(1)}, \ldots, \mathbf{h}^T_{(M)}\right]^T$ and define $h_{ij} = \mathbf{h}_{(i)}\mathbf{g}^*_{(j)}/\|\mathbf{g}_{(j)}\|$, the component of $\mathbf{h}_{(i)}$ along $\mathbf{g}_{(j)}/\|\mathbf{g}_{(j)}\|$. From (21), we have

$$\mathbf{h}_{(i)} = \mathbf{g}_{(i)} + \mathbf{h}_{(i)} \sum_{j=1}^{i-1} \frac{\mathbf{g}^*_{(j)}\mathbf{g}_{(j)}}{\|\mathbf{g}_{(j)}\|^2} \tag{28}$$

$$= \mathbf{g}_{(i)} + \sum_{j=1}^{i-1} h_{ij} \frac{\mathbf{g}_{(j)}}{\|\mathbf{g}_{(j)}\|}. \tag{29}$$

Then, $\mathbf{H}(\mathcal{S}_0)$ can be decomposed as

$$\mathbf{H}(\mathcal{S}_0) = \mathbf{DRQ} \tag{30}$$

where $\mathbf{D}$ is diagonal with $\|\mathbf{g}_{(i)}\|$ as its (i,i)th element, $\mathbf{R}$ is lower triangular

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \epsilon_{21} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & 0 \\ \epsilon_{M1} & \epsilon_{M2} & \cdots & \epsilon_{M,M-1} & 1 \end{bmatrix} \tag{31}$$

where $\epsilon_{ij} = h_{ij}/\|\mathbf{g}_{(i)}\|$, and $\mathbf{Q}$ is a unitary matrix with

$$\mathbf{Q} = \begin{bmatrix} \frac{\mathbf{g}_{(1)}}{\|\mathbf{g}_{(1)}\|} \\ \vdots \\ \frac{\mathbf{g}_{(M)}}{\|\mathbf{g}_{(M)}\|} \end{bmatrix}. \tag{32}$$

Lemma 1: $\mathbf{R}$ is close to diagonal in the sense that its off-diagonal elements are much smaller than unity. Specifically, $|\epsilon_{ij}|$ is upper bounded by

$$|\epsilon_{ij}|^2 < \frac{(M-1)\alpha^2}{1-(M-1)\alpha^2}, \quad 1 \leq i \leq M,\ 1 \leq j \leq i-1. \tag{33}$$

Proof: See Appendix I. ■

Since $\mathbf{H}(\mathcal{S})$ is invertible by construction, the ZFBF precoder is given by

$$\mathbf{W}(\mathcal{S}_0) = \mathbf{H}(\mathcal{S}_0)^{\dagger} = \mathbf{Q}^*\mathbf{R}^{-1}\mathbf{D}^{-1}. \tag{34}$$

Performance loss due to this precoder comes from two factors: 1) since $\mathbf{R}$ is not perfectly diagonal, there would be some loss in effective channel gains and 2) since $|\mathcal{T}_i| < K$ for $i \geq 2$, the multiuser diversity gain of $\log K$ will decrease. We will see, however, that these two factors become negligible as $K \to \infty$.

[5] The choice of α is discussed in Section V.

[6] The fraction of users who survive until the end of the ith iteration is given (lower bounded) by $I_{\alpha^2}(i, M-i)$ in (39). With $\alpha = 0.3$ and $M = 4$, for example, about 25% of the users survive the first iteration, and at least 2% of the original users survive the second iteration.

1) Effective Channel Gain Reduction: Combining (11), (33), and (34), we can prove the following lemma.

Lemma 2: Under ZFBF with SUS algorithm, γ_i, the effective channel gain of the ith selected user, is lower bounded by

$$\gamma_i > \frac{\|\mathbf{g}_{(i)}\|^2}{1+\frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}}. \tag{35}$$

Proof: See Appendix II ■

From the above lemma, we see that the denominator can be made close to unity by choosing α small enough, which implies that the channel gain reduction becomes negligible if we can make α small enough. The above bound, however, is not tight.

2) Multiuser Diversity Gain Reduction: Multiuser diversity gain is related to the size of the set $\mathcal{T}_i$ from which $\mathbf{g}_{(i)}$ is chosen. To estimate the cardinality $|\mathcal{T}_{i+1}|$, we calculate the probability that a random channel $\mathbf{h}\in\mathbb{C}^{1\times M}$ is semiorthogonal to $\{\mathbf{g}_{(j)}, j=1,\ldots,i\}$. Define the set of candidate vectors of $\mathcal{T}_{i+1}$

$$\mathcal{W}_i(\alpha)=\left\{\mathbf{h}\in\mathbb{C}^{1\times M}:\frac{|\mathbf{h}\mathbf{g}_{(j)}^*|}{\|\mathbf{h}\|\|\mathbf{g}_{(j)}\|}<\alpha,\ j=1,\ldots,i\right\}. \tag{36}$$

Denote the subspace spanned by $\{\mathbf{g}_{(j)}, j=1,\ldots,i\}$ as $\mathcal{V}_i=\mathrm{span}\{\mathbf{g}_{(j)}: j=1,\ldots,i\}$, and define

$$\tilde{\mathcal{W}}_i(\alpha)=\left\{\mathbf{h}\in\mathbb{C}^{1\times M}:\frac{|\mathbf{h}\mathbf{v}^*|}{\|\mathbf{h}\|\|\mathbf{v}\|}<\alpha,\ \forall\mathbf{v}\in\mathcal{V}_i\right\}. \tag{37}$$

Clearly, $\tilde{\mathcal{W}}_i(\alpha)\subset\mathcal{W}_i(\alpha)$.

Lemma 3: For a random vector $\mathbf{h}\in\mathbb{C}^{1\times M}$, we have

$$P\{\mathbf{h}\in\tilde{\mathcal{W}}_i(\alpha)\}=F_{2i,2(M-i)}\left(\frac{M-i}{i}\frac{\alpha^2}{1-\alpha^2}\right) \tag{38}$$
$$=I_{\alpha^2}(i,M-i) \tag{39}$$

where $F_{n,m}(x)$ is the cumulative distribution function (CDF) of the F distribution, $I_z(a,b)=B_z(a,b)/B(a,b)$ is the regularized incomplete beta function, $B_z(a,b)$ is the incomplete beta function, and $B(a,b)$ is the (complete) beta function.[7]

Proof: Define $\mathbf{h}_{||}=\mathbf{h}\mathbf{v}^*\mathbf{v}/\|\mathbf{v}\|^2$ and $\mathbf{h}_\perp=\mathbf{h}-\mathbf{h}_{||}$. Then, $\mathbf{h}_\perp\mathbf{v}^*=0, \mathbf{h}_{||}\mathbf{h}_\perp^*=0$, and the condition in (37) is equivalent to $\|\mathbf{h}_{||}\|^2/\|\mathbf{h}_\perp\|^2<\alpha^2/(1-\alpha^2)$. The lemma follows by noting that $\|\mathbf{h}_{||}\|^2$ and $\|\mathbf{h}_\perp\|^2$ are independent χ^2-distributed random variables with $2i$ and $2(M-i)$ degrees of freedom, respectively, and that the ratio of two independent chi-squared random variables follows the F distribution. ■

Applying the law of large numbers, at large K, the order of $|\mathcal{T}_{i+1}|$ is approximately $|\mathcal{T}_{i+1}|\approx KP\{\mathbf{h}\in\mathcal{W}_i(\alpha)\}$. Noting that $I_{\alpha^2}(i,M-i)\geq I_{\alpha^2}(M-1,1)=\alpha^{2(M-1)}$, we have

$$|\mathcal{T}_i|\approx KP\{\mathbf{h}\in\mathcal{W}_{i-1}(\alpha)\} \tag{40}$$
$$\geq KP\left\{\mathbf{h}\in\tilde{\mathcal{W}}_{i-1}(\alpha)\right\} \tag{41}$$
$$=KI_{\alpha^2}(i-1,M-i+1) \tag{42}$$
$$\geq K\alpha^{2(M-1)}. \tag{43}$$

K and α should be chosen such that the lower bound on the right-hand side of (43) is large enough for the law of large numbers to be valid.

3) Asymptotic Performance: In Step 3) of the algorithm, $\mathbf{g}_{(i)}$ is chosen among $\mathbf{g}_k$s in $\mathcal{T}_i$. In Appendix III, we show, using results in extreme order statistics [24], [14, Appendix A], [25], that $\|\mathbf{g}_{(i)}\|^2$, being the maximum of $|\mathcal{T}_i|$ independent identically distributed (i.i.d.) random variables, behaves like $\log|\mathcal{T}_i|$ for sufficiently large K. Formally, $\|\mathbf{g}_{(i)}\|^2$ satisfies

$$P\{\|\mathbf{g}_{(i)}\|^2\geq u_L(i)\}\geq 1-O\left(\frac{1}{\log K}\right) \tag{44}$$

where

$$u_L(i)=\log\frac{KI_{\alpha^2}(i-1,M-i+1)}{(M-i)!} - (M-i)\log\log\frac{KI_{\alpha^2}(i-1,M-i+1)}{(M-i)!} - \log\log\sqrt{K}. \tag{45}$$

Choosing $P_i=\gamma_i P/M$ in (10), we have

$$\mathcal{E}\{R_{\mathrm{ZFBF}}(\mathcal{S}_0)\}$$
$$\geq\mathcal{E}\left\{\sum_{i=1}^{M}\log\left(1+\frac{\gamma_i P}{M}\right)\right\} \tag{46}$$
$$\overset{(a)}{\geq}\mathcal{E}\left\{\sum_{i=1}^{M}\log\left(1+\frac{P}{M}\frac{\|\mathbf{g}_{(i)}\|^2}{1+\frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}}\right)\right\} \tag{47}$$
$$\geq\sum_{i=1}^{M}P\{\|\mathbf{g}_{(i)}\|^2\geq u_L(i)\}\log\left(1+\frac{P}{M}\frac{u_L(i)}{1+\frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}}\right) \tag{48}$$
$$\overset{(b)}{\geq}\sum_{i=1}^{M}\left[1-O\left(\frac{1}{\log K}\right)\right]\log\left(1+\frac{P}{M}\frac{u_L(i)}{1+\frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}}\right) \tag{49}$$
$$\overset{(c)}{\sim}\sum_{i=1}^{M}\log\left(1+\frac{P}{M}\frac{\log\frac{KI_{\alpha^2}(i-1,M-i+1)}{(M-i)!}}{1+\frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}}\right) \tag{50}$$
$$\overset{(d)}{\gtrsim}M\log\left(1+\frac{P}{M}\frac{\log\frac{K\alpha^{2(M-1)}}{(M-1)!}}{1+\frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}}\right) \tag{51}$$
$$\overset{(e)}{\sim}M\log\left(1+\frac{P}{M}\log K\right) \tag{52}$$
$$=\mathcal{E}\{R_{\mathrm{DPC}}\} \tag{53}$$

where $x\gtrsim y$ means $\lim_{K\to\infty}x/y\geq 1$. Inequalities (a) and (b) follow from (35) and (44), respectively. In (c) we used the fact that $\left(1-O(1/\log K)\right)\sim 1$ and $u_L(i)\sim\log\left(KI_{\alpha^2}(i-1,M-i+1)\,/\,(M-i)!\right)$. In (d) we use (43). Finally, (e) can be verified by noting that the difference of the two expressions converges to a constant $M\log\left(1+(M-1)^4\alpha^2/(1-(M-1)\alpha^2)\right)$ as $K\to\infty$.

Since we have shown that $R_{\mathrm{ZFBF}}(\mathcal{S}_0)$ asymptotically has the same sum rate as that of DPC, and $R_{\mathrm{ZFBF}}(\mathcal{S}_0)\leq R_{\mathrm{ZFBF}}\leq R_{\mathrm{DPC}}$, Theorem 1 has been proved.

[7] In MATLAB, use $\mathrm{fcdf}(x,n,m)$ for the F distribution and $\mathrm{betainc}(z,a,b)$ for the regularized incomplete beta function.

TABLE I
COMPLEXITY COMPARISON BETWEEN DPC AND ZFBF-SUS

	DPC	ZFBF-SUS
user selection	Convex optimization problem which can be solved by sum-power iterative waterfilling algorithm with complexity $C_{\text{IW}}K$ [27]	Complexity $C_{\text{SUS}}K$, with $C_{\text{SUS}} \ll C_{\text{IW}}$.
rate and beam weight calculations	MAC-to-BC transformation, which requires a number of matrix multiplications, inversions, and singular value decompositions (SVDs) [18].	one matrix inversion and waterfilling procedure
channel coding	Dirty-paper precoding can be *approximated* using concatenated coding strategies with high complexity [28], [29].	No interference pre-subtraction is required. Conventional single user coding schemes can be applied.

D. Complexity Analysis

In this section, the complexity of the SUS algorithm is analyzed and compared with that of DPC. In DPC, the sum capacity is obtained by (4), which can be solved using standard convex optimization techniques [26]. In fact, an efficient algorithm has been developed to solve (4), utilizing a sum-power iterative waterfilling technique, whose complexity grows as $O(K)$, i.e., linearly with the number of users [27]. The solution to (4) is in the form of a dual MAC channel, so a MAC-to-BC transformation must be applied to obtain the rates, beam directions, and encoding orders of selected users, as described in [18]. Then, each user's stream is encoded using dirty paper precoding [5], where interference from previously encoded users are presubtracted. However, straightforward implementation of dirty paper precoding requires sophisticated random coding and binning strategies which are not practical to implement. Schemes to approximate dirty paper precoding in practice are described in [28] and [29]. These schemes generalize the idea of Tomlinson–Harashima precoding into a multidimensional vector quantization using a complicated concatenated coding structure. The performance of these schemes is quite close to that of DPC. The complexity of these schemes is basically that of regular LDPC-like codes plus a trellis-shaping code for the vector quantization.

Now, let us discuss the complexity of the ZFBF-SUS algorithm. As has been mentioned in Section IV-A, the implementation of ZFBF-SUS consists of two stages: user selection using the SUS algorithm and a beamforming weight vector calculation. First, we note that the latter has a small fixed complexity, requiring only one $M \times M$ matrix inversion $\mathbf{W}(\mathcal{S}_0) = \mathbf{H}(\mathcal{S}_0)^{-1}$ to obtain beamforming weights, and a single waterfilling procedure over M users, as given in (10)–(14), to calculate the optimal power allocation on each subchannel. Henceforth, we concentrate on the complexity of the SUS algorithm. In Step 2) of the SUS algorithm, we need one $(1 \times M) \times (M \times M)$ vector-matrix multiplication per user. Since there are $|\mathcal{T}_i|$ users in the ith iteration, we need $|\mathcal{T}_i|$ matrix multiplications. In Step 3) we need to search for the user with the maximum norm in $\mathcal{T}_i$, whose complexity is linear with $|\mathcal{T}_i|$. In Step 4) we need $|\mathcal{T}_i|$ inner product operations. Since Steps 1)–4) are run at most M times, each time with $|\mathcal{T}_1| = K$, $|\mathcal{T}_2|, \ldots, |\mathcal{T}_M|$ users, we conclude that the computational complexity of running the SUS algorithm is $C\sum_{i=1}^{M}|\mathcal{T}_i|$, where C is a proportionality constant that corresponds to one matrix multiplication, one vector 2-norm calculation, and one inner product. This proportionality constant is much smaller than that required in the sum-power iterative waterfilling, which is a numerical optimization technique that gives an approximate solution after a finite number of iterations (on the order of 10–30, depending on accuracy requirements). Within each of these iterations we need a number of matrix multiplications, matrix inversions, and vector 2-norm calculations for each user, as well as waterfilling over K users to obtain the optimal power allocation. Thus, the proportionality constant C_{IW} for the sum-power iterative waterfilling is much larger than C. Noting from (42) that $|\mathcal{T}_i| \approx KI_{\alpha^2}(i-1, M-i+1)$, the complexity of the SUS algorithm can be expressed as $C\sum_{i=1}^{M}|\mathcal{T}_i| \approx C_{\text{SUS}}K$, where $C_{\text{SUS}} = C\sum_{i=1}^{M} I_{\alpha^2}(i-1, M-i+1)$. For the values of α of interest, $C_{\text{SUS}} < 2C$.[8]

Although this alone gives ZFBF-SUS a large complexity reduction relative to DPC, perhaps the most important complexity disparity between ZFBF and DPC is its encoder and decoder design. Since in ZFBF the multiuser interference is zeroed out by beamforming, conventional single-user coding schemes can be used without modification. In contrast, practical implementation of DPC is still an open problem, and even approximating the DPC requires the use of complicated encoding and decoding schemes, as mentioned in the previous paragraphs. We summarize the complexity comparisons of this section in Table I.

V. NUMERICAL RESULTS

In this section, numerical results for the sum-rate performance of the ZFBF-SUS algorithm are presented. In Fig. 2, we plot $\mathcal{E}\{R_{\text{ZFBF}}(\mathcal{S}_0)\}$, the sum rate of users under the ZFBF-SUS scheme, averaged over channel distributions, as a function of α for $M = 4$, $P = 10$ dB, and K in the range of 10–10 000,[9] where $R_{\text{ZFBF}}(\mathcal{S}_0)$ is obtained from (10) and $\mathcal{S}_0$ from the SUS algorithm. If α is too large, effective channel gains (35) are reduced due to the loss associated with zero-forcing channel inversion, while if α is too small, the multiuser diversity gain (43) decreases. The optimal value of α decreases with K, ranging from 0.2–0.4 for K in the range of 100–100 000.

In Figs. 3 and 4, we compare the sum-rate performance of various MIMO BC strategies. The plots are obtained by averaging over 500 independent channel sets using the optimal α values. The plots show that the performance of the

[8] For example, $C_{\text{SUS}} = 1.48C$ with $\alpha = 0.4$, and $C_{\text{SUS}} = 1.12C$ with $\alpha = 0.2$.

[9] Although practical systems will not be able to accommodate such a large number of users, by looking at the numerical results for a very large number of users we are able to confirm the validity of Theorem 1 and its derivations in Section IV.

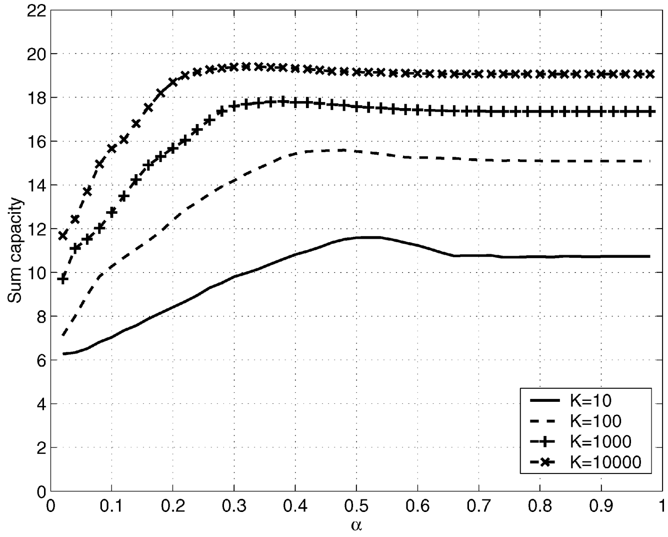

Fig. 2. Sum rate (aggregate data rate of users) of ZFBF-SUS scheme with $M = 4$ and $P = 10$ dB as a function of α. Optimal choice of α ranges 0.2–0.4 for $K \geq 100$.

ZFBF-SUS scheme closely follows the large K asymptotic performance $M \log(1 + (P/M)\log K)$. To obtain the sum capacity of DPC we use the sum-power iterative waterfilling algorithm proposed in [27]. For the optimal ZFBF, we perform an exhaustive search over the entire user set. Due to their computational complexities we could work with only up to 100 users for DPC and the optimal ZFBF. However, it can be inferred that the sum rate of the optimal ZFBF will also converge to $M \log(1 + (P/M)\log K)$, because its performance should lie between those of the ZFBF-SUS and DPC, and for DPC the convergence has already been proven elsewhere [4] as has been explained in Section III-B. Perhaps a more important observation is that the ZFBF scheme performs quite well even for relatively small values of K. This is a big contrast to the random beamforming (RBF) scheme [14], which is also asymptotically optimal at very large K but performs poorly at practical values of K, i.e., for $K \leq 100$.[10] Thus, ZFBF is not only of theoretical interest but also useful for practical systems with $M < K \ll \infty$.

Note that in the figures we are assuming that all the users have equal average received SNRs. Therefore, all the users have the same average rate, which can be obtained by dividing the sum rate by K. In particular, as K increases, the sum rate increases because we have more multiuser diversity. However, the rate of any particular user, given by the sum rate divided by K, will be a decreasing function of K. If the users have heterogeneous SNRs, maximizing the sum rate is not a good design criterion, since users with poor SNRs will experience starvation. This raises a fairness issue among users, which we will discuss in the next section.

VI. Fair Scheduling Algorithms

The schemes discussed so far focus on maximizing the sum rate; fairness among users is an important practical issue

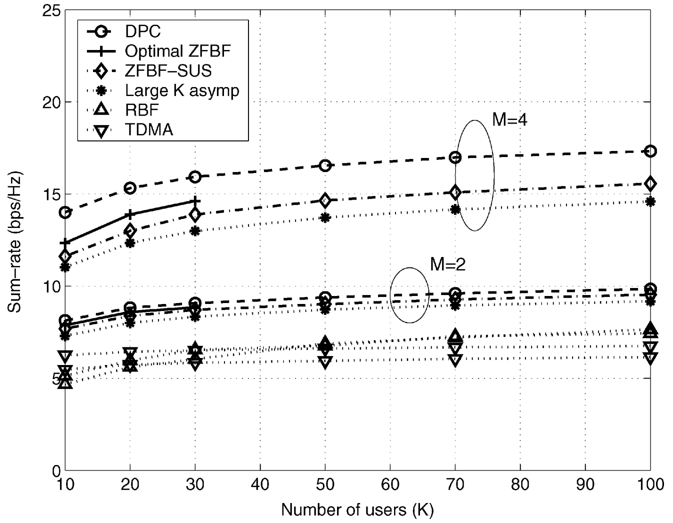

Fig. 3. Sum-rate performance comparison of DPC, ZFBF, ZFBF-SUS, TDMA, and RBF. Also, shown is large K asymptotic performance $M \log(1 + (P/M)\log K)$. Optimal values of α were used for each K. $M = 2, 4$ and $P = 10$ dB.

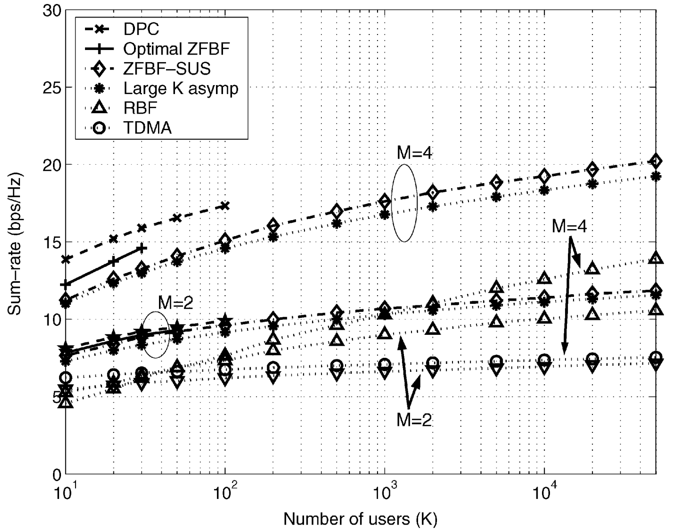

Fig. 4. Sum-rate performance comparison of DPC, ZFBF, ZFBF-SUS, TDMA, and RBF for very large K. Also, shown is large K asymptotic performance $M \log(1 + (P/M)\log K)$. Optimal values of α were used for each K. $M = 2, 4$ and $P = 10$ dB.

when the traffic is delay-constrained or the user channels are heterogeneous. Round-robin scheduling (RRS) is the simplest form of fair scheduling that gives equal opportunities to all the K users. The sum rate of RRS is given by $R_{\mathrm{RRS}} = (1/K)\sum_{k=1}^{K} \log\left(1 + P\|\mathbf{h}_k\|^2\right)$. However, since RRS supports only one user at a time in a TDMA fashion, it does not achieve spatial multiplexing gains. Therefore, regarding R_{RRS} as the performance baseline, we would like to develop a scheduling strategy for our ZFBF technique with similar fairness. We propose two different fair scheduling methods: RR-ZFBF and p PF-ZFBF.

A. RR-ZFBF

In RR-ZFBF, we recursively apply the SUS algorithm. Specifically, we construct a user group $\mathcal{S}_1 = \mathcal{S}_0$ by running the algorithm. Then, we construct the second user group $\mathcal{S}_2$ by

[10] The random beamforming scheme, though, requires a much smaller amount of feedback.

repeating the algorithm for the remaining users, i.e., with an initial user set $\mathcal{T}_1 = \{1, \ldots, K\} - \mathcal{S}_1$. The tth user group $\mathcal{S}_t$ is obtained with $\mathcal{T}_1 = \{1, \ldots, K\} - \bigcup_{i=1}^{t-1} \mathcal{S}_i$. This procedure is repeated until no users are left. Let T be the total number of user groups. A scheduling period consists of T time slots. At the tth time slot, the base station transmits to users in $\mathcal{S}_t$ using a ZFBF precoder $\mathbf{W}(\mathcal{S}_t) = \mathbf{H}(\mathcal{S}_t)^{\dagger}$. Since these users are semiorthogonal, the performance loss due to zero-forcing channel inversion is minimal. Assuming an average power constraint over the scheduling block, we assign fair (equal) powers PT/K to each user. Then, the sum rate of RR-ZFBF is given by

$$R_{\text{RR-ZFBF}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{S}_t|} \log\left(1 + \frac{PT}{K}\gamma_{t,i}\right) \tag{54}$$

where $\gamma_{t,i} = 1/\left[(\mathbf{H}(\mathcal{S}_t)\mathbf{H}(\mathcal{S}_t)^*)^{-1}\right]_{i,i}$. This scheduling scheme essentially combines TDMA and SDMA.

To see how fairness is preserved while giving a larger throughput than RRS, we obtained numerical results for $M = 4$ and $K = 1000$. For RR-ZFBF, there were a total of $T = 257$ user groups, of which 239 groups had four users, 11 groups had three users, four groups had two users, and three groups had one user. Therefore, most of the time, the transmitter could serve $M = 4$ semiorthogonal users simultaneously. The sum rate was 11.9 b/s/Hz at $P = 10$ dB, which was nearly M times RRS's sum rate of 3.3 b/s/Hz at $P/M = 4$ dB SNR, hence, our scheme obtained significant spatial multiplexing gain. In fact, the individual rate for each user was increased by similar amounts (mostly by 3.4 – 4.0), meaning that fairness in the RRS scheme was carried over to RR-ZFBF.

B. PF-ZFBF

A wide class of scheduling problems can be formulated as a weighted sum-rate maximization problem

$$\max_{\substack{\mathcal{S}\subset\{1,\ldots,K\} \\ P_k \geq 0, \sum_{k=1}^{K} P_k = P}} \sum_{k=1}^{K} \mu_k(t) R_k(\mathcal{S}, t) \tag{55}$$

where $\mu_k(t)$ and $R_k(\mathcal{S}, t)$ are the weight and the supported data rate, respectively, of user k at time t, with a scheduling decision $\mathcal{S}$. When $\mu_k(t) = 1$, $\forall k, t$, the problem reduces to the sum-rate maximization criterion discussed in the previous sections. In general, $\mu_k(t)$ can be chosen based on various criteria such as queue lengths [21], [15] for stability or average past throughput [12] for fairness.

Proportional fair scheduling (PFS) is a simple algorithm designed to meet fairness among users while at the same time exploiting the multiuser diversity gain [12]. Specifically, in the original PFS, the base station schedules to the single-user $\mathcal{S} = \{k_{\text{opt}}\}$ with the maximum weighted throughput

$$k_{\text{opt}} = \arg\max_{k\in\{1,\ldots,K\}} \mu_k(t) R_k(\{k\}, t) \tag{56}$$

where the supported data rate $R_k(\{k\}, t)$ is

$$R_k(\{k\}, t) = \log\left(1 + P\|\mathbf{h}_k(t)\|^2\right) \tag{57}$$

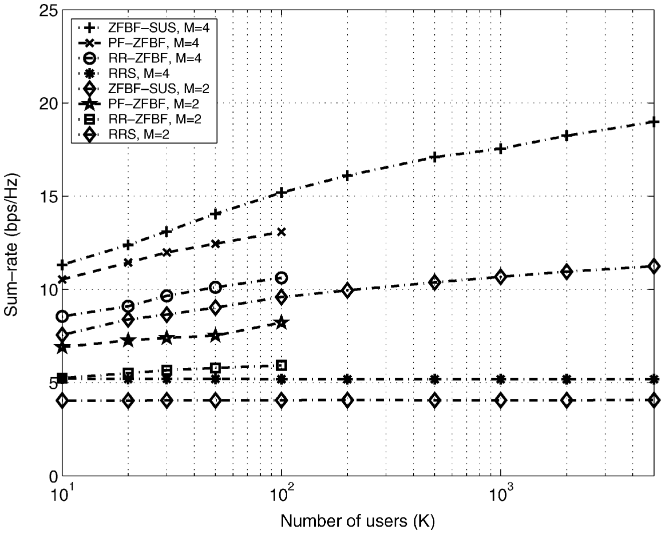

Fig. 5. Sum-rate performance comparison of fair scheduling strategies: PF-ZFBF, RR-ZFBF, and RRS. For comparison the sum-rate of ZFBF-SUS is shown together. Optimal values of α were used for each K. $M = 2$, 4 and $P = 10$ dB.

and the weights, the inverse of the time-averaged past throughput, are updated as

$$\frac{1}{\mu_k(t+1)} = \left(1 - \frac{1}{t_c}\right)\frac{1}{\mu_k(t)} + \frac{1}{t_c}R_k(\mathcal{S}, t), \quad k \in \mathcal{S} \tag{58}$$

$$\frac{1}{\mu_k(t+1)} = \left(1 - \frac{1}{t_c}\right)\frac{1}{\mu_k(t)}, \quad k \notin \mathcal{S} \tag{59}$$

with the averaging window size t_c appropriately chosen.

In PF-ZFBF we extend the above PFS to serve multiple users using ZFBF-SUS. One subtlety in applying the PFS in ZFBF-SUS is that, unlike the original PFS where $R_k(\mathcal{S}, t)$ can be obtained before the scheduling decision is made, in ZFBF-SUS the supported data rates can only be calculated after the user group $\mathcal{S}$ is completely selected. Therefore, in PF-ZFBF, the exact $R_k(\mathcal{S}, t)$ is unknown when the scheduling decision has to be made. Noting, however, that the selected users are semiorthogonal, we may approximate the supported data rate as $\tilde{R}_k(\mathcal{S}, t) = \log\left(1 + (P/M)\|\mathbf{g}_k\|^2\right)$. If the selected users were truly orthogonal and allocated equal powers, we would have $R_k(\mathcal{S}, t) = \tilde{R}_k(\mathcal{S}, t)$. With this approximation, the PF-ZFBF algorithm works as follows.

Step 1) At time t, perform the SUS algorithm to obtain $\mathcal{S} = \mathcal{S}_0$, with the following modification in (22):

$$\pi(i) = \arg\max_{k\in\mathcal{T}_i} \mu_k(t) \log\left(1 + \frac{P}{M}\|\mathbf{g}_k\|^2\right) \tag{60}$$

Step 2) Apply ZFBF to $\mathbf{H}(\mathcal{S})$ to obtain the actual supported data rates $R_k(\mathcal{S}, t)$ of each user $k \in \mathcal{S}$. $\mu_k(t+1)$s are updated as in(58) for $k \in \mathcal{S}$ and as in (59) for $k \notin \mathcal{S}$, using the actual $R_k(\mathcal{S}, t)$s.

C. Performance

The sum rates of RR-ZFBF and PF-ZFBF are compared in Fig. 5. As in Figs. 3–4, all the users were assumed to have the same average SNR and experience independent Rayleigh fading. Compared to the sum-rate maximizing scheduling

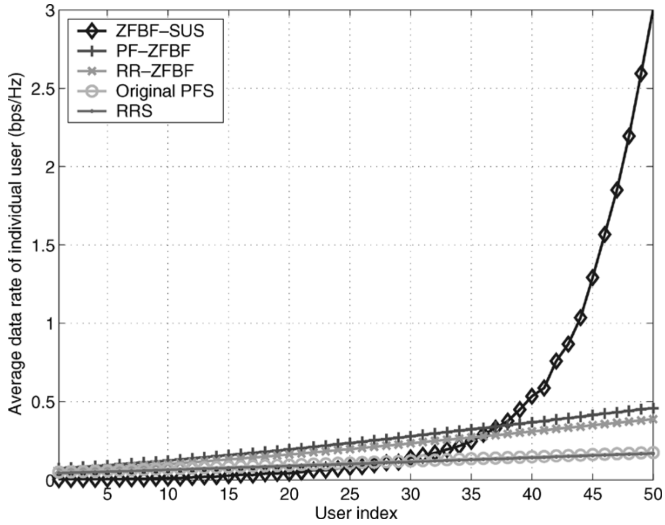

Fig. 6. Fairness comparison of ZFBF-SUS, PF-ZFBF, RR-ZFBF, original PFS, and RRS. Users have unequal average received SNRs ranging from 0 (user 1) to 20 dB (user 50).

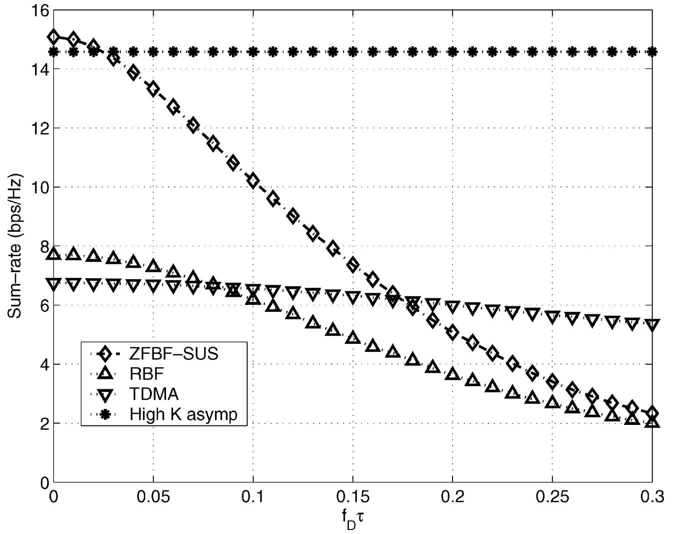

Fig. 7. Sum rate of ZFBF-SUS, RBF, and TDMA with $M = 4$ and $K = 100$ as a function of Doppler-delay product. Average received SNR is $P = 10$ dB.

(ZFBF-SUS), both fair scheduling algorithms suffer rate loss in return for fairness. Among the two schemes, PF-ZFBF is seen to have a better throughput performance than RR-ZFBF. This is because PF-ZFBF benefits from multiuser diversity gain as it tries to choose the best user set through (60), while RR-ZFBF only cares about the orthogonality of user channels and does not exploit fluctuations in channel strengths. Another shortcoming of RR-ZFBF is that the channel must remain unchanged during the entire scheduling period of T time slots. RR-ZFBF, however, provides deterministic fairness, i.e., every user is guaranteed to be scheduled once in every scheduling period. Since a scheduling period consists of T time slots, the worst case delay between consecutive packets is $2T \approx 2K/M$ time slots. Therefore, RR-ZFBF is more suited for delay-constrained traffic. In PF-ZFBF there is no such delay guarantee, although a very high inter-packet delay is unlikely to occur.

To compare fairness of the proposed and other scheduling strategies, in Fig. 6, we plot the time-average data rate that each individual user attains under each scheduling strategy. We use $K = 50$ users with their average received SNRs ranging from 0 to 20 dB, in a log-linear scale. From the figure, we can clearly observe that the proposed fair scheduling schemes achieve similar fairness as RRS, i.e., users have equal chances of being scheduled regardless of their SNRs, and yet all the users have higher throughput than RRS. ZFBF-SUS, on the other hand, strongly favors users with high SNRs and thus causes starvation of users with low SNRs.

VII. Imperfect Channel Knowledge

The discussions so far assume perfect CSI at the transmitter (CSIT). Since this is difficult to obtain in practice, it is important to investigate the effect of imperfect CSIT on the performance of the proposed ZFBF scheme. For single-user MIMO channels, various types of imperfect CSIT, including covariance [30], delayed [31], and quantized feedback [32], have been considered in the literature. Since analytical formulations of these effects are difficult for MIMO BCs with SDMA, in this section we present a preliminary simulation result on the sensitivity of the sum rate to imperfect (outdated) CSIT. Fig. 7 shows the sum rate of ZFBF-SUS, RBF, and TDMA under imperfect CSIT. The plots are generated assuming that the transmitter has an outdated channel knowledge $\{\hat{\mathbf{H}}_k\}$, and its correlation coefficient to the true channel $\{\mathbf{H}_k\}$ is given by $\rho = J_0(2\pi f_D \tau)$, where J_0 is a Bessel function of the first kind of order 0, f_D is the Doppler shift, and τ is the feedback delay. Inaccuracy in CSIT destroys semiorthogonality of users and results in a poor performance of ZFBF. The RBF scheme is also sensitive to outdated CSIT, since the transmitter's beam directions are no longer matched to the targeted users. TDMA is relatively robust to imperfect CSIT because only one data stream is transmitted at one time. Thus, we conclude that high channel accuracy at the transmitter is required for schemes that use multiple beams at the same time.

VIII. Multiple Receive Antennas

While we have proven the asymptotic optimality of ZFBF for only the case with $N_k = 1$, it is not difficult to see that Theorem 1 also holds for general MIMO configurations. Suppose the kth user has $N_k \geq 1$ receive antennas. As an extension to Lemma 3 in [4], we can show that

$$\mathcal{E}\left\{R_{\mathrm{DPC}}\right\} \sim M \log\left(1 + \frac{P}{M} \log \sum_{k=1}^{K} N_k\right). \tag{61}$$

Now, consider the receiver strategy where the N_k receive antennas do not coordinate. In this case, each antenna may be treated as a separate user. Then, since there are effectively $\sum_{k=1}^{K} N_k$ single antenna users, by Theorem 1, ZFBF achieves

$$\mathcal{E}\left\{R_{\mathrm{ZFBF}}\right\} \sim M \log\left(1 + \frac{P}{M} \log \sum_{k=1}^{K} N_k\right). \tag{62}$$

Thus, we again have $\mathcal{E}\left\{R_{\mathrm{ZFBF}}\right\} \sim \mathcal{E}\left\{R_{\mathrm{DPC}}\right\}$, regardless of the number of receive antennas of each user.

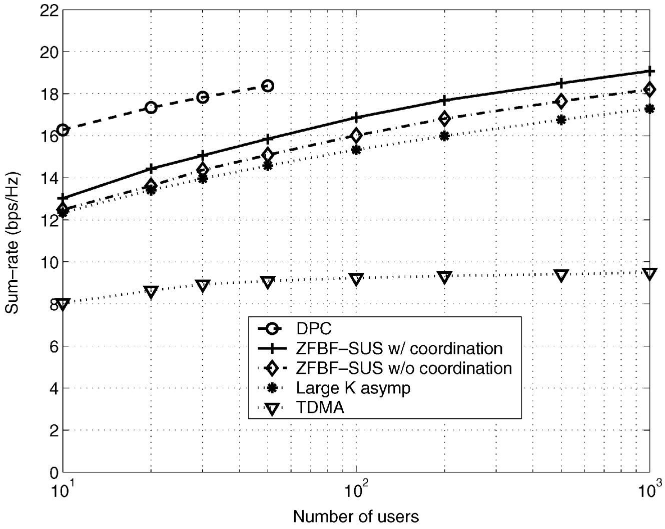

Fig. 8. Sum-rate performance comparison of DPC, ZFBF-SUS with and without receiver coordination, and TDMA. Also, shown is the large K asymptotic performance. $M = 4$, $N = 2$, and $P = 10$ dB.

Although we have shown that the optimal asymptotic performance can be attained without any receiver antenna coordination, coordinating receiver antennas through receiver processing is still beneficial [33]. The receiver processing strategy we use in this section is as follows. Let the singular value decomposition (SVD) of $\mathbf{H}_k$ be $\mathbf{H}_k = \mathbf{U}_k\sqrt{\mathbf{\Lambda}_k}\mathbf{V}_k^*$. The received signal $\mathbf{y}_k$ of each user is multiplied by a receiver shaping matrix $\mathbf{U}_k^*$ [23]. Then, the resulting system can be viewed as a MIMO BC with $\sum_{k=1}^{K} N_k$ single antenna users with channel gains $\sqrt{\lambda_{k,n}}\mathbf{v}_{k,n}^*$ ($\mathbf{v}_{k,n}$ is the nth column of $\mathbf{V}_k$) at the nth stream to the kth user, and thus our ZFBF-SUS scheme can be applied. Fig. 8 compares the sum-rate performances of DPC, TDMA, and ZFBF-SUS with and without receiver coordination. It can be seen that ZFBF with receiver coordination has a slightly better performance than ZFBF without receiver coordination. This is because the receiver coordination presents the transmitter with orthogonal input singular directions ($\mathbf{v}_{k,n}$, $n = 1, \ldots, N_k$), as well as the largest possible channel gain ($\lambda_{k,\max}$) to each user.

IX. Conclusion

We have examined transmit strategies for MIMO BCs with multiple antennas at the base station and a large number of multiantenna users. We have shown that a ZFBF strategy can achieve the same asymptotic sum rate (the aggregate data rate of users) as that of the optimal DPC scheme as the number of users goes to infinity. This is because with a large number of users the transmitter can choose user channels that are nearly orthogonal to one another. We have proposed a low-complexity algorithm for such a semiorthogonal user selection (SUS algorithm). Our proposed ZFBF-SUS scheme thus achieves a sum rate close to the optimal rate promised by DPC, but with much lower complexity. We have also presented fair scheduling schemes based on the ZFBF-SUS algorithm, RR-ZFBF and PF-ZFBF. Numerical results show that ZFBF-SUS (and thus ZFBF) is indeed asymptotically optimal and has fairly good performance for a relatively small number of users. However, its performance suffers under imperfect transmitter channel knowledge. When users are equipped with multiple antennas, receiver antenna coordination further enhances the data rates, although the coordination is not necessary to achieve the asymptotically optimal sum rate.

Although we have discussed only BCs, all the results in this paper are directly applicable to a MAC with a multiantenna receiver and many users (transmitters) with a *common* power source (*sum-power* constraint), with the transmit zero-forcing beamformer in the BC moved to the receiver side in the MAC. By the uplink–downlink duality result [19] in linear beamforming systems and the duality relationship between the capacity (DPC) regions of MIMO-MAC and MIMO-BC [18]–[20], the sum rates of the MAC using DPC and ZFBF, respectively, are the same as the sum rates of the dual BC. Hence, ZFBF achieves the asymptotically optimal sum rate in the sum-power constrained MAC as well.

For most of this paper, we have only investigated users with equal average SNRs. The case of unequal SNR users would be an important extension of the paper. Moreover, although ZFBF is asymptotically optimal in the sum-rate sense, this may not be the case for the weighted sum-rate maximization problem shown in (55), for which the performance gap of ZFBF and DPC needs to be further explored.

Appendix I
Proof of Lemma 1

Suppose $i > j$, i.e., the ith user channel $\mathbf{h}_{(i)}$ is selected after the jth user channel $\mathbf{h}_{(j)}$. Then, by Step 4) of the SUS algorithm, $\mathbf{h}_{(i)}$ must be semiorthogonal to $\mathbf{g}_{(j)}$

$$\frac{|\mathbf{h}_{(i)}\mathbf{g}_{(j)}^*|}{\|\mathbf{h}_{(i)}\|\|\mathbf{g}_{(j)}\|} < \alpha, \qquad \text{for} \quad i > j. \tag{63}$$

Using our definition $h_{ij} = \mathbf{h}_{(i)}\mathbf{g}_{(j)}^*/\|\mathbf{g}_{(j)}\|$, we obtain

$$\begin{aligned}|h_{ij}| &= \frac{|\mathbf{h}_{(i)}\mathbf{g}_{(j)}^*|}{\|\mathbf{g}_{(j)}\|} = \frac{|\mathbf{h}_{(i)}\mathbf{g}_{(j)}^*|}{\|\mathbf{h}_{(i)}\|\|\mathbf{g}_{(j)}\|}\|\mathbf{h}_{(i)}\| \\ &< \alpha\|\mathbf{h}_{(i)}\|, \qquad \text{for} \quad i > j. \end{aligned} \tag{64}$$

Taking squares on both sides, and substituting $\|\mathbf{h}_{(i)}\|^2$ with (29), we have

$$|h_{ij}|^2 < \alpha^2\|\mathbf{h}_{(i)}\|^2 \tag{65}$$

$$= \alpha^2 \left| \mathbf{g}_{(i)} + \sum_{k=1}^{i-1} h_{ik} \frac{\mathbf{g}_{(k)}}{\|\mathbf{g}_{(k)}\|} \right|^2 \tag{66}$$

$$= \alpha^2 \left(\|\mathbf{g}_{(i)}\|^2 + \sum_{k=1}^{i-1} |h_{ik}|^2 \right), \qquad \text{for} \quad i > j \tag{67}$$

where the last equality results from the orthogonality among $\mathbf{g}_{(i)}$ and $\mathbf{g}_{(k)}$s. Taking the summation over $1 \le j \le i-1$, we can write

$$\sum_{j=1}^{i-1} |h_{ij}|^2 < \sum_{j=1}^{i-1} \alpha^2 \left(\|\mathbf{g}_{(i)}\|^2 + \sum_{k=1}^{i-1} |h_{ik}|^2 \right) \tag{68}$$

$$= (i-1)\alpha^2 \left(\|\mathbf{g}_{(i)}\|^2 + \sum_{k=1}^{i-1} |h_{ik}|^2 \right) \tag{69}$$

$$= (i-1)\alpha^2 \left(\|\mathbf{g}_{(i)}\|^2 + \sum_{j=1}^{i-1} |h_{ij}|^2 \right). \tag{70}$$

Rearranging the above, we obtain

$$\sum_{j=1}^{i-1} |h_{ij}|^2 < \frac{(i-1)\alpha^2 \|\mathbf{g}_{(i)}\|^2}{1-(i-1)\alpha^2}. \tag{71}$$

Using this result, we can obtain an upper bound on $|\epsilon_{ij}| = |h_{ij}|/\|\mathbf{g}_{(i)}\|$, $i > j$, as

$$|\epsilon_{ij}|^2 = \frac{|h_{ij}|^2}{\|\mathbf{g}_{(i)}\|^2} \le \frac{\sum_{j=1}^{i-1} |h_{ij}|^2}{\|\mathbf{g}_{(i)}\|^2} \tag{72}$$

$$< \frac{(i-1)\alpha^2}{1-(i-1)\alpha^2} \tag{73}$$

$$\le \frac{(M-1)\alpha^2}{1-(M-1)\alpha^2} \tag{74}$$

which is the desired result.

Appendix II
Proof of Lemma 2

From (11), γ_i may be expressed as

$$\gamma_i = \frac{1}{\left[(\mathbf{H}(\mathcal{S}_0)\mathbf{H}(\mathcal{S}_0)^*)^{-1}\right]_{i,i}} \tag{75}$$

$$= \frac{1}{\left[(\mathbf{DRQQ}^*\mathbf{R}^*\mathbf{D}^*)^{-1}\right]_{i,i}} \tag{76}$$

$$= \frac{\|\mathbf{g}_{(i)}\|^2}{\left((\mathbf{R}^*)^{-1}\mathbf{R}^{-1}\right)_{i,i}}. \tag{77}$$

To find $\mathbf{R}^{-1}$, write $\mathbf{R} = \mathbf{I} + \mathbf{E}$. Note that $\mathbf{E}$ is lower triangular with zeros in the diagonal. Observing $\mathbf{E}^M = \mathbf{0}$, $\mathbf{R}^{-1}$ can be found by

$$\mathbf{R}^{-1} = \sum_{n=0}^{M-1} (-\mathbf{E})^n. \tag{78}$$

By induction, we will show that

$$|(\mathbf{E}^n)_{ij}| < \sqrt{\frac{(M-1)\alpha^2}{1-(M-1)\alpha^2}}, \quad \text{for } 1 \le n \le M-1 \text{ and } i > j \tag{79}$$

for α small enough. This is trivially true for $n = 1$ by (33). Assuming that (79) is true for some $n < M-1$, we have

$$|(\mathbf{E}^{n+1})_{ij}| = \left| \sum_{k=1}^{M} \mathbf{E}_{ik} (\mathbf{E}^n)_{kj} \right| \tag{80}$$

$$\le \sum_{k=1}^{M} |\mathbf{E}_{ik}|\, |(\mathbf{E}^n)_{kj}| \tag{81}$$

$$< M \frac{(M-1)\alpha^2}{1-(M-1)\alpha^2}. \tag{82}$$

Choosing α such that $M < \sqrt{(1-(M-1)\alpha^2)/(M-1)\alpha^2}$, or equivalently $\alpha < \sqrt{1/(M^2+1)(M-1)}$, we obtain (79). Now, $|(\mathbf{R}^{-1})_{ij}|$ for $i > j$ can be upper bounded as

$$|(\mathbf{R}^{-1})_{ij}| = \left| \sum_{k=1}^{M-1} [(-\mathbf{E})^k]_{ij} \right| \tag{83}$$

$$\le \sum_{k=1}^{M-1} |(\mathbf{E}^k)_{ij}| \tag{84}$$

$$< \sqrt{\frac{(M-1)^3\alpha^2}{1-(M-1)\alpha^2}}. \tag{85}$$

Also, note that $(\mathbf{R}^{-1})_{ij} = 0$ for $i < j$ and $(\mathbf{R}^{-1})_{ij} = 1$ for $i = j$. With this bound, the denominator in (77) may be bounded by

$$\left[(\mathbf{R}^*)^{-1}\mathbf{R}^{-1}\right]_{i,i} = \sum_{k=1}^{M} |(\mathbf{R}^{-1})_{ki}|^2 \tag{86}$$

$$= |(\mathbf{R}^{-1})_{ii}|^2 + \sum_{k=1, k\neq i}^{M} |(\mathbf{R}^{-1})_{ki}|^2 \tag{87}$$

$$< 1 + \frac{(M-1)^4\alpha^2}{1-(M-1)\alpha^2}. \tag{88}$$

The lemma follows by substituting this into (77).

Appendix III
Proof of (44)–(45)

Consider $i-1$ randomly chosen orthogonal vectors $\hat{\mathbf{g}}_{(1)}, \ldots, \hat{\mathbf{g}}_{(i-1)} \in \mathbb{C}^{1\times M}$, and define

$$\hat{\mathcal{W}}_{i-1}(\alpha) = \left\{ \mathbf{h} \in \mathbb{C}^{1\times M} : \frac{|\mathbf{h}\mathbf{v}^*|}{\|\mathbf{h}\|\|\mathbf{v}\|} < \alpha, \quad \forall \mathbf{v} \in \hat{\mathcal{V}}_{i-1} \right\} \tag{89}$$

where $\hat{\mathcal{V}}_{i-1} = \mathrm{span}\{\hat{\mathbf{g}}_{(j)} : j = 1, \ldots, i-1\}$. Using a similar procedure as in the proof of [25, Lemma 2], we can show that

$$P\left\{\|\mathbf{g}_{(i)}\|^2 < z\right\} \le P\left\{\|\hat{\mathbf{g}}_{(i)}\|^2 < z\right\}, \quad \forall z > 0 \tag{90}$$

where

$$\hat{\mathbf{g}}_{(i)} = \hat{\mathbf{g}}_{\hat{\pi}(i)} \tag{91}$$

$$\hat{\pi}(i) = \arg\ i\text{th}\max_{k:\mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha)} \|\hat{\mathbf{g}}_k\| \tag{92}$$

$$\hat{\mathbf{g}}_k = \mathbf{h}_k - \sum_{j=1}^{i-1} \frac{\mathbf{h}_k \hat{\mathbf{g}}_{(j)}^*}{\|\hat{\mathbf{g}}_{(j)}\|^2} \hat{\mathbf{g}}_{(j)}, \quad k = 1, \ldots, K. \tag{93}$$

That is, $\hat{\mathbf{g}}_{(i)}$ is defined as the projected channel (projected away from $\hat{\mathbf{g}}_{(1)}, \ldots, \hat{\mathbf{g}}_{(i-1)}$) of the user whose projected channel norm is the ith largest among those whose original channels belong to $\hat{\mathcal{W}}_{i-1}(\alpha)$. Now, define

$$\tilde{\mathbf{g}}_k = \begin{cases} \hat{\mathbf{g}}_k, & \mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha) \\ 0, & \mathbf{h}_k \notin \hat{\mathcal{W}}_{i-1}(\alpha) \end{cases}. \tag{94}$$

With this definition it is clear that $\hat{\pi}(i) = \arg i\text{th}\max_{1\le k\le K} \|\tilde{\mathbf{g}}_k\|$, i.e., $\hat{\mathbf{g}}_{(i)}$ is the ith largest order statistic of $\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_k$. Conditioning on $\mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha)$, the complementary CDF of $\|\tilde{\mathbf{g}}_k\|^2$ is given by

$$\begin{aligned} &P\{\|\tilde{\mathbf{g}}_k\|^2 \geq z\} \\ &= P\left\{\mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha)\right\} P\left\{\|\hat{\mathbf{g}}_k\|^2 \geq z \middle| \mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha)\right\}. \quad (95) \end{aligned}$$

Noting that $\mathbf{h}_k = \hat{\mathbf{g}}_k + (\mathbf{h}_k - \hat{\mathbf{g}}_k)$, and that $\|\hat{\mathbf{g}}_k\|^2$ and $\|\mathbf{h}_k - \hat{\mathbf{g}}_k\|^2$ are independent chi-square random variables with $2(M-i+1)$ and $2(i-1)$ degrees of freedom, respectively, we have

$$\begin{aligned} \mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha) &\Leftrightarrow \frac{\|\mathbf{h}_k - \hat{\mathbf{g}}_k\|}{\|\mathbf{h}_k\|} < \alpha && (96) \\ &\Leftrightarrow \|\hat{\mathbf{g}}_k\|^2 > \frac{1-\alpha^2}{\alpha^2} \|\mathbf{h}_k - \hat{\mathbf{g}}_k\|^2 && (97) \\ &\Leftrightarrow \|\hat{\mathbf{g}}_k\|^2 > \frac{1-\alpha^2}{\alpha^2} \chi^2_{2(i-1)}. && (98) \end{aligned}$$

Therefore

$$\begin{aligned} &P\left\{\|\hat{\mathbf{g}}_k\|^2 \geq z \middle| \mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha)\right\} \\ &= P\left\{\|\hat{\mathbf{g}}_k\|^2 \geq z \middle| \|\hat{\mathbf{g}}_k\|^2 > \frac{1-\alpha^2}{\alpha^2} \chi^2_{2(i-1)}\right\} && (99) \\ &> P\left\{\|\hat{\mathbf{g}}_k\|^2 \geq z\right\} && (100) \end{aligned}$$

and (95) becomes

$$\begin{aligned} &P\left\{\left\|\tilde{\mathbf{g}}_k\right\|^2 \geq z\right\} \\ &> P\left\{\mathbf{h}_k \in \hat{\mathcal{W}}_{i-1}(\alpha)\right\} P\{\|\hat{\mathbf{g}}_k\|^2 \geq z\} && (101) \\ &= I_{\alpha^2}(i-1, M-i+1)\, e^{-z} \sum_{j=0}^{M-i} \frac{z^j}{j!} && (102) \\ &= I_{\alpha^2}(i-1, M-i+1)\, \frac{e^{-z} z^{M-i}}{(M-i)!} \left(1 + O(z^{-1})\right). && (103) \end{aligned}$$

Utilizing [25, Lemma 6], we have

$$P\left\{\|\hat{\mathbf{g}}_{(i)}\|^2 \geq u_L(i)\right\} \geq 1 - O\left(\frac{1}{\log K}\right) \quad (104)$$

where

$$\begin{aligned} u_L(i) = \log &\left(\frac{K I_{\alpha^2}(i-1, M-i+1)}{(M-i)!}\right) \\ &- (M-i) \log\log \left(\frac{K I_{\alpha^2}(i-1, M-i+1)}{(M-i)!}\right) \\ &- \log\log \sqrt{K} \Bigg\}. \quad (105) \end{aligned}$$

Finally, (44) is proved by noting the inequality in (90).

Acknowledgment

The authors would like to thank the editor and anonymous reviewers for their thoughtful reviews and M. A. Maddah-Ali for valuable comments on the proof in Appendix III.

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] N. Jindal and A. Goldsmith, "Dirty paper coding vs. TDMA for MIMO broadcast channels," in *Proc. IEEE Int. Conf. Commun.*, vol. 2, Jun. 2004, pp. 682–686.

[4] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Trans. Commun.*, submitted for publication.

[5] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, pp. 439–441, May 1983.

[6] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Chicago, IL, 2004.

[7] S. Venkatesan and H. Huang, "System capacity evaluation of multiple antenna systems using beamforming and dirty paper coding," Bell Labs.

[8] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1691–1706, Jul. 2003.

[9] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known-interference cancellation," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 6, pp. 802–811, Jun. 2003.

[10] B. Hochwald and S. Vishwanath, "Space-time multiple access: Linear growth in the sum rate," in *Proc. 40th Annual Allerton Conf. Commun., Control, Comput.*, Allerton, IL, Oct. 2002.

[11] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE Int. Conf. Commun.*, vol. 1, Jun. 1995, pp. 331–335.

[12] 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.

[13] R. W. Heath, Jr., M. Airy, and A. J. Paulraj, "Multiuser diversity for MIMO wireless systems with linear receivers," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2001, pp. 1194–1199.

[14] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side infonnation," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[15] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel," in *Proc. Allerton Conf. Commun., Control, Comput.*, Allerton, IL, Oct. 2004.

[16] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[17] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 51, no. 6, pp. 684–702, Jun. 2003.

[18] 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, no. 10, pp. 2658–2668, Oct. 2003.

[19] P. Viswanath and D. N. C. 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.

[20] W. Yu and J. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.

[21] H. Viswanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink wireless systems," in *Proc. Allerton Conf. Commun., Control, Comput.*, Allerton, IL, 2001.

[22] Z. Tu and R. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Commun. Lett.*, vol. 7, no. 8, pp. 370–372, Aug. 2003.

[23] R. Zhang, Y. C. Liang, and J. Cioffi, "Throughput comparison of wireless downlink transmission schemes with multiple antennas," in *Proc. IEEE Int. Conf. Commun.*, 2005.

[24] H. A. David, *Order Statistics*. New York: Wiley, 1980.

[25] M. A. Maddah-Ali, M. Ansari, and A. K. Khandani, "An efficient signaling scheme for MIMO broadcast systems: Design and performance evaluation," *IEEE Trans. Inf. Theory*. [Online]. Available: http://www.cst.uwaterloo.ca/pub_tech_rep.html.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[27] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[28] W. Yu, D. Varodayan, and J. Cioffi, "Trellis and convolutional precoding for transmitter-based interference presubtraction," *IEEE Trans. Commun.*, vol. 53, no. 7, pp. 1220–1230, Jul. 2005.

[29] U. Erez and S. Brink, “Approaching the dirty paper limit for canceling known interference,” in *Proc. Allerton Conf. Commun., Control, Comput.*, Allerton, IL, Oct. 2003.

[30] S. A. Jafar and A. Goldsmith, “Transmitter optimization and optimality of beamforming for multiple antenna systems,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 7, pp. 1165–1175, Jul. 2004.

[31] S. Venkatesan, S. Simon, and R. Valenzuela, “Capacity of a Gaussian MIMO channel with nonzero mean,” in *Proc. IEEE Veh. Technol. Conf.*, vol. 3, Oct. 2003, pp. 1767–1771.

[32] D. J. Love, R. W. Heath, Jr., and T. Strohmer, “Grassmannian beamforming for multiple-input multiple-output wireless systems,” *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[33] Q. Spencer, A. L. 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.

Taesang Yoo (S’04) received the B.S. degree (Hons.) in electrical engineering from Seoul National University, Seoul, Korea, in 1998, and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 2003. He is currently working towards the Ph.D. degree at Stanford University.

He was a Summer Intern in the Wireless Communications Group, Lucent Bell Laboratories, Holmdel, NJ, in 2005. In Summer 2002, he was an Intern at Qualcomm, Campbell, CA. From 2000 to 2001, he was with Xeline, Seoul, Korea, where he worked on the design of powerline communications chipsets. He was also an Engineer at DSI, Seoul, Korea, from 1998 to 2000. His research interests include multiple antenna systems, wireless ad hoc networks, and communication theory.

Andrea Goldsmith (S’90–M’93–SM’99–F’05) received the B.S., M.S., and Ph.D. degrees in electrical engineering from University of California at Berkeley.

She is an Associate Professor of Electrical Engineering at Stanford University, Stanford, CA, and was previously an Assistant Professor of Electrical Engineering at California Institute of Technology, Pasadena. She has also held industry positions at Maxim Technologies and AT&T Bell Laboratories. Her research includes work on the capacity of wireless channels and networks, wireless communication and information theory, adaptive resource allocation in wireless networks, multiantenna wireless systems, energy-constrained wireless communications, wireless communications for distributed control, and cross-layer design for cellular systems, ad hoc wireless networks, and sensor networks.

Dr. Goldsmith is a Fellow of Stanford University, and currently holds Stanford’s Bredt Faculty Development Scholar Chair. She has received several awards for her research, including the National Academy of Engineering Gilbreth Lectureship, the Alfred P. Sloan Fellowship, the Stanford Terman Fellowship, the National Science Foundation CAREER Development Award, and the Office of Naval Research Young Investigator Award. She was also a corecipient of the 2005 IEEEE Communications Society and Information Theory Society Joint Paper Award. She is currently an Editor for the *Journal on Foundations and Trends in Communications and Information Theory and in Networks*, and was previously an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, and for the *IEEE Wireless Communications Magazine*. She is also very active in the IEEE and currently serves on the Board of Governors for both the Information Theory and Communications Societies.