Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 479357, 17 pages
doi:10.1155/2008/479357

Research Article

Uplink SDMA with Limited Feedback: Throughput Scaling

Kaibin Huang, Robert W. Heath Jr., and Jeffrey G. Andrews

Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-0240, USA

Correspondence should be addressed to Kaibin Huang, huangkb@mail.utexas.edu

Received 15 June 2007; Accepted 23 October 2007

Recommended by Christoph F. Mecklenbräuker

Combined space division multiple access (SDMA) and scheduling exploit both spatial multiplexing and multiuser diversity, increasing throughput significantly. Both SDMA and scheduling require feedback of multiuser channel sate information (CSI). This paper focuses on uplink SDMA with limited feedback, which refers to efficient techniques for CSI quantization and feedback. To quantify the throughput of uplink SDMA and derive design guidelines, the throughput scaling with system parameters is analyzed. The specific parameters considered include the numbers of users, antennas, and feedback bits. Furthermore, different SNR regimes and beamforming methods are considered. The derived throughput scaling laws are observed to change for different SNR regimes. For instance, the throughput scales logarithmically with the number of users in the high SNR regime but double logarithmically in the low SNR regime. The analysis of throughput scaling suggests guidelines for scheduling in uplink SDMA. For example, to maximize throughput scaling, scheduling should use the criterion of minimum quantization errors for the high SNR regime and maximum channel power for the low SNR regime.

Copyright © 2008 Kaibin Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

In a wireless communication system, using the spatial degrees of freedom, a base station with multiantennas can communicate with multiple users in the same time and frequency slot. This method, known as *space division multiple access* (SDMA), significantly increases throughput. SDMA is capable of achieving multiuser channel capacity with only one-end joint processing at the base station by employing *dirty paper coding* for the downlink [1] or successive interference cancelation for the uplink [2]. Despite being suboptimal, SDMA with the linear beamforming constraint has attracted extensive research recently due to its low-complexity and satisfactory performance (see, e.g., [3–5]). In a system with a large number of users, the simplicity of beamforming SDMA facilitates its joint designs with scheduling [6–8]. Integrating SDMA and scheduling achieves both the multiplexing and multiuser diversity gains [6, 8, 9], leading to high throughput. This paper considers an uplink SDMA system with scheduling. Specifically, this paper characterizes how the throughput of uplink SDMA scales with different system parameters. These parameters include the number of antennas, the number of users, and the amount of channel state information (CSI) feedback.

Both uplink SDMA and scheduling require CSI of the multiuser uplink channels at the base station. In the presence of line-of-sight propagation, the base station estimates the directions of arrival of different users, and uses this information for beamforming and scheduling [10, 11]. For channels with rich scattering (non-line-of-sight), the base station can estimate uplink channels using pilot symbols transmitted by scheduled users [12–14]. Nevertheless, for a large number of users, scheduled users constitute only a small subset of users, but joint SDMA and scheduling require CSI of all users. Therefore, CSI feedback from all users is required if the user pool is large.

Two CSI feedback methods exist, namely, limited feedback [15] and analog feedback [16]. Analog feedback involves uplink transmission of pilot symbols from the mobile users and thereby enables channel estimation at the base station [16]. Alternatively, limited feedback replaces pilot symbols with quantized CSI [15]. The relative efficiency of these two types of feedback overhead, namely, pilot symbols and quantized CSI, is unclear but is outside the scope of this paper. The use of limited feedback requires channel reciprocity (in, e.g., time division multiplexing (TDD) systems), which enables users to acquire uplink CSI through downlink channel estimation. Compared with analog feedback, limited

feedback supports flexible feedback rates and CSI protection using error-control coding. For these advantages, limited feedback is considered in this paper. The required assumption on the existence of channel reciprocity is made in this paper.

To maximize throughput, the design of SDMA with limited feedback requires joint optimization of scheduling, beamforming, and CSI quantization algorithms. This optimization problem is difficult and remains open. Nevertheless, it is a much easier task to design an SDMA system that achieves the optimum throughput scaling with key system parameters such as the feedback rate, the number of users, and the antenna array size. The analysis of throughput scaling laws provides useful guidelines for designing uplink SDMA with limited feedback. Therefore, such analysis forms the theme of this paper.

1.1. Prior work and motivation

The prior work on throughput scaling laws of SDMA with limited feedback targets the downlink [6, 8, 17]. The existing analytical approach is to use the *extreme value theory* [6, 8], but this approach is not directly applicable for uplink SDMA as explained below. The key to this approach is the derivation of the *probability density function* (pdf) of the signal-to-interference-noise ratio (SINR). This SINR PDF allows the application of extreme value theory for analyzing the throughput scaling law. The above approach is feasible for downlink SDMA because the SINR of a scheduled user depends only on this user's CSI [6, 8]. In contrast, for uplink SDMA, this SINR is a function of the CSI of all scheduled users. Such a discrepancy is due to the difference between the downlink and uplink. To be specific, both the signal and interference received by a user (the base station) propagate through the same channel (different channels) in the downlink (uplink). Consequently, the derivation of the SINR pdf for uplink SDMA is complicated because of its dependence on the specific scheduling algorithm. This motivates us to seek new tools for analyzing the throughput scaling laws for uplink SDMA.

Two beamforming and scheduling methods, *zero-forcing beamforming* [6, 18] and *orthogonal beamforming* [8, 17, 19], are being discussed for enabling downlink SDMA with limited feedback in the 3GPP-LTE standard [19, 20]. Due to the uplink-downlink difference mentioned above, the scaling laws for downlink SDMA in [6, 8, 17] cannot be directly extended to the uplink counterpart. Furthermore, the scaling law for orthogonal beamforming in the interference-limited regime remains unknown even for downlink SDMA. This motivates us to consider both orthogonal and zero-forcing beamforming in the analysis of uplink SDMA. Furthermore, the throughput scaling analysis covers high SNR (interference limited), normal SNR, and low SNR (noise limited) regimes.

1.2. Contributions

To discuss the contributions of this paper, the system model is summarized as follows. The uplink SDMA system model includes a base station with multiantennas and users with single-antennas. The multiuser channels are assumed to follow the i.i.d. Rayleigh distribution. The CSI feedback of each user consists of a quantized channel-direction vector and two real scalars, namely, the quantization error and the channel power, which can be assumed perfect since they require much less feedback than the vector. Moreover, both orthogonal [8, 17] and zero-forcing beamforming [6, 21] are considered for beamforming at the base station.

The main contributions of this paper are the asymptotic throughput scaling laws for uplink SDMA with limited feedback in different SNR regimes and for both orthogonal and zero-forcing beamforming. The derivation of the throughput scaling laws makes use of new analytical tools including the Vapnik-Chervonenkis theorem [22] and the bins-and-balls model [23] for analyzing multiuser limited feedback. Our results are summarized as follows.

(1) In the high SNR regime and for orthogonal beamforming, an upper and a lower bound are derived for the throughput scaling factor. These bounds show that the throughput scales *logarithmically* with both the number of users U and the quantization codebook size N. Furthermore, the linear scaling factor is smaller than the number of antennas N_t, indicating the loss in the spatial multiplexing gain.

(2) In the high SNR regime and for zero-forcing beamforming, the exact throughput scaling factor is derived, which provides the same observations as for orthogonal beamforming. To be specific, the throughput scales logarithmically with both U and N. The linear factor of the asymptotic throughput is smaller than N_t.

(3) In the normal SNR regime, for both orthogonal and zero-forcing beamforming, the throughput is shown to scale *double logarithmically* with U and linearly with N_t.

(4) The same results are obtained for the lower SNR regime.

The analysis of the throughput scaling laws provides the following guidelines for designing uplink SDMA with limited feedback. In the high SNR regime, the scheduling algorithm should select users with minimum quantization errors. Thus, feedback of channel power for scheduling is unnecessary. In the lower SNR regime, the scheduled users should be those with maximum channel power. Consequently, scheduling requires no feedback of quantization errors. In the normal SNR regime, the scheduling criterion should include both channel power and quantization errors. This implies that the feedback of both types of CSI is needed.

The remainder of this paper is organized as follows. The system model is described in Section 2. Background on limited feedback, scheduling, and beamforming is provided in Section 3. Analytical tools are discussed in Section 4. Using these tools, the asymptotic throughput scaling of uplink SDMA is analyzed in Sections 5, 6, and 7, respectively, for the high, normal, and low SNR regimes. Numerical results are presented in Section 8, followed by concluding remarks in Section 9.

2. SYSTEM DESCRIPTION

The uplink SDMA system considered in this paper is illustrated in Figure 1. In this system, U backlogged users each with a single antenna attempt to communicate with a base station with N_t antennas. For each time slot, up to N_t users are scheduled for uplink SDMA transmission. Users learn the scheduling decisions from the indices of scheduled users broadcast by a base station. The base station separates the data packets of scheduled users by receive beamforming. The base station requires the CSI feedback from all users for scheduling and beamforming. Each user sends back CSI using limited feedback as elaborated later. Two approaches for scheduling and beamforming based on limited feedback are analyzed in this paper, namely, *orthogonal beamforming* [8, 17] and *zero-forcing beamforming* [6, 21], which are discussed, respectively, in Sections 3.3.1 and 3.3.2.

Assuming the presence of channel reciprocity (hence a time-division multiplexing (TDD) system), each user estimates the downlink channel, equivalently the uplink channel, using pilot symbols periodically broadcast by the base station. For simplicity, we make the following assumption.

Assumption 1. Each user has perfect CSI of the corresponding uplink channel.

This assumption simplifies analysis by allowing omission of channel estimation errors. Consider a system with a large number of users. Even by exploiting channel reciprocity, the base station can acquire the CSI of only the scheduled uplink users, which is a small subset of users. Nevertheless, the base station requires the CSI of all users for scheduling and beamforming, which motivates the CSI feedback from all users. Each user relies on a finite-rate feedback channel for CSI feedback, thus limited feedback is used for efficiently quantizing CSI for satisfying the finite-rate constraint.

The uplink channel of each user is modeled as a frequency-flat block-fading vector channel. By blocking fading, channel realizations for different time slots are independent. Consequently, the uplink channel of the uth user can be represented by a random vector $\mathbf{h}_u$. To simplify our analysis, we make the following assumption.

Assumption 2. The vector channel of each user, $\mathbf{h}_u$ where $u = 1, 2, \ldots, U$, is an i.i.d vector with complex Gaussian coefficients $\mathcal{CN}(0, 1)$.

This assumption is commonly made in the literature of multiuser diversity [7, 8, 21, 24]. For analysis, the channel vector $\mathbf{h}_u$ is decomposed into *channel shape* and *channel power*, defined as $\mathbf{s}_u = \mathbf{h}_u/\|\mathbf{h}_u\|$ and $\rho_u = \|\mathbf{h}_u\|^2$, respectively.

Based on the above model, the vector of multiantenna observations at the base station, denoted as $\mathbf{y}$, can be written as

$$\mathbf{y} = \sum_{u\in\mathcal{A}} \sqrt{P\rho_u}\mathbf{s}_u x_u + \boldsymbol{\nu}, \tag{1}$$

where $\mathcal{A}$ is the index set of scheduled users, x_u is the data symbol of the uth user, and $\boldsymbol{\nu}$ is the AWGN vector. Furthermore, the recovered data symbol for the scheduled uth user after beamforming is given as

$$\hat{x}_u = \mathbf{v}_u^\dagger \mathbf{y} = \sqrt{P\rho_u}\mathbf{v}_u^\dagger \mathbf{s}_u x_u + \sum_{m\in\mathcal{A}/\{u\}} \sqrt{P\rho_m}\mathbf{v}_u^\dagger \mathbf{s}_m x_m + \nu_u, \tag{2}$$

where $\mathbf{v}_u$ is the beamforming vector used for retrieving the data symbol of the uth user.

3. LIMITED FEEDBACK, SCHEDULING, AND BEAMFORMING

This section presents the analytical framework for limited feedback, scheduling, and beamforming for uplink SDMA. SINR and throughput are important quantities for scheduling at the base station. Their exact values are unknown to the base station because of imperfect CSI feedback. The approximated SINR and throughput, named *expected* SINR and *expected throughput*, are discussed in Sections 3.1 and 3.2, respectively. These new quantities are computable at the base station using limited feedback.

Based on limited feedback, the beamforming vectors of scheduled users are computed at the base station to satisfy the following constraint:

$$\mathbf{v}_u \perp \hat{\mathbf{s}}_{u'} \quad \forall u, u' \in \mathcal{A},\ u \neq u', \tag{3}$$

where $\mathbf{v}_u$ is the beamforming vector, $\hat{\mathbf{s}}_u$ the quantized channel-shape, and $\mathcal{A}$ the index set of scheduled users. This constraint has been also used for downlink SDMA with limited feedback [7, 8, 17, 21]. For perfect feedback ($\mathbf{s}_u = \hat{\mathbf{s}}_u$), the above constraint ensures no interference between scheduled users. In Section 3.3, two beamforming approaches for satisfying (3), namely, *orthogonal beamforming* and *zero-forcing beamforming*, are introduced. In addition, the compatible scheduling methods are also described.

3.1. Expected SINR

In this section, the expected SINRs of scheduled users are defined, which are computable using limited feedback. Given the index set of scheduled users $\mathcal{A}$ and corresponding beamforming vectors $\{\mathbf{v}_u\}$, as in [6, 21], the SINR is obtained from (2) as

$$\mathrm{SINR}_u = \frac{\gamma\rho_u \left|\mathbf{v}_u^\dagger \mathbf{s}_u\right|^2}{1 + \gamma\sum_{m\in\mathcal{A}, m\neq u}\rho_m \epsilon_m \beta_{m,u}}, \tag{4}$$

where the *signal-to-noise ratio* (SNR) $\gamma = P/\sigma_\nu^2$, and $\mathbf{s}_u$ and ρ_u are, respectively, the channel shape and power of the uth user, $\epsilon_u = \sin^2(\angle(\mathbf{s}_u, \hat{\mathbf{s}}_u))$ is the quantization error of the channel shape. Moreover, $\beta_{m,u}$ is a Beta random variable that is independent of ϵ_m and has the cumulative density function (CDF) $\Pr(\beta_{m,u} \leq \beta_0) = \beta_0^{N_t - 1}$.

The direct feedback of SINRs in (4) by users is infeasible as computation of SINRs requires multiuser CSI and such information is unavailable to individual users. Note that the SINR feedback is feasible for downlink SDMA since the SINR

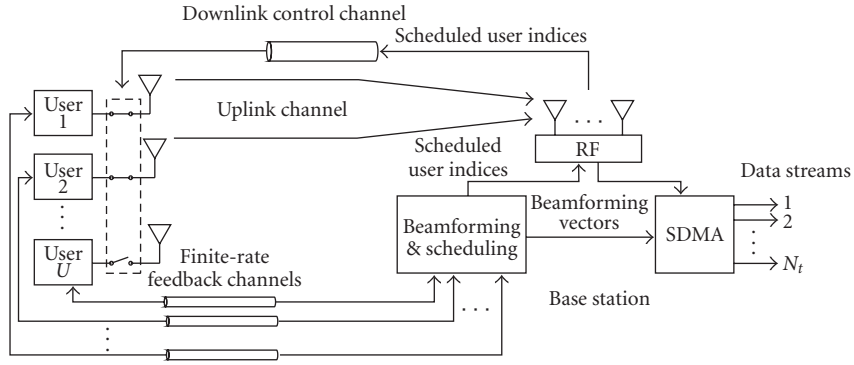


Figure 1: Uplink SDMA system with limited feedback.

depends only on single-user CSI [8] or approximately so [6]. Therefore, we require that the expected SINR is computable at the base station using individual users' CSI feedback.

The expected SINR is defined as follows, which is computable from the feedback of channel power $\{\rho_u\}$ and channel-shape quantization errors $\{\epsilon_u\}$ by users. In addition, the feedback of quantized channel shapes allows the base station to compute beamforming vectors $\{\mathbf{v}_u\}$ that satisfy the constraint in (3). As feedback of a scalar requires potentially much fewer bits than that of a vector, the following assumption is made throughout this paper unless specified otherwise.

Assumption 3. The feedback of channel power $\{\rho_u\}$ and channel-shape quantization errors $\{\epsilon_u\}$ from all users are perfect.

Depending on the operational SNR regime, either of these two types of scalar feedback can be avoided as shall be discussed later. Given Assumption 3, limited feedback in this paper focuses on quantization and feedback of channel shapes. Under Assumption 3, the expected SINR for the uth user, denoted as Ψ_u, is defined as

$$\Psi_u = \frac{\gamma \rho_u}{1 + \gamma \sum_{m \in \mathcal{A}, m \neq u} \rho_m \epsilon_m}. \tag{5}$$

3.2. Expected throughput

In this section, the expected throughput that approximates the exact one is defined as follows:

$$R = \mathbf{E}\left[\sum_{u \in \mathcal{A}} \log\left(1 + \Psi_u\right)\right], \tag{6}$$

where Ψ_u is defined in (5) and $\mathcal{A}$ is the index set of scheduled users. This quantity is estimated by the base station using limited feedback and for a given set of scheduled users.

Next, the expected throughput is shown to converge to the actual one when the number of users is large. Therefore, the expected throughput can replace the actual one in the asymptotic analysis of throughput scaling, which significantly simplifies our analysis. To obtain the desired result, a useful lemma from [21] is provided below.

Lemma 1. *Let $\epsilon(N)$ be the minimum of N i.i.d. Beta random variables. The following inequalities hold:*

$$\begin{aligned} \mathbf{E}[-\log(\epsilon(N))] &\leq \frac{\log N + 1}{N_t - 1}, \\ \mathbf{E}[\epsilon(N)] &< (N)^{-1/(N_t - 1)}. \end{aligned} \tag{7}$$

Let φ_u denote the angle between the beamforming vector and quantized channel shape of the uth scheduled user, hence $\varphi_u = \angle(\mathbf{v}_u, \hat{\mathbf{s}}_u)$. Using this lemma, the following result on the difference between the expected and the exact throughput is proved.

Proposition 1. *If $\varphi_u \leq \varphi_0$, $\epsilon_u \leq \theta_0$, and $(\varphi_0 + \theta_0) < \pi/2$, then*

$$|R - C| \leq \max\left\{2\log\cos(\varphi_0 + \theta_0), \frac{N_t}{N_t - 1}[\log(N_t - 1) + 1]\right\}, \tag{8}$$

where C is the exact throughput given as

$$C = \mathbf{E}\left[\sum_{u \in \mathcal{A}} \log\left(1 + \mathrm{SINR}_u\right)\right]. \tag{9}$$

The proof is given in Appendix A. As shown in subsequent sections, the expected throughput R increases continuously with the number of users U. Consequently, from Proposition 1, the expected throughput R has the same asymptotic scaling factor as the exact throughput in (9).

3.3. Beamforming methods

The orthogonal and zero-forcing beamforming methods are commonly used in the literature of downlink SDMA with limited feedback [6, 8, 17, 18, 21]. These methods are adopted in this paper for uplink SDMA as elaborated in Sections 3.3.1 and 3.3.2, respectively.

The main difference between orthogonal and zero-forcing beamforming lies in their use of the quantizer codebook. For orthogonal beamforming, the codebook of unitary vectors provides potential beamforming vectors. In other words, quantized CSI of scheduled users directly provides their beamforming vectors. For zero-forcing beamforming,

the codebook is used in the traditional way as in vector quantization. Beamforming vectors are computed from quantized CSI using the zero-forcing method.

3.3.1. Orthogonal beamforming

In this section, orthogonal beamforming for downlink SDMA with limited feedback is discussed. The orthogonal beamforming method is characterized by the following constraint [8, 17]:

$$\text{(orthogonal beamforming)} \begin{cases} \hat{\mathbf{s}}_u \perp \hat{\mathbf{s}}_{u'} & \forall u, u' \in \mathcal{A},\ u \neq u', \\ \mathbf{v}_u = \hat{\mathbf{s}}_u & \forall u \in \mathcal{A}. \end{cases} \tag{10}$$

The above constraint can be implemented using the following joint design of limited feedback, beamforming, and scheduling (see, e.g., [17]). First, the channel shape of each user is quantized using a codebook that is comprised of multiple orthonormal vector sets. Let $\mathcal{F}$ denote the codebook, $N = |\mathcal{F}|$ the codebook size, and $M := N/N_t$ the number of orthonormal sets in $\mathcal{F}$. Moreover, let $\mathbf{v}_n^{(m)}$ denote the nth member of the mth orthonormal set in $\mathcal{F}$. Thus, $\mathcal{F} = \{\mathbf{v}_n^{(m)}, 1 \leq n \leq N_t, 1 \leq m \leq M\}$. As in [17], the M orthonormal vector sets of $\mathcal{F}$ are generated randomly and independently using a method such as that in [25]. Consider the quantization of $\mathbf{s}_u$, the channel shape of the uth user. Following [26], the quantizer function is given as

$$\hat{\mathbf{s}}_u = \arg \max_{\mathbf{v} \in \mathcal{F}} |\mathbf{v}^\dagger \mathbf{s}_u|^2, \tag{11}$$

where $\hat{\mathbf{s}}_u$ represents the quantized channel shape. The quantization error is given as $\epsilon_u = |\hat{\mathbf{s}}^\dagger \mathbf{s}|^2$. The quantized channel shapes $\{\hat{\mathbf{s}}_u\}$ as well as channel power $\{\rho_u\}$ and quantization error $\{\epsilon_u\}$ are sent back from the users to the base station.

The base station constrains the quantized channel shapes of scheduled users to belong to the same orthonormal set in the codebook $\mathcal{F}$. Furthermore, the quantized channel shapes of scheduled users are applied as beamforming vectors. Thereby, the orthogonal beamforming constraint in (10) is satisfied. Under this constraint and for the criterion of maximizing throughput, the expected throughput defined in (6) can be written as

$$R_{\text{or}} = \mathbf{E}\left[\max_{1 \leq m \leq M} \max_{\substack{u_n \in \mathcal{I}_n^{(m)} \\ n=1,\ldots,N_t}} \sum_{n=1}^{N_t} \log\left(1 + \Psi_{u_n}\right) \right], \tag{12}$$

where Ψ_{u_n} is the scheduling metric defined in (5). The user index set $\mathcal{I}_n^{(m)}$, which groups users with identical quantized channel shapes, is defined as

$$\mathcal{I}_n^{(m)} = \{1 \leq u \leq U \mid \hat{\mathbf{s}}_u = \mathbf{v}_n^{(m)}\}, \quad 1 \leq m \leq M,\ 1 \leq n \leq N_t. \tag{13}$$

3.3.2. Zero-forcing beamforming

In this section, the zero-forcing beamforming method for SDMA with limited feedback [6, 21] is introduced, which satisfies the following constraint:

$$\text{(zero-forcing beamforming)} \begin{cases} \angle(\hat{\mathbf{s}}_u, \hat{\mathbf{s}}_{u'}) \geq \varphi_0 \\ \quad \forall u, u' \in \mathcal{A},\ u \neq u', \\ \mathbf{v}_u \perp \hat{\mathbf{s}}_{u'} \\ \quad \forall u, u' \in \mathcal{A},\ u \neq u'. \end{cases} \tag{14}$$

The constant $0 < \varphi_0 < 1$, which is usually large, ensures that the quantized channel shapes of scheduled users are well separated in angles [6]. The second condition of the above constraint is satisfied by computing beamforming vectors $\{\mathbf{v}_u, u \in \mathcal{A}\}$ from $\{\hat{\mathbf{s}}_u, u \in \mathcal{A}\}$ using the zero-forcing method [6, 21]. Following [6, 21], the channel shape of each user is quantized using the random vector quantization method, where the codebook $\mathcal{F}$ consists of N i.i.d. isotropic unitary vectors.

To derive an expression of the expected throughput for the criterion of maximizing throughput, define all subsets of users whose quantized channel shapes satisfy the first condition of the beamforming constraint in (14) as follows:

$$\begin{aligned} \{\mathcal{B}\} = \{\mathcal{B} \subset \mathcal{U} \mid |\mathcal{B}| \leq N_t, \\ \angle(\hat{\mathbf{s}}_u, \hat{\mathbf{s}}_{u'}) \geq \varphi_0\ \forall u, u' \in \mathcal{B},\ u \neq u'\}. \end{aligned} \tag{15}$$

In terms of the above subsets, the expected throughput can be written as

$$R_{\text{zf}} = \mathbf{E}\left[\max_{\mathcal{A} \subset \{\mathcal{B}\}} \sum_{u \in \mathcal{A}} \log\left(1 + \Psi_u\right) \right], \tag{16}$$

where the expected SINR Ψ_u is given in (5).

4. BACKGROUND: ANALYTICAL TOOLS

In this section, two analytical tools are provided for analyzing the throughput scaling laws in the sequel. In Section 4.1, the bins-and-balls model is discussed, which models multiuser limited feedback. In Section 4.2, the theory of uniform convergence in the weak law of large numbers is introduced. This theory is useful for characterizing the number of users whose channel shapes lie in a same Voronoi cell.

4.1. Bins and balls

In this section, a bins-and-balls model for multiuser feedback of quantized channel shapes is introduced. This model provides a useful tool for analyzing throughput scaling law for orthogonal beamforming in Section 5.1. In this model as illustrated in Figure 2, U balls are thrown into $N + 1$ bins: N small bins and one big one, whose total volume is equal to one.

Some useful results are derived using the bins-and-balls model. Let the probability that a ball falls into a specific bin

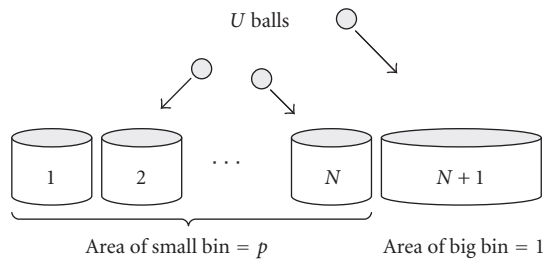


Figure 2: The bins-and-balls model for multiuser feedback of quantized channel shapes.

be equal to p for each small bin and q for the big bin, hence $q = 1 - Np$. The first question to ask is *how many small bins are nonempty?* The answer to this question is provided in the following lemma, obtained Using the Chebychev's inequality [23].

Lemma 2. *Denote* $\widetilde{p} = 1-(1-p)^U$. *The number of nonempty small bins W satisfies*

$$\Pr\left(W \geq N\widetilde{p} - \sqrt{\log N(N\widetilde{p} - N\widetilde{p}^2)}\right) \geq 1 - \frac{1}{\log N}. \quad (17)$$

Next, consider clusters of N_t neighboring small bins. In Section 5.1, each cluster is related to an orthonormal vector set in the quantizer codebook for orthogonal beamforming. Each cluster is said to be nonempty if it contains no empty bins. Then, the second question to ask is *how many clusters are nonempty?* The answer is provided in the following corollary of Lemma 2.

Corollary 1. *Denote the number of nonempty clusters of small bins as Q. Then Q satisfies*

$$\Pr\left(Q \geq M\widetilde{p}^{N_t} - \sqrt{\log M(M\widetilde{p}^{N_t} - M\widetilde{p}^{2N_t})}\right) \geq 1 - \frac{1}{\log M}, \quad (18)$$

where M is the total number of clusters.

4.2. Uniform convergence in weak law of large numbers

In this section, a lemma on the uniform convergence in the weak law of large numbers [22] is obtained by generalizing [27, Lemma 4.8]. This lemma given below is useful for analyzing the number of users whose channel shapes lie in one of a set of congruent disks on the surface of a hyper sphere. Such analysis will appear frequently in the subsequent throughput analysis.

Lemma 3 (Gupta and Kumar). *Consider U random points uniformly distributed on the surface of a unit hyper-sphere in* C^{N_t} *and N disks on the sphere surface that have equal volume denoted as A. Let* T_n *denote the number of points belong to the nth disk. For every* $\tau_1, \tau_2 > 0$*:*

$$\Pr\left(\sup_{1\leq n\leq N}\left|\frac{T_n}{U} - A\right| \leq \tau_1\right) > \tau_2 \quad U \geq U_o, \quad (19)$$

where

$$U_o = \max\left\{\frac{3}{\tau_1}\log\frac{16c}{\tau_2}, \frac{4}{\tau_1}\log\frac{2}{\tau_2}\right\}, \quad (20)$$

and c is a constant.

5. THROUGHPUT SCALING: HIGH SNR

In this section, the throughput scaling law of uplink SDMA in the high SNR regime ($\gamma \gg 1$) is analyzed. The expected SINR in (5) for this regime is simplified as

$$\Psi_u^{(\alpha)} = \frac{\rho_u}{\sum_{m\in\mathcal{A}, m\neq u}\rho_m\epsilon_m}, \quad (21)$$

where the superscript (α) is added to indicate the high SNR regime. Using the analytical tools discussed in Section 4, the throughput scaling laws are derived in Sections 5.1 and 5.2 for orthogonal and zero-forcing beamforming, respectively.

5.1. Throughput scaling for orthogonal beamforming

In this section, we analyze the throughput scaling laws for orthogonal beamforming in the high SNR regime. Two cases are considered. First, both the number of users U and the quantization codebook size N are large. For this case, we derive an upper and a lower bounds for the throughput scaling factor as functions of U and N. Second, U is large but N is fixed. For this case, the exact throughput scaling factor in terms of U is obtained.

5.1.1. $U\to\infty$ and $N\to\infty$

To derive the throughput scaling law for $U\to\infty$ and $N\to\infty$, the following approach is adopted. First, we derive an upper bound for the throughput scaling factor of the expected throughput, which is defined in (6). To avoid confusion, the expected throughput is denoted as $R_{\text{or}}^{(\alpha)}$ where the superscript specifies the high SNR regime and the subscript indicates orthogonal beamforming. Second, an achievable lower bound is obtained by constructing a suboptimal scheduling algorithm. Last, the throughput scaling law for $R_{\text{or}}^{(\alpha)}$ is shown to hold for the exact throughput.

An upper bound for scaling factor of $R_{\text{or}}^{(\alpha)}$ is derived as follows. To avoid considering any specific scheduling algorithm in the derivation, the following assumption is made.

Assumption 4. The channel power of a scheduled user is lower-bounded as:

$$\rho_u \geq \frac{1}{\log U + c} \quad \forall u \in \mathcal{A}. \quad (22)$$

This assumption is justifiable under the current design criterion of maximizing throughput. Under this criterion, as U grows, the channel power of scheduled users increases but the lower bound in (22) converges to zero. Since $\rho_u \geq 0$ and we are interested in the case of $U\to\infty$, Assumption 4 is justified. Using this assumption, an upper bound for the scaling factor of $R_{\text{or}}^{(\alpha)}$ is derived and shown in the following lemma.

Lemma 4. *In the high SNR regime and for the case of $U\to\infty$ and $N\to\infty$, the scaling factor of the expected throughput $R_{\text{or}}^{(\alpha)}$ in* (6) *is upper bounded as*

$$\lim_{\substack{U\to\infty\\ B\to\infty}} \frac{R_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))(\log U+\log N)} \le 1. \tag{23}$$

The proof is given in Appendix B.

Next, an achievable lower bound for the scaling factor of $R_{\text{or}}^{(\alpha)}$ is obtained. The direct derivation of a scheduling algorithm for maximizing the scaling factor of $R_{\text{or}}^{(\alpha)}$ in (6) is very difficult if not impossible. To overcome this difficulty, we argue that it is unnecessary to consider channel power in scheduling. In the sequel, we prove that the scheduling neglecting channel power leads to a reasonable lower bound of the optimum throughput scaling factor for orthogonal beamforming. The reason for the above argument is that scheduling users with largest channel power can at most increase the scaling factor by only $O(\log\log U)$ since the largest power scales as $\log U$ [8]. Such an increment is negligible because the expected scaling factor is $O(\log U)$ as shown in Lemma 4. Thus, to achieve the optimum throughput scaling, using minimum quantization errors $\{\epsilon_u\}$ as the scheduling criterion suffices. In the high SNR regime that is interference limited, such a criterion minimizes interference caused by quantization errors. The use of only quantization errors as the scheduling criterion leads to the following lower bound for $R_{\text{or}}^{(\alpha)}$. Let $\chi_1^2,\chi_2^2,\ldots,\chi_{N_t}^2$ denote a sequence of chi-squared random variables representing the channel power of scheduled users. From (6) and (21),

$$\begin{aligned}
R_{\text{or}} &\ge \mathbf{E}\left[\max_{1\le m\le M}\max_{\substack{u_k\in\mathcal{I}_k^{(m)}\\ k=1,\ldots,N_t}}\sum_{n=1}^{N_t}\log\left(1+\frac{\chi_n^2}{\sum_{k=1,k\ne n}^{N_t}\chi_k^2\epsilon_{u_k}}\right)\right]\\
&\ge \mathbf{E}\left[\max_{1\le m\le M}\sum_{n=1}^{N_t}\log\left(1+\frac{\chi_n^2}{\sum_{k=1,k\ne n}^{N_t}\chi_k^2\min_{u\in\mathcal{I}_k^{(m)}}\epsilon_u}\right)\right]\\
&\ge N_t\mathbf{E}\left[\max_{1\le m\le M}\right.\\
&\qquad\left.\times\log\left(1+\frac{\chi_n^2}{\max_{1\le n\le N_t}\min_{u\in\mathcal{I}_n^{(m)}}\epsilon_u\sum_{k=1,k\ne n}^{N_t}\chi_k^2}\right)\right]\\
&= N_t\mathbf{E}\left[\log\left(1+\frac{\chi_n^2}{\epsilon^\star\sum_{k=1,k\ne n}^{N}\chi_k^2}\right)\right],
\end{aligned} \tag{24}$$

where

$$\epsilon^\star=\min_{1\le m\le M}\max_{1\le n\le N_t}\min_{u\in\mathcal{I}_n^{(m)}}\epsilon_u. \tag{25}$$

A scheduling algorithm directly follows from the throughput lower bound in (24). Define

$$m^\star=\arg\min_{1\le m\le M}\left(\max_{1\le n\le N_t}\min_{u\in\mathcal{I}_k^{(m)}}\epsilon_u\right). \tag{26}$$

Then the scheduled user set $\mathcal{A}$ is given as

$$\mathcal{A}=\left\{\arg\min_{u\in\mathcal{I}_n^{(m^\star)}}\epsilon_u,\ 1\le n\le N_t\right\}. \tag{27}$$

Using this scheduled algorithm, an achievable lower bound of the throughput scaling factor is obtained and shown in the following lemma.

Lemma 5. *In the high SNR regime and for the case of $U\to\infty$ and $N\to\infty$, the scaling factor of the expected throughput $R_{\text{or}}^{(\alpha)}$ in* (6) *is lower-bounded as*

$$\lim_{\substack{U\to\infty\\ N\to\infty}}\frac{R_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))\log U+(1/(N_t-1))\log N}\ge 1. \tag{28}$$

The proof is given in Appendix C. The proof procedure involves using the bins-and-balls model and Lemma 1 in Section 4.1.

Proposition 1 implies the identical throughput scaling factors for the expected throughput $R_{\text{or}}^{(\alpha)}$ and the exact one, denoted as $C_{\text{or}}^{(\alpha)}$, because their difference is no more than a constant. By combining Proposition 1, Lemmas 5 and 4, the main result of this section is obtained and summarized in the following theorem.

Theorem 1. *In the high SNR regime and for the case of $U\to\infty$ and $N\to\infty$, the scaling law of the throughput for orthogonal beamforming is given as*

$$\begin{aligned}
&\lim_{\substack{U\to\infty\\ N\to\infty}}\frac{C_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))\log U+(N_t/(N_t-1))\log N}\le 1,\\
&\lim_{\substack{U\to\infty\\ N\to\infty}}\frac{C_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))\log U+(1/(N_t-1))\log N}\ge 1.
\end{aligned} \tag{29}$$

A few remarks are in order.

(i) The bounds in (29) agree on that the throughput scaling factor with respect to U is $(N_t/(N_t-1))\log U$.

(ii) The lower and the upper bounds in (29) differ by N_t times in the throughput scaling factor with respect to N. The smaller scaling factor in the constructive lower bound is due to the use of a suboptimal scheduling algorithm. The design of a scheduling algorithm for achieving the upper bound for the scaling factor in (29) is a topic for future investigation.

(iii) No feedback of channel power is required for achieving the lower bound for the throughput scaling factor in (29), because scheduling is independent of channel power.

5.1.2. $U\to\infty$ and N fixed

In this section, the throughput scaling law for orthogonal beamforming is analyzed for the high SNR regime and the case where the codebook size N is fixed and the number of users $U\to\infty$.

The upper bound of the throughput scaling factor is shown in the following lemma. The proof can be easily modified from that for Lemma 4 by substituting $\lim_{U\to\infty} \log N/\log U = 0$.

Lemma 6. *In the high SNR regime and with N fixed, the throughput scaling factor for orthogonal beamforming is upper-bounded as*

$$\lim_{U\to\infty} \frac{R_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))\log U} \leq 1. \tag{30}$$

Next, the equality in (30) is shown to hold using the following scheduling algorithm. First, among users belonging to the index set $\mathcal{I}_n^{(m)}$, the one with the smallest quantization error is selected. Second, among the selected users corresponding to the index sets $\{\mathcal{I}_n^{(m)}\}$, an arbitrary set of users with orthogonal quantized channel shapes are scheduled and these orthogonal vectors are applied as their beamforming vectors. Using this scheduling algorithm, the index set of scheduled users can be written as $\mathcal{A} = \{\arg\min_{u\in\mathcal{I}_n^{(m)}} \epsilon_u,\ 1 \leq n \leq N_t\}$. Based on the above scheduling algorithm and from (6), the expected throughput is bounded as

$$R_{\text{or}}^{(\alpha)} \geq N_t \mathbf{E}\left[\log\left(1 + \frac{\chi_n^2}{\sum_{k=1,k\neq n}^{N} \chi_k^2 \min_{u\in\mathcal{I}_k^{(m)}} \epsilon_u}\right)\right]. \tag{31}$$

Using the above throughput lower bound and Lemma 6, the following lemma is proved.

Lemma 7. *The upper bound of the throughput scaling factor in* (30) *is achievable:*

$$\lim_{U\to\infty} \frac{R_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))\log U} = 1. \tag{32}$$

The proof is given in Appendix D. This proof makes use of the theory of uniform convergence in the weak law of large numbers as discussed in Section 4.2.

By combining Lemma 7 and Proposition 1, the main result of this section is obtained and summarized in the following theorem.

Theorem 2. *In the high SNR regime ($\gamma \gg 1$) and with a fixed codebook size N, the throughput scaling law for orthogonal beamforming is*

$$\lim_{U\to\infty} \frac{C_{\text{or}}^{(\alpha)}}{(N_t/(N_t-1))\log U} = 1. \tag{33}$$

Two remarks are given.

(i) The current throughput scaling factor is identical to the first terms of the bounds in (29) corresponding to the case of $N\to\infty$.

(ii) For $N_t \geq 3$, the linear scaling factor in (33), namely, $N_t/(N_t-1)$, is smaller than N_t, which is the number of available spatial degrees of freedoms. This indicates the loss in multiplexing gain for $N_t \geq 3$.

5.2. Throughput scaling for zero-forcing beamforming

In this section, the scaling law for zero-forcing beamforming in the high SNR regime is analyzed. Two cases are considered: (1) $U\to\infty$ and $N\to\infty$ and (2) $U\to\infty$ and N is fixed, which are jointly analyzed due to their similarity in analysis. Denote the expected and the exact throughput for zero-forcing beamforming in the high SNR regime as R_{zf}^{α} and C_{zf}^{α}.

The upper bounds of the throughput scaling factor for orthogonal beamforming in Lemmas 4 and 6 can be shown to hold for zero-forcing beamforming by trivial modifications of the proofs. Thus,

$$\begin{aligned} \lim_{\substack{U\to\infty \\ N\to\infty}} \frac{R_{\text{zf}}^{(\alpha)}}{(N_t/(N_t-1))(\log U + \log N)} &\leq 1, \\ \lim_{U\to\infty} \frac{R_{\text{zf}}^{(\alpha)}}{(N_t/(N_t-1))\log U} &\leq 1. \end{aligned} \tag{34}$$

The above upper bounds for the throughput scaling factor of zero forcing beamforming can be achieved using the following scheduling algorithm. Consider an arbitrary basis of $\mathbb{C}^{N_t}$, denoted as $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{N_t}\}$. Using this basis, we define the following index sets:

$$\mathcal{J}_k = \{1 \leq u \leq U \mid 1 - |\mathbf{q}_n^{\dagger}\hat{\mathbf{s}}_u|^2 \leq \tau_o\} \quad 1 \leq k \leq N_t, \tag{35}$$

where $\tau_o = \sin^2((\pi/4) - (\varphi_o/2)) = (1 + \sin(\varphi_o))/2$ and $\hat{\mathbf{s}}_u$ is the quantized channel shape. The purpose of these index sets is to select users who satisfy the zero-forcing beamforming constraint in (14). Among the users in each of the index sets $\{\mathcal{J}_k\}$, the one with the smallest quantization error is scheduled. In other words, the index set of the scheduled users is

$$\mathcal{A} = \left\{\arg\min_{u\in\mathcal{J}_k} \epsilon_u,\ 1 \leq k \leq N_t\right\}. \tag{36}$$

The beamforming vectors of the scheduled users are computed from their quantized channel shapes using the zero-forcing method. From the above, scheduling algorithm results in the following throughput lower bound:

$$R_{\text{zf}}^{(\alpha)} \geq N_t \mathbf{E}\left[\log\left(1 + \frac{\chi_n^2}{\sum_{k=1,k\neq n}^{N} \chi_k^2 \min_{u\in\mathcal{J}_k} \epsilon_u}\right)\right]. \tag{37}$$

Using the above throughput lower bound, we prove the following theorem.

Theorem 3. *In the high SNR regime, the throughput scaling law for zero-forcing beamforming is given as follows.*

(1) *For $U\to\infty$, $N\to\infty$,*

$$\lim_{\substack{U\to\infty \\ N\to\infty}} \frac{C_{\text{zf}}^{(\alpha)}}{(N_t/(N_t-1))\log U + (N_t/(N_t-1))\log N} = 1. \tag{38}$$

(2) *For $U\to\infty$, N fixed*

$$\lim_{\substack{U\to\infty \\ N\to\infty}} \frac{C_{\text{zf}}^{(\alpha)}}{(N_t/(N_t-1))\log U} = 1. \tag{39}$$

The proof is given in Appendix E. The proof uses the uniform convergence in the weak law of large numbers. As before, Proposition 1 is applied to equate the scaling laws between the expected and the exact throughput.

A few remarks are in order.

(i) For $U\to\infty$, $N\to\infty$, the throughput scaling factor for zero-forcing beamforming upper bounds that for orthogonal beamforming (cf. (29)). Note that this does not imply the former is larger since the achievability of the same scaling factor for orthogonal beamforming is unknown.

(ii) The same scaling laws as in (3) have been also proved for downlink SDMA with limited feedback [6]. They are derived using a different approach based on the extreme value theory, though. This similarity demonstrates uplink-downlink duality.

(iii) As for orthogonal beamforming, the scheduling algorithm, which achieves the above scaling laws for zero-forcing beamforming, requires no feedback of channel power.

6. THROUGHPUT SCALING: NORMAL SNR

In this section, the throughput scaling law for uplink SDMA in the normal SNR regime is analyzed. In this regime, neither the noise nor the interference dominates, thus the SINR and scheduling metric are given, respectively, in (4) and (5). The throughput scaling law for orthogonal beamforming and zero-forcing beamforming are analyzed separately in Sections 6.1 and 6.2.

6.1. Orthogonal beamforming

In this section, the throughput scaling factor for orthogonal beamforming is obtained by deriving an upper bound and an achievable lower bound of this factor.

The upper bound of the scaling factor is given in the following lemma. This upper bound also holds for the low SNR regime and the zero-forcing beamforming.

Lemma 8. *For both the normal and low SNR regimes, the throughput scaling factors for both orthogonal and zero-forcing beamforming are upper-bounded as*

$$\lim_{U\to\infty}\frac{R_{\mathrm{or/zf}}}{N_t\log\log U}\le 1. \tag{40}$$

The proof is similar to that for Lemma 4 and hence omitted. In the proof, the upper bound of the throughput scaling factor in (40) is derived by omitting interference. This implies that reducing interference by increasing the codebook size N has no effect on this upper bound. Thus it is unnecessary to consider the case of $N\to\infty$ in the analysis for the normal SNR regime.

The scheduling algorithm for achieving the equality in (40) is provided as follows. Define the user index sets

$$\mathcal{T}_n^{\prime(m)} = \left\{1\le u\le U \mid \mathbf{s}_u\in\mathcal{B}_n^{(m)}\left(\frac{1}{(\log U)^{N_t-1}}\right)\right\} \tag{41}$$

and a scalar $U_\beta := \exp(-d_{\min}/4)$. Then $\mathcal{T}_n^{\prime(m)}\subset\mathcal{I}_n^{(m)}$ for all $U\ge U_\beta$. From each set $\mathcal{T}_n^{\prime(m)}$, the user with the maximum channel power is selected. Next, among the selected users, up to N_t users are scheduled using the criterion of maximizing throughput. Using this scheduling algorithm and from (12), a lower-bound of the throughput is obtained as

$$\begin{aligned}R_{\mathrm{or}}&\ge\mathbf{E}\left[\max_{1\le m\le M}\sum_{n=1}^{N_t}\log\left(1+\frac{\gamma\max_{u\in\mathcal{T}_n^{(m)}}\rho_u}{1+\gamma\sum_{k=1,k\ne u}^{N_t}\max_{u'\in\mathcal{T}_k^{(m)}}\rho_{u'}(1/\log U)}\right)\right]\\&\qquad U\ge U_\beta\\&\ge\mathbf{E}\left[\sum_{n=1}^{N_t}\log\left(1+\frac{\gamma\max_{u\in\mathcal{T}_n^{(m)}}\rho_u}{1+\gamma\sum_{k=1,k\ne u}^{N_t}\max_{u'\in\mathcal{T}_k^{(m)}}\rho_{u'}(1/\log U)}\right)\right]\\&\qquad U\ge U_\beta.\end{aligned} \tag{42}$$

Using the above lower bound, we prove the following theorem.

Theorem 4. *In the normal SNR regime, the scaling law for orthogonal beamforming is*

$$\lim_{U\to\infty}\frac{C_{\mathrm{or}}}{N_t\log\log U}=1. \tag{43}$$

The proof is given in Appendix F. Again, the proof relies on the uniform convergence in the weak law of large numbers.

A few remarks are in order.

(i) The throughput in the normal SNR regime scales as $\log\log U$ but that in the high SNR regime increases as $\log U$. Therefore, the throughput scaling rate is much higher in the high SNR regime than in the normal SNR regime.

(ii) The scaling law in Theorem 4 shows the full multiplexing gain.

(iii) Besides quantized channel shapes, feedback of both channel power and quantization errors from users are required.

6.2. Zero-forcing beamforming

This section focuses on the throughput scaling law for zero-forcing beamforming in the normal SNR regime. A scheduling algorithm for achieving the scaling upper bound in Lemma 8 is constructed as follows. Define the index sets, $\{\mathcal{T}_n\}_{n=1}^N$, similar to (41) but based on the RVQ codebook for zero-forcing beamforming (cf. Section 3.3.2). Next, define a new index set

$$\mathcal{L}_k=\mathcal{J}_k\cap\left(\bigcup_{n=1}^{N}\mathcal{T}_n\right),\quad 1\le k\le N_t, \tag{44}$$

where $\mathcal{J}_k$ is given in (35). From users in each of the sets $\{\mathcal{L}_k\}$, the one with the maximum channel power is scheduled. Thus, the index set of scheduled users is given as

$$\mathcal{A} = \left\{ \max_{u \in \mathcal{L}_k} \rho_u,\ 1 \le k \le N_t \right\}. \tag{45}$$

Using the above scheduling algorithm, we obtain the following theorem by proving the achievability of the throughput-scaling upper bound in Lemma 8.

Theorem 5. *In the normal SNR regime, the scaling law for zero-forcing beamforming is*

$$\lim_{U \to \infty} \frac{C_{\text{zf}}}{N_t \log \log U} = 1. \tag{46}$$

The proof is given in Appendix G. The proof involves repeated applications of Lemma 3, which show the uniform convergence of the numbers of users in the index sets $\{\mathcal{T}_n\}$ and $\mathcal{J}_n$ defined (35), respectively.

Comparing Theorems 4 and 5, the same scaling law holds for both orthogonal and zero-forcing beamforming in the normal SNR regime. Furthermore, this scaling law is identical to that for downlink SDMA with limited feedback [6, 8, 17].

7. THROUGHPUT SCALING: LOW SNR

In this section, the analysis of the throughput scaling law for uplink SDMA focuses on the lower SNR regime where channel noise is dominant. In this regime, the expected SINR in (5), denoted as $\Psi^{(\beta)}$, reduces to $\gamma \rho_u$. The following analysis is presented in Sections 7.1 and 7.2, which correspond, respectively, to orthogonal and zero-forcing beamforming.

7.1. Orthogonal beamforming

In the lower SNR regime, the throughput scaling law for orthogonal beamforming is obtained by achieving the upper bound for the throughput scaling factor in Lemma 8 using a specific scheduling algorithm. Denote the expected and exact throughput as $R_{\text{or}}^{(\beta)}$ and $C_{\text{or}}^{(\beta)}$, respectively.

A suitable scheduling algorithm can be modified from that in Section 6.1 by replacing the index sets in (41) with the following ones:

$$\check{\mathcal{T}}_n^{(m)} = \{1 \le u \le U \mid \mathbf{s}_u \in \mathcal{B}_n^{(m)}((d_{\min}/4)^{N_t-1})\}, \quad 1 \le m \le M, 1 \le n \le N_t. \tag{47}$$

Note that $\check{\mathcal{T}}_n^{(m)} \cap \check{\mathcal{T}}_{n'}^{(m')} = \varnothing$ for all $(m,n) \ne (m',n')$. The modified scheduling algorithm leads to the following throughput lower bound:

$$R_{\text{or}}^{(\beta)} \ge N_t \mathbf{E}\left[\log\left(1 + \gamma \max_{u \in \check{\mathcal{T}}_n^{(m)}} \rho_u\right)\right]. \tag{48}$$

Using the above throughput lower bound, the throughput scaling law is obtained and summarized in the following theorem.

Theorem 6. *In the low SNR regime, the scaling law of uplink SDMA with orthogonal beamforming is given as*

$$\lim_{U \to \infty} \frac{C_{\text{or}}^{(\beta)}}{N_t \log \log U} = 1. \tag{49}$$

The proof is similar to that for Theorem 4. Specifically, the proof uses the result of the extreme value theory in (B.6) and Lemma 3 of the uniform convergence in the weak law of large numbers. The details of the proof are omitted.

Comparing Theorems 4 and 6, the scaling laws in the normal and the low SNR regimes are identical. The intuition is that the interference power decreases continuously with U. Thus, for a large U, both the low and normal SNR regimes become noise limited, resulting in the same throughput scaling laws.

7.2. Zero-forcing beamforming

As in the last section, the derivation of the throughput scaling law for zero-forcing beamforming in the low SNR regime relies on the use of a specific scheduling for achieving the scaling upper bound in Lemma 8. This scheduling algorithm is simplified from that in Section 6.2 as follows. For the current algorithm, the scheduled users are selected from the index sets $\{\mathcal{J}_k\}$ in (35) rather than $\{\mathcal{L}_k\}$ as in Section 6.2. Consequently, the index set of scheduled users is

$$\mathcal{A} = \left\{ \max_{u \in \mathcal{L}_k} \rho_u,\ 1 \le k \le N_t \right\}. \tag{50}$$

Using the above scheduling algorithm, we prove the following theorem.

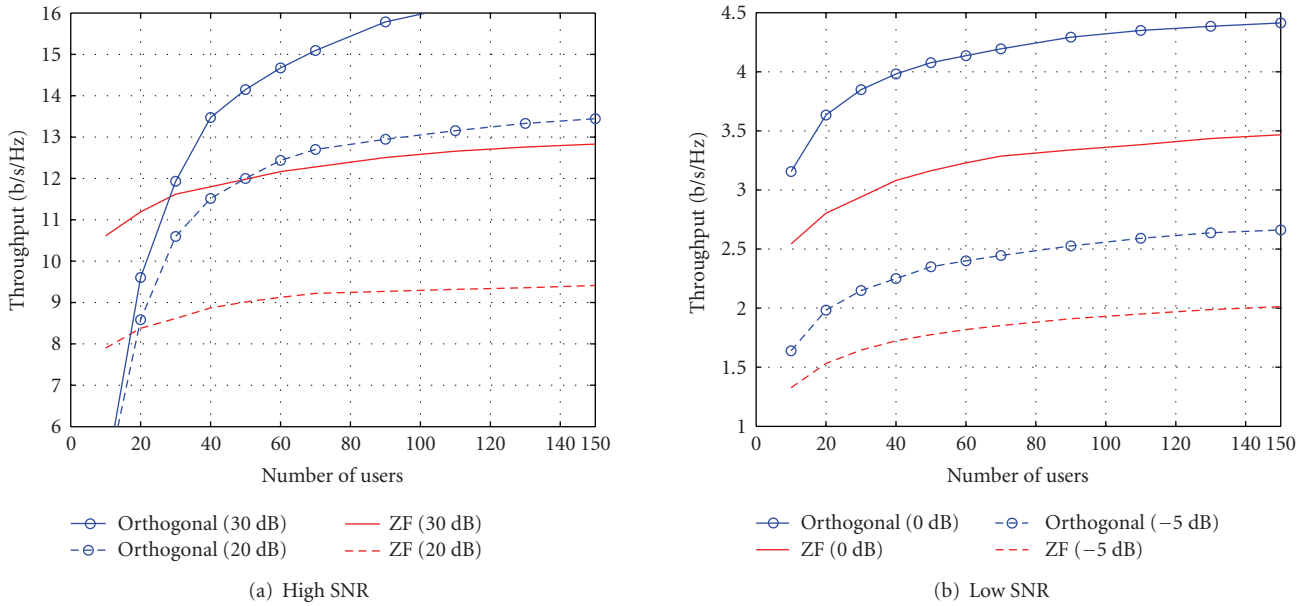
Theorem 7. *In the low SNR regime, the scaling law for zero-forcing beamforming is*

$$\lim_{U \to \infty} \frac{C_{\text{zf}}}{N_t \log \log U} = 1. \tag{51}$$

The proof is a simplified version of that for Theorem 7 due to the similarity in scheduling algorithms. Unlike the previous proof, the current proof requires only one-time application of Lemma 3. Similar remarks for Theorem 6 are also applicable here.

8. NUMERICAL RESULTS

In this section, based on simulation, orthogonal and zero-forcing beamforming are compared in terms of uplink SDMA throughput for an increasing number of users U. Such a comparison is to evaluate the throughput difference between orthogonal and zero-forcing beamforming in the practical regime of U. Note that the throughput scaling laws derived in previous sections indicate the same slopes for the throughput versus U curves for both beamforming methods in the asymptotic regime of U. Furthermore, uplink SDMA with limited feedback is compared with uplink channel-aware random access proposed in [28], which requires no CSI feedback.



(a) High SNR (b) Low SNR

FIGURE 3: Throughput comparisons between orthogonal and zero-forcing beamforming for uplink SDMA in (a) the high SNR regime and (b) the low SNR regime. The number of antennas at the base station is $N_t = 2$ and the quantizer codebook size is $N = 8$. The plotted values in brackets specify the SNR values in dB.

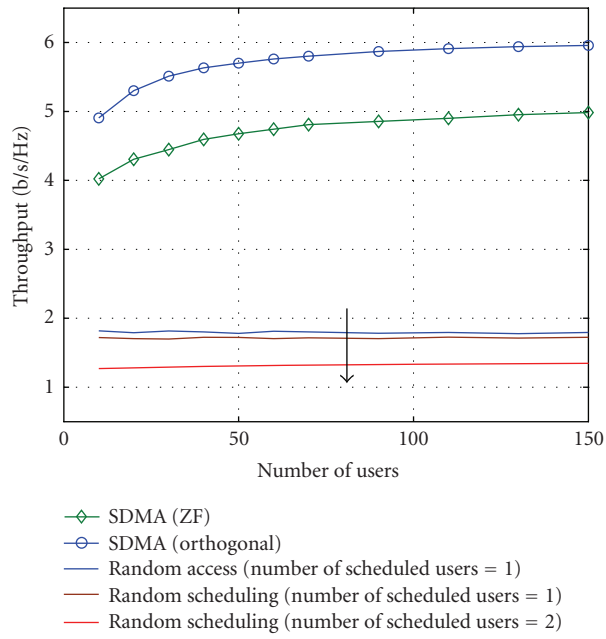


FIGURE 4: Throughput comparisons between uplink SDMA with limited feedback, SDMA with random scheduling, and uplink random access in [28]. The number of antennas at the base station is $N_t = 2$; the quantizer codebook size is $N = 8$; the SNR = 5 dB.

Orthogonal and zero-forcing beamforming are compared for both the high and the low SNR regimes. For simulation, the scheduling criterion is minimum quantization error in the high SNR regime and maximum channel power in the low SNR regime. These criteria are shown to achieve optimum throughput scaling in Sections 5 and 7. For zero-forcing beamforming, the scheduling algorithms are modified from that proposed in [6] by using the above criterions in greedy-search scheduling. For orthogonal beamforming, the scheduling algorithms are identical to those proposed in Sections 5.1 and 7.1. The throughput of orthogonal and zero-forcing beamforming are compared in Figure 3 for an increasing number of users U. For this comparison, the number of antennas is $N_t = 2$, the quantizer codebook size is $N = 8$, and the SNRs are $\{-5, 0\}$ dB for the low SNR regime and $\{20, 30\}$ dB for the high SNR regime. Several observations are made from Figure 3. First, as shown Figure 3(a) for the high SNR regime, orthogonal beamforming provides higher (smaller) throughput than zero-forcing beamforming if the number of users is large (small). The crossing point between the curves for orthogonal and zero-forcing beamforming is at $U = 20$ for SNR = 20 dB and at $U = 28$ for SNR = 30 dB. Second, from Figure 3(b) for the low SNR regime, orthogonal beamforming always achieves higher throughput than zero-forcing beamforming. Note that for $U \to \infty$, the curves for orthogonal and zero-forcing beamforming have identical slops according to the throughput scaling laws.

In Figure 4, the throughput of uplink SDMA is compared with that of SDMA with random scheduling and uplink random access [28], both of which require no CSI feedback. For SDMA with random scheduling, a random set of users is scheduled and their beamformers are columns of a random orthonormal basis. Note that with single-scheduled users, SDMA with random scheduling reduces to TDMA. For uplink random access, transmitting users are selected distributively using a channel power threshold, which increases with the total number of users [28]. For fair comparison, the uplink random access design originally proposed in [28] for

SISO channels is modified to allow transmit beamforming at each user who has N_t antennas. For uplink SDMA with limited feedback, the scheduling algorithms used in the previous comparison for the low SNR regime are applied. The simulation parameters are SNR = 5 dB, $N_t = 2$, and $N = 8$. Several observations are made from Figure 4. First, the throughput for uplink SDMA is much higher than that of SDMA with random scheduling and uplink random access. The throughput gains of uplink SDMA result from scheduling at the base station and the support of N_t simultaneous users. Second, the throughput of SDMA with random scheduling and uplink random access is insensitive to changes on the number of users U for the following reasons. Without giving preference to users with large channel power, random scheduling is incapable of exploiting multiuser diversity. Next, uplink random access achieves the throughput scaling of log log U but such a function grows extremely slowly with U. In summary, uplink SDMA outperforms SDMA with random scheduling and uplink random access in [28] by a large margin at the expense of finite-rate feedback from each user. Note that it is possible to schedule feedback users so as to constraint the total feedback overhead for uplink SDMA by following an approach similar to those proposed in [18, 29].

9. CONCLUSION

In this paper, the scaling law of uplink SDMA with limited feedback is analyzed for different SNR regimes and both orthogonal and zero-forcing beamforming. In the high SNR regime and for orthogonal beamforming, for an increasing quantizer codebook size, the throughput scales logarithmically with both the number of users and the codebook size; for a fixed codebook size, the throughput scales logarithmically only with the codebook size. For both cases, the linear scaling factor is smaller than the number of antennas, indicating the loss in spatial multiplexing gain. Similar results are obtained for zero-forcing beamforming. In the normal SNR regime, for both orthogonal zero-forcing beamforming, the throughput is found to scale double logarithmically with the number of users and linearly with the number of antennas. The same results are obtained for the low SNR regime.

Simulation results suggest that orthogonal and zero-forcing beamforming achieve different uplink throughput in nonasymptotic regimes even though they may follow the same throughput scaling laws asymptotically. For a small SNR or a large SNR coupled with many users, orthogonal beamforming outperforms zero-forcing beamforming. The reverse is true for a large SNR and a small number of users.

The analysis in this paper opens several interesting topics for future investigation. First, how to design a scheduling algorithm for uplink SDMA with orthogonal beamforming that achieves the optimum throughput scaling factor in the high SNR regime? Second, how to design scheduling algorithms for maximizing uplink SDMA throughput in practical regimes? Note that the scheduling algorithms discussed in this paper only ensure the asymptotic throughput scaling. Third, how to select feedback users for reducing the sum feedback rate along the vein of [18, 29]? Last, what is the relative efficiency of limited and analog feedback?

APPENDICES

A. PROOF OF PROPOSITION 1

Using the triangular inequality, $|\angle(\mathbf{v}_u, \hat{\mathbf{s}}_u) - \angle(\mathbf{s}_u, \hat{\mathbf{s}}_u)| \le \angle(\mathbf{v}_u, \mathbf{s}_u) \le \angle(\mathbf{v}_u, \hat{\mathbf{s}}_u) + \angle(\mathbf{s}_u, \hat{\mathbf{s}}_u)$. By definitions of φ_u and θ_u, the above expression can be rewritten as

$$|\varphi_u - \theta_u| \le \angle(\mathbf{v}_u, \mathbf{s}_u) \le \varphi_u + \theta_u. \tag{A.1}$$

From the given condition $\varphi_u + \theta_u \le \varphi_0 + \theta_0 < \pi/2$ and (A.1), $\cos(\angle(\mathbf{v}_u, \mathbf{s}_u)) \ge \cos(\varphi_0 + \theta_0)$. Using this inequality, (9), and (4), then

$$\begin{aligned} C &= \mathbf{E}\left[\sum_{u\in\mathcal{A}} \log\left(1 + \frac{\gamma\rho_u \cos^2(\angle(\mathbf{v}_u, \mathbf{s}_u))}{1+\gamma\sum_{m\in\mathcal{A}, m\neq u}\rho_m\epsilon_m\beta_{m,u}}\right)\right] \\ &\overset{(a)}{\ge} \mathbf{E}\left[\sum_{u\in\mathcal{A}} \log\left(1 + \frac{\gamma\rho_u \cos^2(\varphi_0 + \theta_0)}{1+\gamma\sum_{m\in\mathcal{A}, m\neq u}\rho_m\epsilon_m}\right)\right] \\ &\ge \mathbf{E}\Bigg[|\mathcal{A}| \log(\cos^2(\varphi_0 + \theta_0)) \\ &\qquad + \sum_{u\in\mathcal{A}} \log\left(1 + \frac{\gamma\rho_u}{1+\gamma\sum_{m\in\mathcal{A}, m\neq u}\rho_m\epsilon_m}\right)\Bigg] \\ &= |\mathcal{A}| \log(\cos^2(\varphi_0 + \theta_0)) + R \\ &\ge \log(\cos^2(\varphi_0 + \theta_0)) + R. \end{aligned} \tag{A.2}$$

For (a), note that $\beta_{m,n} \le 1$. Next, an upper bound is obtained for the throughput C as follows:

$$\begin{aligned} C &\le \mathbf{E}\left[\sum_{u\in\mathcal{A}} \log\left(1 + \frac{\gamma\rho_u}{1+\min_{m\in\mathcal{A}, m\neq u}\beta_{m,u}\gamma\sum_{m\in\mathcal{A}, m\neq u}\rho_m\epsilon_m}\right)\right] \\ &\le \mathbf{E}\Bigg[\sum_{u\in\mathcal{A}} \log\Bigg(\left(1\Big/\min_{\substack{m\in\mathcal{A}\\ m\neq u}}\beta_{m,u}\right) \\ &\qquad + \left(\gamma\rho_u\Big/\left(\min_{\substack{m\in\mathcal{A}\\ m\neq u}}\beta_{m,u} + \min_{\substack{m\in\mathcal{A}\\ m\neq u}}\beta_{m,u}\gamma\sum_{\substack{m\in\mathcal{A}\\ m\neq u}}\rho_m\epsilon_m\right)\right)\Bigg)\Bigg] \\ &\le \mathbf{E}\left[\sum_{u\in\mathcal{A}} \log\left(1 + \frac{\gamma\rho_u}{1+\gamma\sum_{m\in\mathcal{A}, m\neq u}\rho_m\epsilon_m}\right)\right] \\ &\quad + |\mathcal{A}|\mathbf{E}\left[-\log\left(\min_{m\in\mathcal{A}, m\neq u}\beta_{m,u}\right)\right] \\ &\overset{(a)}{\le} R + |\mathcal{A}|\frac{\log(|\mathcal{A}|-1)+1}{N_t - 1} \\ &\le R + \frac{N_t}{N_t - 1}[\log(N_t - 1) + 1], \end{aligned} \tag{A.3}$$

where the inequality (a) is obtained by applying Lemma 1. Combining (A.2) and (A.3) gives the desired result.

B. PROOF OF LEMMA 4

From (6) and given Assumption 4,

$$R \overset{(a)}{\leq} \mathbf{E}\left[\max_{1\leq m\leq M}\sum_{n=1}^{N_t}\log\left(\frac{1}{\max_{u'\in\mathcal{I}_k^{(m)}}\epsilon_{u'}} + \frac{(\log U+c)\max_{u\in\mathcal{I}_n^{(m)}}\rho_u}{\min_{u'\in\mathcal{I}_k^{(m)}}\epsilon_{u'}}\right)\right]$$

$$\leq \mathbf{E}\left\{\max_{1\leq m\leq M}\sum_{n=1}^{N_t}\left[\log\left(1+(\log U+c)\max_{u\in\mathcal{I}_n^{(m)}}\rho_u\right) - \log\left(\min_{u'\in\mathcal{I}_k^{(m)}}\epsilon_{u'}\right)\right]\right\}$$

$$\overset{(b)}{\leq} N_t\mathbf{E}\left[\log\left(1+(\log U+c)\max_{1\leq u\leq U}\rho_u\right)\right] - N_t\mathbf{E}\left[\log\left(\min_{1\leq m\leq M}\min_{u'\in\mathcal{I}_k^{(m)}}\epsilon_{u'}\right)\right] \tag{B.4}$$

$$= N_t\mathbf{E}\left[\log\left(1+(\log U+c)\max_{1\leq u\leq U}\rho_u\right)\right] - \Pi. \tag{B.5}$$

The inequality (a) follows from $(\max_{u'\in\mathcal{I}_k^{(m)}}\epsilon_{u'} \leq 1)$. The inequality (b) is obtained by moving the "max" operator in (B.4) into the summation term. The definition of Π in (B.5) is obvious.

The following result is well known from extreme value theory (see, e.g., [8, Equation (A10)]):

$$\Pr\left(\left|\max_{1\leq u\leq U}\rho_u - \log U\right| < O(\log\log U)\right) > 1 - O\left(\frac{1}{\log U}\right). \tag{B.6}$$

From (B.5) and (B.6),

$$\begin{aligned} R &\leq N_t\log\{1+(\log U+c)[\log U + O(\log\log U)]\}\\ &\quad\times\left[1-O\left(\frac{1}{\log U}\right)\right]\\ &\quad + N_t\mathbf{E}\left[\log\left(1+(\log U+c)\sum_{u=1}^{U}\rho_u\right)\right]O\left(\frac{1}{\log U}\right)+\Pi\\ &\leq O(\log\log U) + N_t\log\{1+(\log U+c)U\mathbf{E}[\rho_u]\}\\ &\quad\times O\left(\frac{1}{\log U}\right)+\Pi = O(\log\log U)\\ &\quad + N_t\log[1+(\log U+c)UN_t]O\left(\frac{1}{\log U}\right)+\Pi. \end{aligned} \tag{B.7}$$

Last, a close-form expression is derived for Π defined in (B.5). Since $\bigcup_{1\leq m\leq M}\mathcal{I}_n^{(m)} \in \{u \mid 1\leq u\leq U\}$, Π is upper-bounded as

$$\begin{aligned} \Pi &\leq N_t\mathbf{E}\left[-\log\left(\min_{1\leq u\leq U}\epsilon_u\right)\right]\\ &\overset{(a)}{=} N_t\mathbf{E}\left[-\log\left(\min_{1\leq u\leq U}\min_{1\leq n\leq M}N_t^{-1/(N_t-1)}\beta_{u,n}\right)\right]\\ &\overset{(b)}{=} \frac{N_t}{N_t-1}\left[\log(MU)+\log N_t+\frac{1}{N_t-1}\right]. \end{aligned} \tag{B.8}$$

The equality (a) is a property of the quantization for orthogonal beamforming [17] where $\{\beta_{u,n}\}$ are i.i.d. delta random variables, (b) is obtained by applying Lemma 1. The desired result follows from (B.8) and (B.7).

C. PROOF OF LEMMA 5

The proof is divided according to three cases: $N = o(U)$, $N = \Theta(U)$, and $N = O(U)$. Only the proof for the case $N = o(U)$ is presented below and those for other two cases are omitted due to their similarity.

To begin, a bins-and-balls model is constructed for multiuser feedback of quantized channel shapes as follows. In this model, the U balls of the channel shapes of U users, which are i.i.d. points, uniformly distributed on the surface of the unit hyper sphere. The small bins (cf. Section 4.1) are N congruent disks on the unit hyper-sphere as defined below:

$$\mathcal{B}_n^{(m)}(A) = \{\mathbf{s}\in\mathbb{C}^{N_t} \mid \|\mathbf{s}\|^2 = 1,\ 1-\left|\mathbf{s}^\dagger\mathbf{v}_n^{(m)}\right|^2 \leq A^{1/(N_t-1)}\},\quad 1\leq m\leq M,\ 1\leq n\leq N_t, \tag{C.9}$$

where $A = 1/U$ is the disk volume. Note that the volume of the big bin is $1-N/U$. Following this definition, each disk (or small bin) is centered at a code vector in the codebook $\mathcal{F}$ (cf. Section 3.3.1) and has a volume $1/U$. The set of balls inside the small bin $\mathcal{B}_n^{(m)}$ is specified by the following index set:

$$\mathcal{T}_n^{(m)} = \{1\leq u\leq U \mid \mathcal{B}_n^{(m)}(U^{-1})\}. \tag{C.10}$$

Therefore, the number of balls in $\mathcal{B}_n^{(m)}$ is $T_n^{(m)} = |\mathcal{T}_n^{(m)}|$. Define the mth cluster of small bins as $\{\mathcal{B}_n^{(m)},\ 1\leq n\leq N_t\}$. Furthermore, define the index set for nonempty clusters:

$$\mathcal{Q} = \{1\leq m\leq M:\ \mathcal{T}_n^{(m)}\neq\varnothing\ \forall 1\leq n\leq N_t\}. \tag{C.11}$$

Thus the number of nonempty clusters is $Q = |\mathcal{Q}|$. The above bins-and-balls model allows us to apply Lemma 1 for characterizing Q. Specifically, Q satisfies (18) with $p = 1/U$.

Next, we derive the probability that a small bin lies inside a Voronoi cell, namely, $\Pr(\mathcal{T}_n^{(m)} \subset \mathcal{I}_n^{(m)})$, where $\mathcal{T}_n^{(m)}$ and $\mathcal{I}_n^{(m)}$ are defined in (C.10) and (13), respectively. This probability conditioned on a nonempty bin $\mathcal{T}_n^{(m)}\neq\varnothing$ is given in the following lemma.

Lemma 9. *The index sets $\mathcal{T}_n^{(m)}$ and $\mathcal{I}_n^{(m)}$ have the following relationship:*

$$\Pr(\mathcal{T}_n^{(m)}\subset\mathcal{I}_n^{(m)} \mid \mathcal{T}_n^{(m)}\neq\varnothing) \geq \left(1-\frac{N_t4^{N_t-1}}{U}\right)^{M-1}. \tag{C.12}$$

Proof. Define $\sin^2\theta = U^{-1/(N_t-1)}$. Given $\mathcal{T}_n^{(m)}\neq\varnothing$, a sufficient condition for $u\in\mathcal{T}_n^{(m)} \Rightarrow u\in\mathcal{I}_n^{(m)}$ is $1-|\mathbf{v}^\dagger\mathbf{v}_n^{(m)}|^2 \geq \sin^2(2\theta)$

for all $\mathbf{v} \in \mathcal{F}$ and $\mathbf{v} \neq \mathbf{v}_n^{(m)}$, whose proof is straightforward and hence omitted. Using this sufficient condition,

$$\begin{aligned}\Pr\left(\mathcal{T}_n^{(m)} \subset \mathcal{I}_n^{(m)} \mid \mathcal{T}_n^{(m)} \neq \varnothing\right) &\geq \Pr\left(1 - \left|\mathbf{v}^\dagger \mathbf{v}_n^{(m)}\right|^2 \geq \sin^2(2\theta)\ \forall \mathbf{v} \in \mathcal{F}, \mathbf{v} \neq \mathbf{v}_n^{(m)}\right) \\ &\overset{(a)}{=} \left[1 - N_t(\sin 2\theta)^{2(N_t-1)}\right]^{M-1} \\ &\geq \left[1 - N_t\left(4\sin^2\theta\right)^{(N_t-1)}\right]^{M-1},\end{aligned} \tag{C.13}$$

where (a) is a property of the quantization codebook for orthogonal beamforming, which consists of M randomly generated orthonormal sets [17]. The desired result follows from the last equation and the definition of $\sin\theta$. □

To use the result based on the bins-and-balls model, the following variable is defined by replacing $\mathcal{I}_n^{(m)}$ in (25) with $\mathcal{T}_n^{(m)}$:

$$\widetilde{\epsilon} = \min_{1\leq m\leq M} \max_{1\leq n\leq N_t} \min_{u\in\mathcal{T}_k^{(m)}} \epsilon_u. \tag{C.14}$$

A useful result is provided in the following lemma.

Lemma 10. *The mean of* $\widetilde{\epsilon}$ *in* (C.14) *is upper-bounded as*

$$\mathbf{E}[\,\widetilde{\epsilon}\,] \leq U^{-1/(N_t-1)}\mathbf{E}[Q^{-1/N_t(N_t-1)}], \tag{C.15}$$

where $Q := |\mathcal{Q}|$ *and* $\mathcal{Q}$ *is defined in* (C.11).

Proof. From (C.14) and the definition of $\mathcal{Q}$ in (C.11),

$$\mathbf{E}[\,\widetilde{\epsilon}\,] \leq \mathbf{E}\left[\min_{m\in\mathcal{Q}} \max_{1\leq n\leq N_t} \min_{u\in\mathcal{T}_k^{(m)}} \epsilon_u\right]. \tag{C.16}$$

For $u \in \mathcal{T}_n^{(m)}$, $\epsilon_u \cong U^{-1/(N_t-1)}\beta$, where β is a beta random variable (cf. Lemma 1) and $\cong$ denotes the equivalence in distribution. Therefore, from (C.16) and the definition of $\mathcal{T}_n^{(m)}$ in (C.10), then

$$\begin{aligned}\mathbf{E}[\,\widetilde{\epsilon}\,] &\leq \mathbf{E}\left[\min_{m\in\mathcal{Q}} \max_{1\leq n\leq N_t} \epsilon_u \mid u \in \mathcal{T}_k^{(m)}\right] \\ &= \mathbf{E}\left[\min_{m\in\mathcal{Q}} \max_{1\leq n\leq N_t} U^{-1/(N_t-1)}\beta_{m,n}\right].\end{aligned} \tag{C.17}$$

Next, a close-form expression is derived for the lower bound in (C.17). Since

$$\Pr\left(\max_{1\leq n\leq N_1} \beta_{m,n} \leq \epsilon_0\right) = \epsilon_0^{(N_t-1)N_1}, \tag{C.18}$$

we have $\Pr(\min_{1\leq m\leq N_2} \max_{1\leq n\leq N_1} \beta_{m,n} \geq \epsilon_0) = (1 - \epsilon_0^{(N_t-1)N_1})^{N_2}$. Using the above CDF and following the similar procedure in the proof of Lemma 1 (cf. [21]), we obtain that

$$\mathbf{E}\left[\min_{m\in\mathcal{Q}} \max_{1\leq n\leq N_t} \beta_{m,n}\right] = \mathbf{E}[Q^{-1/N_t(N_t-1)}]. \tag{C.19}$$

The desired result follows from the last equation and (C.17). □

Using the above results, the lower bound of the scaling factor of the expected throughput $R_{\text{or}}^{(\alpha)}$ is readily obtained as follows. From (24),

$$\begin{aligned}R_{\text{or}}^{(\alpha)} &\geq N_t\mathbf{E}[\log(\chi_n^2)] - N_t\mathbf{E}\left[\log\left(\sum_{\substack{k=1\\k\neq n}}^{N}\chi_k^2\right)\right] - N_t\mathbf{E}[\log\epsilon^\star] \\ &= O(1) - N_t\mathbf{E}[\log\epsilon^\star] \\ &\overset{(a)}{\geq} O(1) - N_t\log\mathbf{E}[\epsilon^\star] \\ &\geq O(1) - N_t\log\mathbf{E}[\widetilde{\epsilon}]\Pr\left(\widetilde{\epsilon} = \epsilon^\star\right) \\ &\overset{(b)}{\geq} O(1) - N_t\log\left(U^{-1/(N_t-1)}\mathbf{E}[Q^{-1/N_t(N_t-1)}]\right)\Pr\left(\widetilde{\epsilon} = \epsilon^\star\right) \\ &\overset{(c)}{\geq} O(1) - N_t\log\left\{U^{-1/(N_t-1)} \times\left[Mp - \sqrt{\log M\,\mathrm{var}(Q)}\right]^{-1/N_t(N_t-1)}\right\} \\ &\quad\times\Pr\left(\epsilon^\star = \widetilde{\epsilon}\right)\Pr\left(Q \geq Mp - \sqrt{\log M\mathrm{var}(Q)}\right) \\ &\overset{(d)}{\geq} O(1) + \left[\left(\frac{N_t}{N_t-1}\right)\log U + \frac{1}{N_t-1}\log M + O(1)\right] \\ &\quad\times\left(1 - \frac{N_t 4^{N_t-1}}{U}\right)^{M-1}\left(1 - \frac{1}{\log M}\right).\end{aligned} \tag{C.20}$$

The inequality (a) is the result of the Jensen's inequality; (b) is obtained by using Lemma 10; (c) results from Lemma 1. The inequality (d) is obtained using Lemma 9, (25), and (C.14). The desired result in Lemma 5 follows from the last inequality.

D. PROOF OF LEMMA 7

The idea for proof is summarized as follows. Consider a set of disks as defined in (C.9). A user is said to be in a disk if his/her channel shape belongs to the disk. First, the uniform convergence of the numbers of users in the N disks is shown using Lemma 3. Second, for a large number of users, a disk is shown to lie inside a corresponding Voronoi cell. With this result, considering only users in the disks rather than all results in a throughput lower bound that is tight for a large number of users.

Consider a set of N disks $\{\mathcal{B}_n^{(m)}(1/\log U)\}$ as defined in (C.9), each has a volume of $1/\log U$. A corollary of Lemma 3 is provided as follows. Define the index set of the users in the disk $\mathcal{B}_n^{(m)}(1/\log U)$:

$$\widehat{\mathcal{T}}_n^{(m)} = \left\{1 \leq u \leq U \mid \mathbf{s}_u \in \mathcal{B}_n^{(m)}\left(\frac{1}{\log U}\right)\right\}, \quad 1 \leq m \leq M,\ 1 \leq n \leq N_t. \tag{D.21}$$

The following corollary follows from Lemma 3 by substituting $A = 1/\log U$ and $\tau_1 = \tau_2 = 1/\log U$.

Corollary 2. *The numbers of users belonging to the index sets* (D.21) *satisfy*

$$\Pr\left(\min_{m,n} |\widehat{\mathcal{T}}_n^{(m)}| \geq \frac{2U}{\log U}\right) \geq 1 - \frac{1}{\log U} \quad \forall U \geq U_o, \tag{D.22}$$

where U *is defined in* (20) *with* $\tau_1 = \tau_2 = 1/\log U$.

Next, for a sufficiently large number of users, the disk $\mathcal{B}_n^{(m)}(1/\log U)$ is shown to lie inside the corresponding Voronoi cell. Define the minimum distance of the codebook $\mathcal{F}$ as

$$d_{\min} = \min_{\mathbf{v},\mathbf{v}' \in \mathcal{F}} 1 - |\mathbf{v}^\dagger \mathbf{v}'|^2. \tag{D.23}$$

Therefore, $\mathbf{s}_u \in \mathcal{B}_n^{(m)}((d_{\min}/4)^{N_t-1}) \Rightarrow u \in \mathcal{U}_{m,n}$. Using this fact and (D.21), there exists U_α such that for all $U \geq U_\alpha$, $\widehat{\mathcal{T}}_{m,n} \in \mathcal{U}_{m,n}$. In other words, users in the disk $\mathcal{B}_n^{(m)}$ must also lie in the corresponding Voronoi cell. Using this fact, a throughput lower bound follows by replacing $\mathcal{U}_{m,n}$ in (32) with $\widehat{\mathcal{T}}_{m,n}$:

$$R_{\text{or}}^{(\alpha)} \geq \mathbf{E}\left[\sum_{n=1}^{N_t} \log\left(1 + \frac{\chi_n^2}{\sum_{k=1,k\neq n}^{N} \chi_k^2 \min_{u\in\widehat{\mathcal{T}}_n^{(m)}} \epsilon_u}\right) \,\middle|\, \widehat{\mathcal{T}}_n^{(m)} \neq \varnothing \ \forall n,m\right] \Pr\left(\widehat{\mathcal{T}}_n^{(m)} \neq \varnothing \ \forall n,m\right) \quad \forall U \geq U_\alpha. \tag{D.24}$$

By applying Corollary 2 on (D.24),

$$\begin{aligned} R_{\text{or}}^{(\alpha)} &\geq \mathbf{E}\left[\sum_{n=1}^{N_t} \log\left(1 + \frac{\chi_n^2}{\sum_{k=1,k\neq n}^{N} \chi_k^2 \min_{u\in\mathcal{T}_{m,n}} \epsilon_u}\right) \,\middle|\, |\mathcal{T}_{m,n}| \geq \frac{U}{2\log U} \ \forall n,m\right]\left(1 - \frac{1}{2\log U}\right) \quad \forall U \geq \max\{U_o, U_\alpha\} \\ &\geq -N_t \mathbf{E}\left[\log\left(\sum_{\substack{k=1\\k\neq n}}^{N} \chi_k^2 \min_{u\in\widehat{\mathcal{T}}_n^{(m)}} \epsilon_u\right) \,\middle|\, |\mathcal{T}_{m,n}| \geq \frac{U}{2\log U} \ \forall n,m\right]\left(1 - \frac{1}{2\log U}\right) \\ &\quad + N_t \mathbf{E}[\log(\chi_n^2)]\left(1 - \frac{1}{2\log U}\right) \quad \forall U \geq \max\{U_o, U_\alpha\}. \end{aligned} \tag{D.25}$$

Using $\mathbf{E}[\log(\chi_n^2)] = O(1)$ and by applying Jensen's inequality, from (D.25),

$$\begin{aligned} R_{\text{or}}^{(\alpha)} &\geq -N_t \log\left(\sum_{\substack{k=1\\k\neq n}}^{N} \mathbf{E}[\chi_k^2] \mathbf{E}\left[\min_{u\in\mathcal{T}_{m,n}} \epsilon_u \mid |\widehat{\mathcal{T}}_n^{(m)}| \geq \frac{U}{2\log U}\right]\right) \times \left(1 - \frac{1}{2\log U}\right) + O(1) \\ &\geq -N_t \log\left(N_t(N_t-1)\mathbf{E}\left[\min_{u\in\mathcal{T}_{m,n}} \epsilon_u \mid |\widehat{\mathcal{T}}_n^{(m)}| \geq \frac{U}{2\log U}\right]\right) \times \left(1 - \frac{1}{2\log U}\right) + O(1). \end{aligned} \tag{D.26}$$

Furthermore, using Lemma 1,

$$R_{\text{or}}^{(\alpha)} \geq -N_t \log\left[N_t^2 \left(\frac{U}{2\log U}\right)^{-1/(N_t-1)}\right]\left(1 - \frac{1}{2\log U}\right) + O(1). \tag{D.27}$$

The desired result follows from the above equation.

E. PROOF OF THEOREM 3

The proof procedure is similar to that in Appendix D. To apply the theory of uniform convergence in the weak law of large numbers, the following corollary of Lemma 3 is obtained.

Corollary 3. *The number of users in the index sets* $\{\mathcal{J}_k\}$ *satisfies the following property:*

$$\Pr\left(\min_{1\leq n\leq N_t} |\mathcal{J}_n| \geq \tau_0^{N_t-1} U\right) < 1 - \frac{1}{U} \quad \forall U > U_o, \tag{E.28}$$

where U_o *is from* (20) *with* $\tau_1(U) = \tau_2(U) = 1/U$.

Using this corollary and (38),

$$\begin{aligned} R_{\text{zf}}^{(\alpha)} &\geq \mathbf{E}\left[\sum_{n=1}^{N_t} \log\left(1 + \frac{\chi_n^2}{\sum_{k=1,k\neq n}^{N} \chi_k^2 \min_{u\in\mathcal{J}_n} \epsilon_u}\right) \,\middle|\, \mathcal{J}_n \neq \varnothing \ \forall n,m\right] \Pr\left(\mathcal{J}_n \neq \varnothing \ \forall n,m\right) \\ &\geq \mathbf{E}\left[\sum_{n=1}^{N_t} \log\left(\frac{\gamma\chi_n^2}{\sum_{k=1,k\neq n}^{N} \chi_k^2 \min_{u\in\mathcal{J}_n} \epsilon_u}\right) \,\middle|\, \mathcal{J}_n \geq \tau_o^{N_t-1} U - 1 \ \forall n,m\right]\left(1 - \frac{1}{U}\right). \end{aligned} \tag{E.29}$$

Following similar steps in Appendix D, we obtain that

$$R_{\text{zf}}^{(\alpha)} \geq O(1) - N_t \log\left[\left(\frac{NU}{\tau_o^{N_t-1}}\right)^{-1/(N_t-1)}\right]\left(1 - \frac{1}{U}\right). \tag{E.30}$$

It follows that

$$\lim_{\substack{U\to\infty\\ N\to\infty}} \frac{R_{\mathrm{zf}}^{(\alpha)}}{(N_t/(N_t-1))\log U + (N_t/(N_t-1))\log N} \geq 1,$$
$$\lim_{U\to\infty} \frac{R_{\mathrm{zf}}^{(\alpha)}}{(N_t/(N_t-1))\log U} \geq 1. \tag{E.31}$$

From the above inequalities, (34), and Proposition 1, we obtain the desired scaling factors.

F. PROOF OF THEOREM 4

From the definition in (41) and by applying Lemma 3,

$$\Pr\left(|\mathcal{T}_n^{(m)}| \geq \frac{U}{(\log U)^{N_t-1}}\right) > 1 - \frac{1}{2(\log U)^{N_t-1}} \quad \forall U > U_o, \tag{F.32}$$

where U_o is from (20) with $\tau_1 = \tau_2 = 1/2(\log U)^{N_t-1}$. From (42) and (F.32),

$$\begin{aligned} R &\geq N_t \mathbf{E}\left[\log\left(1+\frac{\gamma \max_{u\in\mathcal{T}_n^{(m)}}\rho_u}{1+\gamma\sum_{k=1,k\neq u}^{N_t}\max_{u'\in\mathcal{T}_k^{(m)}}\rho_{u'}(1/\log U)}\right)\middle| \right. \\ &\qquad \left. |\mathcal{T}_n^{(m)}| \geq \frac{U}{(\log U)^{N_t-1}}\right]\left(1-\frac{1}{2(\log U)^{N_t-1}}\right) \\ &\qquad\qquad \forall U > U_1 \\ &\overset{(a)}{\geq} N_t \mathbf{E}\left[\log\left(1+\frac{\log\widetilde{U} - O(\log\log\widetilde{U})}{1/\gamma + [\log\widetilde{U} + O(\log\log\widetilde{U})](1/\log U)}\right)\right] \\ &\quad \times\left(1-\frac{1}{2(\log U)^{N_t-1}}\right)\left[1-O\left(\frac{1}{\log U}\right)\right]^{N_t} \\ &\qquad \text{where } \widetilde{U} = \frac{U}{(\log U)^{N_t-1}}. \end{aligned} \tag{F.33}$$

It follows from the last equation that $\lim_{U\to\infty}(R/N_t \log\log U) \geq 1$. The desired result is obtained by combining the above inequality, Lemma 8, and Proposition 1.

G. PROOF OF THEOREM 5

Define the index set

$$\mathcal{L}_n = \left\{1 \leq u \leq U \mid \mathbf{s}_u \in \mathcal{B}_n\left(\frac{1}{2(\log U)^{N_t-1}}\right)\right\} \quad 1 \leq n \leq N_t. \tag{G.34}$$

By applying Lemma 3,

$$\Pr\left(|\mathcal{L}_n| \geq \frac{U}{(\log U)^{N_t-1}}\right) > 1 - \frac{1}{2(\log U)^{N_t-1}} \quad \forall U > U_1, \tag{G.35}$$

where $U_1 = \max\{(3/2)(\log U)^{N_t-1}\log[10c(\log U)^{N_t-1}], (4/2)(\log U)^{N_t-1}\log[2(\log U)^{N_t-1}]\}$. There exists U_0 such that $\mathcal{L}_n \cap \mathcal{L}_n' = \varnothing$ for all $n \neq n'$.

Next, define $J_n = |\mathcal{J}_n \cap (\bigcup_{n=1}^{N_t}\mathcal{L}_n)|$ and $L = |\bigcup_{n=1}^{N_t}\mathcal{L}_n|$. Again, by applying Lemma 3,

$$\Pr\left(J_n \geq \tau_o^{N_t-1}L - \log L\right) > 1 - \frac{\log L}{L} \quad \forall L \geq L_1, \tag{G.36}$$

where $L_1 = \max\{(3L/\log L)\log[10c(L/\log L)], (4L/\log L)\log[2(L/\log L)]\}$. Denote $\widetilde{U} = U/(\log U)^{N_t-1}$:

$$\begin{aligned} R &\geq \sum_{n=1}^{N_t}\mathbf{E}\left[\log\left(1+\frac{\gamma\max_{u\in\mathcal{J}_n}\rho_u}{1+\gamma\sum_{k=1,k\neq n}^{N}\max_{u'\in\mathcal{J}_k}\rho_{u'}(2^{-1/(N_t-1)}/\log U)}\right)\middle| \right. \\ &\qquad \left. L \geq \widetilde{U}\right]\Pr(L \geq \widetilde{U}) \\ &\geq \sum_{n=1}^{N_t}\mathbf{E}\left[\log\left(1+\frac{\gamma\max_{u\in\mathcal{J}_n}\rho_u}{1+\gamma\sum_{k=1,k\neq n}^{N}\max_{u'\in\mathcal{J}_k}\rho_{u'}(2^{-1/(N_t-1)}/\log U)}\right)\middle| \right. \\ &\qquad \left. J_n \geq \tau_o^{N_t-1}\widetilde{U} - \log\widetilde{U}\right] \\ &\Pr\left(J_n \geq \tau_o^{N_t-1}\widetilde{U} - \log\widetilde{U} \mid L \geq \widetilde{U}\right)\Pr(L \geq \widetilde{U}) \\ &\geq \sum_{n=1}^{N_t}\mathbf{E}\left[\log\left(1+\frac{\gamma\log U}{1+\gamma\log U(2^{-1/(N_t-1)}/\log U)}\right)\right] \\ &\qquad\qquad U \longrightarrow \infty. \end{aligned} \tag{G.37}$$

The desired result following from the last inequality and Proposition 1.

ACKNOWLEDGMENTS

Kaibin Huang is the recipient of a Motorola Partnerships in Research Grant. This work is funded by the National Science Foundation under Grants nos. CCF-514194 and CNS-435307.

REFERENCES

[1] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.

[2] J. G. Andrews, "Interference cancellation for cellular systems: a contemporary overview," *IEEE Wireless Communications Magazine*, vol. 12, no. 2, pp. 19–29, 2005.

[3] D. Gesbert, M. Kountouris, R. W. Heath Jr., C.-B. Chae, and T. Sälzer, "Shifting the MIMO paradigm," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 36–46, 2007.

[4] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[5] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, 2004.

[6] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1478–1491, 2007.

[7] W. Choi, A. Forenza, J. G. Andrews, and R. W. Heath Jr., "Opportunistic space division multiple access with beam selection," *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2371–2380, 2007.

[8] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[9] M. Kountouris, R. De Francisco, D. Gesbert, D. T.M. Slock, and T. Sälzer, "Efficient metrics for scheduling in MIMO broadcast channels with limited feedback," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 3, pp. 109–112, 2007.

[10] S. C. Swales, M. A. Beach, D. J. Edwards, and J. P. McGeehan, "Performance enhancement of multibeam adaptive basestation antennas for cellular land mobile radio systems," *IEEE Transactions on Vehicular Technology*, vol. 39, no. 1, pp. 56–67, 1990.

[11] S. Anderson, M. Millnert, M. Viberg, and B. Wahlberg, "adaptive array for mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 1, pp. 230–236, 1991.

[12] H. Yin and H. Liu, "Performance of space-division multiple-access (SDMA) with scheduling," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 611–618, 2002.

[13] J. H. Winters, J. Salz, and R. D. Gitlin, "Impact of antenna diversity on the capacity of wireless communication systems," *IEEE Transactions on Communications*, vol. 42, no. 234, pp. 1740–1751, 1994.

[14] F. Shad, T. D. Todd, V. Kezys, and J. Litva, "Dynamic slot allocation (DSA) in indoor (SDMA)/(TDMA) using a smart antenna basestation," *IEEE/ACM Transactions on Networking*, vol. 9, no. 1, pp. 69–81, 2001.

[15] D. J. Love, R. W. Heath Jr., W. Santipach, and M. L. Honig, "What is the value of limited feedback for MIMO channels?" *IEEE Communications Magazine*, vol. 42, no. 10, pp. 54–59, 2004.

[16] T. L. Marzetta and B. M. Hochwald, "Fast transfer of channel state information in wireless systems," *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1268–1278, 2006.

[17] K. Huang, J. G. Andrews, and R. W. Heath Jr., "Orthogonal beamforming for SDMA downlink with limited feedback," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 3, pp. 97–100, 2007.

[18] C. Swannack, G. W. Wornell, and E. Uysal-Biyikoglu, "MIMO broadcast scheduling with quantized channel state information," in *Proceedings of the IEEE International Symposium on Information Theory*, pp. 1788–1792, July 2006.

[19] Samsung Electronics, "Downlink MIMO for EUTRA," Technical Specification TSG RAN WG1 # 44/R1-060335, 3GPP, 2006.

[20] Motorola Inc, "Downlink MIMO summary," Technical specification TSG RAN WG1 # 49-bis/R1-072693, 3GPP, 2007.

[21] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5059, 2006.

[22] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Applicationss*, vol. 16, no. 2, pp. 264–280, 1971.

[23] M. Mitzenmacher and E. Upfal, *Probability and Computing*, Cambridge University Press, New York, NY, USA, 2005.

[24] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[25] K. Zyczkowski and M. Kus, "Random unitary matrices," *Journal of Physics*, vol. A27, pp. 4235–4245, 1994.

[26] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2735–2747, 2003.

[27] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[28] X. Qin and R. A. Berry, "Distributed approaches for exploiting multiuser diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 392–413, 2006.

[29] K.-B. Huang, R. W. Heath Jr., and J. G. Andrews, "SDMA with a sum feedback rate constraint," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3879–3891, 2007.