Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 854120, 14 pages
doi:10.1155/2008/854120

Research Article

Scheduling for Multiuser MIMO Downlink Channels with Ranking-Based Feedback

Marios Kountouris,[1] Thomas Sälzer,[2] and David Gesbert[3]

[1] *Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA*
[2] *France Telecom Research and Development, Issy-les-Moulineaux, 92794, France*
[3] *Mobile Communications Department, Eurecom Institute, Sophia-Antipolis, F-06904, France*

Correspondence should be addressed to Marios Kountouris, mkountouris@ece.utexas.edu

Received 28 June 2007; Revised 10 October 2007; Accepted 26 January 2008

Recommended by Christoph Mecklenbräuker

We consider a multi-antenna broadcast channel with more single-antenna receivers than transmit antennas and partial channel state information at the transmitter (CSIT). We propose a novel type of CSIT representation for the purpose of user selection, coined as ranking-based feedback. Each user calculates and feeds back the rank, an integer between 1 and W + 1, of its instantaneous channel quality information (CQI) among a set of W past CQI measurements. Apart from reducing significantly the required feedback load, ranking-based feedback enables the transmitter to select users that are on the highest peak (quantile) with respect to their own channel distribution, independently of the distribution of other users. It can also be shown that this feedback metric can restore temporal fairness in heterogeneous networks, in which users' channels are not identically distributed and mobile terminals experience different average signal-to-noise ratio (SNR). The performance of a system that performs user selection using ranking-based CSIT in the context of random opportunistic beamforming is analyzed, and we provide design guidelines on the number of required past CSIT samples and the impact of finite W on average throughput. Simulation results show that feedback reduction of order of 40−50% can be achieved with negligible decrease in system throughput.

1. INTRODUCTION

Multiple-input multiple-output (MIMO) communication systems have the potential to offer high spectral efficiency as well as link reliability. In multiuser MIMO downlink networks, the spatial degrees of freedom offered by multiple antennas can be advantageously exploited to enhance the system capacity by scheduling multiple users by means of space division multiple access (SDMA) to simultaneously share the spatial channel [1, 2]. As the capacity-achieving dirty paper coding (DPC) approach [3] is rather complex to be implemented, several practical downlink transmission techniques have been lately proposed. Downlink linear precoding, although suboptimal, has been shown to achieve a large fraction of DPC capacity, exhibiting reduced complexity [4–6]

The advantages promised by multiuser MIMO systems unfortunately come at the cost of perfect channel state information at transmitter (CSIT) in order to properly serve the spatially multiplexed users. Recent information-theoretic results reveal the cardinal importance of CSIT in multiuser MIMO precoding. If a base station (BS) with M transmit antennas communicating with K single-antenna receivers has perfect CSI, a multiplexing gain of $\min(M, K)$ can be achieved. Nevertheless, if the transmitter has imperfect channel knowledge, the full multiplexing gain is severely reduced at high signal-to-noise ratio (SNR) [7], whereas if there is complete lack of CSI knowledge, the multiplexing gain collapses to one [8]. The approximation of close-to-perfect CSI at the receiver (CSIR) is often reasonable, especially for downlink channels, where a common pilot can be employed for channel estimation by a large number of users. However, this assumption is often unrealistic and impractical at the transmitter side. In a time-division duplex (TDD) system, close-to-perfect CSIT can be obtained by exploiting the channel reciprocity. In the context of frequency-division

duplex (FDD) systems, CSIR is obtained through training, whereas obtaining CSIT generally requires feedback reporting from each mobile station (MS).

Providing CSIT at the BS poses serious challenges in practical settings where the channel information needs be conveyed via a limited feedback channel in the uplink. The requirement of CSIT feedback in multiuser MIMO configurations places a significant burden on uplink capacity in most systems, exacerbated in systems with wideband (e.g., OFDM) communication or high mobility (such as 3GPP-LTE, WiMax). The often unrealistic assumption of close-to-perfect CSIT, as well as the considerable capacity gap between full and no CSIT, have motivated research work on feedback reduction schemes. Inspection of recent literature reveals several different schools of thought on limited feedback, including vector quantization, dimension reduction, adaptive feedback, contention-based feedback, statistical feedback, and opportunistic SDMA. A tutorial on multiuser MIMO with limited feedback can be found in [9]. One line of work, often referred to as limited feedback approach, attempts to reduce the amount of feedback per user by means of quantization of CSI parameters. Limited feedback approaches, imposing a bandwidth constraint on the feedback channel, have been proposed for MIMO point-to-point systems [10–13], where each user feeds back finite-precision CSI on its channel direction by quantizing its normalized channel vector to the closest vector contained in a predetermined codebook. An extension of the limited feedback model for multiple antenna broadcast channels for the case of $K = M$ is made in [14, 15]. In [14], it is shown that the feedback load per mobile must increase approximately linearly with the number of transmit antennas and the average transmit power (in dB) in order to achieve the full multiplexing gain, and consequently performance close to that with full CSIT [14]. For instance, in a 6-transmit antenna system operating at 10 dB, each user has to report 17 bits. A feedback reduction technique for MIMO broadcast channels exploiting multiple antennas at the receiver side as a means to improve the quality of channel estimate conveyed back to the BS is proposed in [16].

A popular, very low-rate feedback technique, coined as opportunistic random beamforming, was initially proposed for single-beam setting [17] and later generalized for an SDMA setting in [18]. In this scheme, once M orthonormal beams are generated randomly, each user calculates its signal-to-interference plus noise ratio (SINR) for each of the M beams and feeds back its best SINR value along with the corresponding beam index. The best user on each beam is then scheduled. By means of multiuser diversity [19], this scheme is shown to yield the optimal capacity growth of $M \log\log K$ for large number of users. However, the sum rate performance of this scheme is quickly degrading with decreasing number of users.

An alternative approach, referred to as selective or threshold-based feedback, allows a user to send back information depending on whether its current channel conditions exceed a certain threshold or not. This feedback reduction algorithm was first proposed in [20] for a downlink single-input, single-output (SISO) system and SNR-dependent thresholds. This method is shown to reduce statistically the required total amount of feedback by means of multiuser diversity. The feedback rate can be further reduced, at the cost of feedback delay by using an adaptive threshold [21]. The selective feedback idea was extended for MISO systems in [22]. In [23], a scheme based on [17] and one bit feedback was shown to achieve the optimal capacity growth rate when $K \to \infty$. A scheme based on multibeam random beamforming was proposed in [24, 25] where it was proved that a deterministic feedback of $\log_2(1 + M)$ bits per user is enough to guarantee the optimal scaling law for single-antenna receivers and fixed M.

A common limitation of the above feedback reduction techniques is that the total feedback rate grows linearly with the number of users, thus reducing the effective system throughput when the number of users is large. SDMA under a sum feedback rate constraint is considered in [26], in which threshold-based feedback on the channel quality and the channel direction is used for feedback reduction in order to satisfy a sum feedback rate constraint. Differently from the previous approaches in which users are assumed to send feedback through dedicated channels, the authors in [27] consider a contention-based feedback protocol, in which users compete to gain access in a shared medium. In this system, the feedback resources are fixed random access minislots, and active users attempt to convey feedback messages only if their channel gain is above a threshold.

In this paper, we take on a completely different approach for feedback reduction compared to the existing ones. Our work is building upon recently proposed ideas in the context of scheduling [28]. In [28], a so-called "score-based" opportunistic scheduler was proposed for realistic scenarios with asymmetric fading statistics and data rate constraints. Similar distribution-based schedulers have also been proposed in [29–31] as a means to schedule a user whose instantaneous rate is in the highest quantile of its distribution. Interestingly, these works were solely focused on scheduling at the transmitter side, and not in the context of feedback reduction nor that of MIMO systems. We consider the problem of feedback reduction in a downlink multiple antenna communication system, in which a BS equipped with M antennas communicates with $K \geq M$ single-antenna users. It is assumed that the receivers have perfect channel state information (CSI) while the BS relies only on partial CSI, conveyed through a feedback channel. In the lines of [32, 33], we adopt a two-stage approach by splitting the feedback between a first stage of scheduling (or "user selection") and a second transmission/precoding design (or "user serving") stage. During the scheduling phase, all active users are allowed to feedback some kind of finite-rate channel quality information (CQI), whereas in the second step, information on the transmission rate is requested only from the $M \ll K$ selected users. We focus on the first phase and we propose a new CQI representation metric as a means to reduce significantly the burden on uplink feedback channel rate.

The contributions of this paper are as follows.

(i) We propose a new concept of CSIT representation, coined as "ranking-based feedback," for the sole purpose of user selection as a means to reduce the required feedback load. The ranking-based CSIT consists of an integer value that represents the rank of each user's instantaneous CQI among a number of stored CQI values observed over the W past slots.

(ii) The key advantage of the proposed method is two fold: (1) the ranking-based feedback is already in digital form which helps for further compression and simple scalar quantization. (2) the ranking-based feedback provides not only information about the channel quality at any instant but also about the relative quality level, in a way that is independent of the users' fading statistics, thus providing inherent fairness. This type of limited feedback enables the base station to select users that are on the peak of their own channel distribution, independently of the channel conditions of other users.

(iii) We analyze the sum-rate performance of a multi-antenna downlink system with multiple orthogonal beams as in [18], in which users are selected during the scheduling phase based on ranking-based CSIT. Furthermore, we provide analytic expressions for the sum rate when W is finite.

(iv) We compare the performance against standard random beamforming schemes using SINR feedback metric for user selection, and we quantify the effect of finite W and the error introduced in the scheduling decisions compared to the optimal case of $W \to \infty$.

(v) We present an additional merit of ranking-based CSIT in a heterogeneous network by showing that such form of feedback can provide temporal fairness among users, as the probability of a user to be selected is $1/K$, independently of the other users' channel distributions and its own average SNR (pathloss).

The remainder of this paper is organized as follows. The system model is described in Section 2, and in Section 3 the proposed ranking-based feedback framework is presented. The system rate of a system employing ranking-based feedback metric for user selection is analyzed in Section 4. Extensions to codebook-based SDMA schemes are provided in Section 5, and the proposed feedback concept is applied to a heterogeneous network in Section 6. The performance of the proposed feedback reduction technique is numerically evaluated in Section 7, and, finally, Section 8 concludes the paper.

2. SYSTEM MODEL

We consider a multiple antenna downlink channel in which a base station (transmitter) equipped with M antennas communicates with K single-antenna users (receivers). The received signal $y_k(t)$ of the kth user at time slot t is mathematically described as

$$y_k(t) = \mathbf{h}_k^H(t)\mathbf{x}(t) + n_k(t), \quad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{C}^{M\times 1}$ is the vector of transmitted symbols at time slot t, $\mathbf{h}_k(t) \in \mathbb{C}^{M\times 1}$ is the channel vector from the transmitter to the kth receiver, and $n_k(t)$ is additive white Gaussian noise at receiver k. We assume that each of the receivers has perfect and instantaneous knowledge of its own channel $\mathbf{h}_k$, and that n_k is independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian with zero mean and unit variance. The covariance matrix of the transmitted signal is $\mathbf{\Sigma}_x = \mathbb{E}(\mathbf{x}\mathbf{x}^H)$. The transmitter is subject to a total power constraint P, which implies $\mathrm{Tr}(\mathbf{\Sigma}_x) \leq P$, where $\mathrm{Tr}(\cdot)$ is the trace operator. We consider an i.i.d. block Rayleigh flat fading model, where the channel is invariant during each coded block, but is allowed to vary independently from block to block. We also assume that the number of mobiles is greater than or equal to the number of transmit antennas, that is, $K \geq M$, and that the BS selects for transmission $\mathcal{M}$ out of K users, with $1 \leq \mathcal{M} \leq M$.

Notation 1. We use bold upper and lower case letters for matrices and column vectors, respectively. $(\cdot)^*$, $(\cdot)^T$, and $(\cdot)^H$ stand for conjugate, transpose, and Hermitian transpose, respectively. $\mathbb{E}(\cdot)$ denotes the expectation operator. The ℓ^2-norm of the vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|$, and $\angle(\mathbf{x}, \mathbf{y})$ represents the angle between vectors $\mathbf{x}$ and $\mathbf{y}$. The $\log(\cdot)$ refers to the natural logarithm while the base 2 logarithm is denoted $\log_2(\cdot)$.

3. RANKING-BASED FEEDBACK FRAMEWORK

In this section, we present the concept of *ranking-based feedback* and its intrinsic advantages when it is used as a user selection metric during the scheduling stage. For simplicity of exposition, we study its use in the particular context of random beamforming (RBF), however, as it is shown later, the ideas could be generalized to various downlink precoding scenarios.

3.1. Random beamforming system model

In the random opportunistic beamforming scheme, only $\mathcal{M}$, $1 \leq \mathcal{M} \leq M$, spatially separated users access the channel simultaneously. The transmitter generates $\mathcal{M}$ mutually orthogonal random beams, as proposed in [17] for $\mathcal{M} = 1$ and in [18] for the multibeam case of $\mathcal{M} = M$. The transmitted signal is given by

$$\mathbf{x}(t) = \sum_{m=1}^{\mathcal{M}} \mathbf{q}_m(t) s_m(t), \tag{2}$$

where $s_m(t)$ is the transmit symbol associated to the mth beam, and $\mathbf{q}_m \in \mathbb{C}^{M\times 1}$ is the beamforming vector for the mth beam in slot t. The random orthonormal vectors are

generated as isotropically distributed. The SINR of the kth user on beam m is given by

$$\mathrm{SINR}_{k,m} = \frac{\left|\mathbf{h}_k^H \mathbf{q}_m\right|^2}{\sum_{j\neq m}\left|\mathbf{h}_k^H \mathbf{q}_j\right|^2 + \mathcal{M}/P}, \quad m = 1,\ldots,\mathcal{M}. \tag{3}$$

For $\mathcal{M} = 1$, (3) represents the received SNR given by $\mathrm{SNR}_k = P|\mathbf{h}_k^H \mathbf{q}_m|^2$.

3.2. Ranking-based scheduling

Let $\mathcal{G}$ be the set of all possible subsets of disjoint indices among the complete set of user indices $\{1,\ldots,K\}$ and let $\mathcal{S} \in \mathcal{G}$ be one such group of $|\mathcal{S}| = \mathcal{M} \leq M$ users selected for transmission at a given time slot. In the proposed CSIT framework, we assume a two-step feedback approach by splitting the feedback resource into two stages (scheduling followed by transmission). In the scheduling stage, all K active users compete for medium access and each user k is allowed to report instantaneous CQI, denoted as γ_k, which is a certain function of the channel, that is, $\gamma_k = f(\mathbf{h}_k)$. This CQI metric can generally take on any form of channel information representation. For instance, in a time-division multiple access (TDMA) context, γ_k may represent the SNR or the transmission rate of user k, whereas in an SDMA variant, the CQI can be the received SINR (achievable or estimated). This channel quality metric is used *solely* for purposes of user selection during the scheduling stage. Given a set of $\mathcal{M}$ preselected users, a second-step exploiting precoding is applied to serve the selected users. The second-step precoding matrix may require variable levels of additional CSIT feedback to be computed, depending on design. Here, we assume that the second-step beamformer is the same as the one used in the scheduling step and the selected users feed back their transmission rate. Alternatively, the need for a second stage in order to inform the BS on the transmission rate can potentially be circumvented by assuming that the cumulative distribution functions (CDFs) of different users are known a priori at the transmitter. This assumption can be justified in systems where the statistical reciprocity between the downlink and uplink channels allows the BS to estimate the distributions by aggregating each user's CQI feedback.

At time instant t, each user measures its CQI on each of $\mathcal{M}$ randomly generated beams (columns of the first-stage precoding matrix). In addition to the instantaneous CQI value on each beam m, $\{\gamma_{k,m}(t)\}_{m=1}^{\mathcal{M}}$, each user also keeps record of a set of past CQI values, denoted as $\mathcal{W}_{k,m}$, observed over a window of size W, that is, $\mathcal{W}_{k,m} = \{\gamma_{k,m}(t-1), \gamma_{k,m}(t-2),\ldots,\gamma_{k,m}(t-W+1)\}$. Then, each user, say the kth, calculates the ranking (order) $r_{k,m}(t) \in \{1,\ldots,W+1\}$ of its current CQI metric $\gamma_{k,m}(t)$ on beam m among the W past values contained in the set $\mathcal{W}_{k,m}$. In other words, if $\gamma_{k,m}(t)$ is the third largest value within the set of W latest measured values, $r_{k,m}(t) = 3$. The rank value of user k at slot t on beam m is mathematically given by [28]

$$\begin{aligned} r_{k,m}(t) = 1 &+ \sum_{w=1}^{W-1} 1\{\gamma_{k,m}(t) < \gamma_{k,m}(t-w)\} \\ &+ \sum_{w=1}^{W-1} 1\{\gamma_{k,m}(t) = \gamma_{k,m}(t-w)\} Z_w, \end{aligned} \tag{4}$$

where Z_w are i.i.d. random variables on $\{0,1\}$ with $\Pr\{Z_w = 0\} = 1/2$ corresponding to the case where the instantaneous CQI is equal to one or several of the past values, in which either rank value is randomly chosen with equal probability.

The key ideas are as follows:

(1) each user selects its minimum rank value over the beams, that is,

$$r_k(t) = \min_{m=1,\ldots,\mathcal{M}} r_{k,m}(t); \tag{5}$$

(2) each user, instead of reporting directly its maximum CQI value over the beams, feeds back a quantized value $\hat{r}_k(t)$ of the integer $r_k(t)$, along with the beam index m in which the ranking value is minimum, that is,

$$\hat{r}_k(t) = \mathcal{Q}(r_k(t)), \tag{6}$$

where $\mathcal{Q}(\cdot)$ represents an $N = 2^B$-level quantizer. Thus, the feedback load per user is $\lceil \log_2 N \rceil$ bits for the ranking and $\lceil \log_2 M \rceil$ bits for the index of its preferred beam.

At the transmitter side, the scheduler assigns each beam m to the user k_m^* with the minimum reported ranking value, that is,

$$k_m^*(t) = \arg\min_{1\leq k\leq K} \hat{r}_k(t). \tag{7}$$

As stated before, once the users $\{k_m^*(t)\}_{m=1}^{\mathcal{M}}$ are selected based on ranking-based CSIT, they are polled and requested to report the transmission rate that can be supported by their instantaneous channel conditions.

The length of the observation window provides a measure of how accurately the channel distribution is monitored by the user. The larger the W, the better a user can track the distribution of its CQI process, thus identifying more accurately the peaks with respect to its own distribution. In other words, ranking-based CSIT enables each user to have an estimate of the quantile of its CQI using W previous CQI samples, where the sample quantile of order p is defined as the statistical functional $\hat{F}_W^{-1}(p) = \inf\{x : \hat{F}_W(x) \geq p\}$ for $p \in (0,1)$ and $\hat{F}_W(\cdot)$ denoting the empirical distribution function of W samples. More formally, for a process $(Y(t), t \geq 0)$ with stationary and independent increments with $Y(0) = 0$, the p-quantile of $(Y(s), 0 \leq s \leq t)$ for $0 < p < 1$ is defined by $M(p,t) = \inf\{x : \int_0^t \mathbf{1}(Y(s) \leq s) ds > pt\}$. In the asymptotic case of $W \to \infty$, the observation window captures the entire distribution and corresponds to the case in which ranking-based CSIT gives exact information on the CDF of the CQI process. In this

case, the user with the minimum ranking-based CQI value is the one whose instantaneous CQI is in the highest quantile.

4. PERFORMANCE EVALUATION

In this section, we evaluate the average rate of a system employing random opportunistic beamforming in which ranking-based feedback is used as user selection metric. We assume that the CQI takes on the form of user rate, that is, $\gamma_{k,m} = \log_2(1 + \mathrm{SINR}_{k,m})$. Let $X_{k,m}$ denote the rate process of the kth user rate on the mth beam with CDF denoted as $F_{X_{k,m}}(\cdot)$. The distribution function is assumed to be strictly increasing and continuous, such that its inverse $F^{-1}_{X_{k,m}}(\cdot)$ exists. In the following sections, unless otherwise stated, we assume a homogenous network where all users have equal average SNR (i.i.d. channel statistics). The case of independent but not identically distributed (non-i.i.d.) channel statistics is studied in Section 6.

4.1. Asymptotic optimality of ranking-based feedback for large window size W

For finite window size W, ranking-based CSIT enables each user to estimate the quantile of its instantaneous CQI based on W samples of its empirical CQI process. For fixed x, the number of random variables (r.v.) X_i such that $X_i \leq x$ follows a binomial distribution with probability of "success" $p = F(x)$, thus the random variable $\hat{F}^W_X(x)$ follows a binomial distribution with possible values $0, 1/W, \ldots, 1$. In this section, we examine the behavior of the empirical function $\hat{F}^W_X(x)$ for W increasing and show how likely is $\hat{F}^W_X(x)$ to be close to $F(x)$ for arbitrary large W and x fixed.

Let the collection of r.v. $\mathcal{X} = \{X_t : t \in \mathbb{N}^+\}$ be a discrete-time stochastic process for each user defined on the same probability space. $\mathcal{X}$ is assumed stationary and ergodic and for exposition convenience we omitted the user index k from the stochastic process. The random sample of i.i.d. r.v. $X_1, X_2, \ldots, X_W$ is an empirical process, whose empirical distribution $\hat{F}^W_X(\cdot)$ is defined as the CDF that puts mass $1/W$ at each sample point X_i, that is,

$$\hat{F}^W_X(x) = \frac{1}{W}\sum_{i=1}^{W} \mathbb{I}\{X_i \leq x\}, \tag{8}$$

where $\mathbb{I}\{X_i \leq x\}$ is an indicator function defined as

$$\mathbb{I}\{X_i \leq x\} = \begin{cases} 1, & X_i \leq x, \\ 0, & X_i > x. \end{cases} \tag{9}$$

Proposition 1. *In a system where users have i.i.d. channel statistics, user selection based on ranking-based feedback converges to the capacity-optimal max-rate scheduling for $W \to \infty$.*

Proof. See Appendix A. □

4.2. Average sum rate for infinite observation window size W

In this section, we study the average sum rate of a system using ranking-based feedback as a user selection metric in the large W regime. Assuming W to be infinitely large, we can easily see that user selection based on ranking-based CSIT is equivalent to minimum complementary (CCDF) scheduling. This means that if $r_{k,m}$ captures the distribution of received SINR process $\Gamma_{k,m}$, then $\lim_{W\to\infty}(r_{k,m}/W) = \overline{F}_{\Gamma_{k,m}}(\gamma_{k,m})$, where $\overline{F}_{\Gamma_{k,m}}(\gamma_{k,m}) = 1 - F_{\Gamma_{k,m}}(\gamma_{k,m})$ is the complementary CDF of CQI metric $\gamma_{k,m}$. Hence, as shown in Proposition 1, selecting on each beam m the user k^*_m with the minimum ranking value is equivalent to selecting the user with the minimum tail of CDF, that is

$$\begin{aligned} k^*_m &= \arg\min_{1\leq k\leq K} \overline{r}_{k,m}(t) \\ &= \arg\min_{1\leq k\leq K} 1 - F_{\Gamma_{k,m}}(\gamma_{k,m}(t)) \\ &= \arg\max_{1\leq k\leq K} F_{\Gamma_{k,m}}(\gamma_{k,m}(t)) \quad m = 1, \ldots, \mathcal{M}, \end{aligned} \tag{10}$$

where $\overline{r}_{k,m}(t)$ is the normalized ranking value and $\gamma_{k,m}(t)$ is the realization of $\Gamma_{k,m}$ at slot t.

The rate of user k on beam m, prior to channel-aware scheduling, is given by

$$\begin{aligned} \mathcal{R}_{k,m} &= \int_0^\infty \log_2(1+\gamma) f_{\Gamma_{k,m}}(\gamma) d\gamma \\ &= \int_0^1 \log_2\left(1 + F^{-1}_{\Gamma_{k,m}}(\overline{r})\right) d\overline{r}, \end{aligned} \tag{11}$$

where $f_{\Gamma_{k,m}}(\cdot)$ is the probability density function (pdf) of CQI metric γ.

Consider a homogeneous system (i.i.d. channel distributions) and that the user on the highest quantile is scheduled on each beam m, then the average sum rate is given by the following proposition.

Proposition 2. *The average sum rate, $\mathcal{R}$, of a homogeneous system in which user selection is performed based on ranking-based feedback is given by*

$$\mathcal{R} = \mathcal{M}K \int_0^1 \log_2\left(1 + F^{-1}_\Gamma(z)\right) z^{K-1} dz. \tag{12}$$

Proof. The proof is straightforward by changing the variable $F_\Gamma(\gamma) = z$ in the sum rate given by $\mathcal{R} = \mathcal{M}\int_0^\infty \log_2(1+\gamma) dF^K_\Gamma$, where F^K_Γ is the CDF of the best user selected among K i.i.d. users with common parent distribution $F_\Gamma(\gamma)$.

Note that similar result has been derived in [29]. Therein, the authors derive the average user rate for the general case where the channel distributions are not necessarily identically distributed and $\mathcal{M} = 1$. Proving that the probability that user k is selected at time slot t given that the user rate $X_k(t) = x_k$ is $\Pr\{k^*(t) = k \mid X_k(t) = x_k\} = F^{K-1}_{X_k}(x_k)$, they showed that the average rate of a user is given by $R_k = \int_0^1 u^{K-1} F^{-1}_{X_k}(u) du$.

Equation (12) does not always result in closed-form expressions. For instance, the sum rate of multibeam RBF

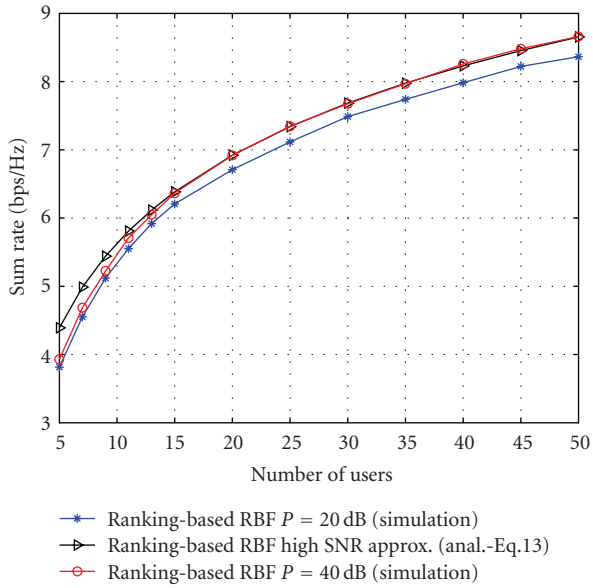

Figure 1: Sum rate performance comparison of analytic high SNR approximate solution (13) for RBF ($M = 4$ beams) using ranking-based CSIT as user selection metric to simulated results using Monte Carlo.

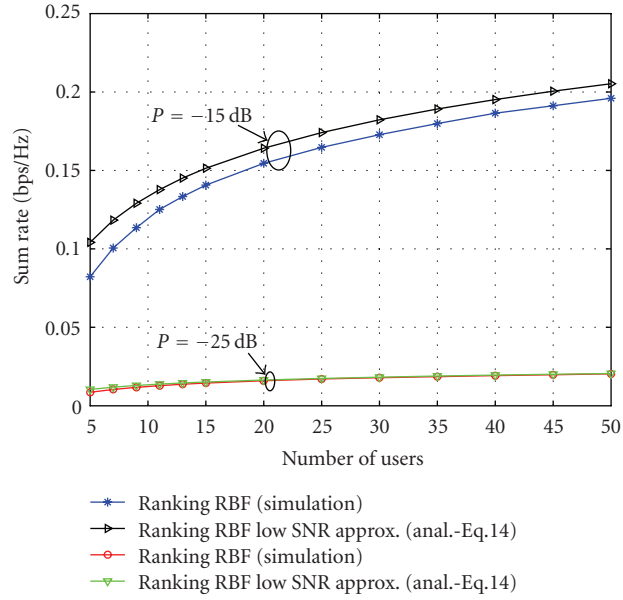

Figure 2: Sum rate performance comparison of analytic low SNR approximate solution (14) for RBF ($M = 4$ beams) using ranking-based CSIT as user selection metric to simulated results using Monte Carlo.

given by $\mathcal{R}_{\text{RBF}} = \mathcal{M}K\int_0^1 F_{X_k}^{-1}(u)u^{K-1}du$, where $F_{X_k}^{-1}(u)$ is the inverse of $F_{X_k}(u) = 1 - e^{-M/P}e^{-2^u M/P}/2^{(M-1)u}$, requires numerical calculation. Nevertheless, analytic sum-rate expressions can be derived in specific regimes, such as the high- and low-power regions. □

Corollary 1. *At high SNR ($P \to \infty$), the average sum rate of multibeam random beamforming with M beams and ranking-based user selection is given by*

$$\mathcal{R}_{\text{high}} = \frac{M}{M-1}\frac{H_K}{\log(2)}, \tag{13}$$

where $H_K = \sum_{k=1}^{K}(1/k)$ is the kth harmonic number.

Proof. When $P \to \infty$, the CDF of SINR can be approximated by $F_\Gamma(\gamma) = 1 - 1/(1+\gamma)^{M-1}$ Thus, by Proposition 2, the sum-rate is given by $\mathcal{R} = -(M/(M-1))\,K\int_0^1 \log_2(1-z)z^{K-1}dz = (M/(M-1))(H_K/\log(2))$. □

Corollary 2. *At low SNR ($P \to 0$), the average sum rate of multibeam random beamforming with M beams and ranking-based user selection is given by*

$$\mathcal{R}_{\text{low}} = \log_2(e)PH_K, \tag{14}$$

where H_K is the kth harmonic number.

Proof. When $P \to 0$, the CDF of SINR can be approximated by $F_\Gamma(\gamma) = 1 - e^{-M\gamma/P}$. Using the first-order Taylor series expansion of the logarithm, that is, $\log(1+x) \approx x$ for small x, we have $\mathcal{R} = -(PK/\log(2))\int_0^1 \log(1-z)z^{K-1}dz = \log_2(e)PH_K$.

The analytic expressions of the above two corollaries ((13) and (14)) are compared to simulated results in Figures 1 and 2. □

4.3. Average sum rate for finite observation window size W

Let $X_{k_m^*}(t)$ denote the rate process of the user k selected on beam m with with distribution function $F_{X_{k_m^*}}(x)$. The expected rate $\mathcal{R}_{k,m}$ of kth user when scheduled on beam m is given by

$$\begin{aligned}\mathcal{R}_{k,m} &= \mathbb{E}\{X_{k_m^*}(t)\} \\ &= \int_0^\infty \Pr\Big\{\max_{1\le k\le K} X_{k,m}(t) > x\Big\}dx.\end{aligned} \tag{15}$$

Proposition 3. *The average sum rate $\mathcal{R}$ of a system generating $\mathcal{M}$ random orthonormal beams and scheduling $\mathcal{M}$ users among K active users based on ranking-based feedback with observation window W is given by*

$$\begin{aligned}\mathcal{R} = \sum_{m=1}^{\mathcal{M}}\Bigg(&\int_0^\infty \left(1 - \left(F_{X_{k_m^*}}(x)\right)^W\right)dx \\ &- \sum_{w=1}^{W}\left(\frac{W-w}{W}\right)^K \int_0^\infty F_{w,m}(x)dx\Bigg),\end{aligned} \tag{16}$$

where $F_{w,m}(x) = \binom{W}{w}\left(F_{X_{k_m^*}}(x)\right)^{W-w}\left(1 - F_{X_{k_m^*}}(x)\right)^w$, and $F_{X_{k_m^*}}(x) = [\Pr\{X_{k,m} \le x\}]^K$.

Proof. See Appendix B. □

For instance, based on the above proposition the throughput $\mathcal{R}_{\text{TDMA}}$ of single-beam RBF is given by

$$\mathcal{R}_{\text{TDMA}} = \sum_{w=0}^{W}\left[1-\left(\frac{W-w}{W}\right)^{K}\right]\binom{W}{w} \times \int_0^{\infty}\left(F_{X_{k_m^*}}(x)\right)^{W-w}\left(1-F_{X_{k_m^*}}(x)\right)^{w}dx, \tag{17}$$

as $F_{X_{k_m^*}}(x) = (1-e^{-(2^x-1)/P})^K$. Unfortunately, (17) does not seem to have closed-form representation for exponentially distributed channel gains. However, in the high power regime, the following series representation can be obtained.

Corollary 3. *At high SNR ($P \to \infty$), the average sum rate R_{high}^W of multibeam random beamforming with $M = 2$ beams, finite W, and ranking-based user selection, are given by*

$$\mathcal{R}_{high}^W = 2\sum_{w=1}^{W}\binom{W}{w}\left[1-\left(\frac{W-w}{W}\right)^{K}\right] \times \frac{\Gamma(Kw-1)\Gamma(KW-Kw+1)}{\Gamma(KW)}. \tag{18}$$

For large enough W, a good approximation of the binomial distribution is given by the normal distribution (De Moivre-Laplace theorem). Let $q = F_{X_{k_m^*}}(x)$ and $p = 1 - F_{X_{k_m^*}}(x)$, then $F_{w,m}(x)$ can be approximated by

$$F_{w,m}(x) \approx \frac{1}{\sqrt{2\pi Wpq}}e^{-(w-Wp)^2/2Wpq}, \tag{19}$$

which simplifies the calculation of integral in (16) as $\int_0^{\infty}F_{w,m}(x)dx = Q(\sqrt{2Wp/q})$, where $Q(\cdot)$ is the standard normal CDF.

4.4. Performance reduction bound for finite window size W

In this section, we provide a bound on the ratio of the empirical distribution observed over W samples by the actual CDF ($W\to\infty$) as a means to quantify the throughput reduction using ranking-based CSIT calculated over finite W. Intuitively, the rate performance is a monotonically decreasing function with W, thus for W decreasing, the performance degradation is increased.

A bound on the difference between the rate when each user knows perfectly its CDF and the throughput when ranking-based feedback is based on the empirical distribution of each user's channel distribution over W samples does not seem tractable. The main difficulty is that the user rate distribution, as $F_{X_{k,m}}(x)$, is not a linear function of the CQI distribution, that is, $F_{X_{k,m}}(x) = F_{\Gamma_{k,m}}(2^x - 1)$. Nevertheless, a bound on the the ratio $\mathcal{F}(W,K) = \widehat{F}_{X_{k_m^*}}^W(x)/F_{X_{k_m^*}}(x)$, where $\widehat{F}_{X_{k_m^*}}^W(\cdot)$ is rate distribution seen by user k when is scheduled based on ranking-based feedback estimated using W samples is derived in [31].

Proposition 4. *For a system with K active users employing ranking-based CSIT observed over W past values, the ratio $\mathcal{F}(W,K)$ is lower bounded as*

$$\mathcal{F}(W,K) \geq \left(1-\left(\frac{W}{W+1}\right)^{K}\right)\frac{W+1}{K} \leq \left(1-e^{-K/W}\right)\frac{W+1}{K}, \tag{20}$$

where the Bernoulli inequality is used for bounding $(W/(W+1))^K$.

Expanding $e^{-K/W}$ in Taylor series, we have that $(1-e^{-K/W})((W+1)/K) = (1-(K/2W))((W+1)/W) \gtrsim 1 - K/(W+2)$. Hence, for fixed throughput reduction, the number of samples W required to be stored in memory has to scale almost linearly with the number of active users K in the system.

In addition to the previous bound, a sharp nonasymptotic bound can be derived based on the Dvoretsky-Kiefer-Wolfowitz (DKW) inequality [34, 35].

Theorem 1. *Let $X_1, X_2, \ldots, X_W \sim F_{X_{k,m}}$, then for any $\epsilon > 0$*

$$\Pr\left\{\sup_x |\widehat{F}_{X_{k,m}}^W(x) - F_{X_{k,m}}(x)| > \epsilon\right\} \leq 2e^{-2W\epsilon^2}. \tag{21}$$

Based on Theorem 1, we can construct a confidence set that gives us a measure of the required window size W. Given $\alpha \in (0,1)$, say that a random set $S(x)$ is a $(1-\alpha)$ confidence set for the parameter θ if

$$\Pr\{\theta \in S(x)\} \geq 1-\alpha. \tag{22}$$

Define two sequences $\ell_1(x) = \max\{\widehat{F}_{X_{k,m}}^W(x) - \epsilon_W, 0\}$ and $\ell_2(x) = \min\{\widehat{F}_{X_{k,m}}^W(x) + \epsilon_W, 1\}$ with $\epsilon_W = \sqrt{(1/2W)\log(2/\alpha)}$. Then, for any F, we have that

$$\Pr\{\ell_1(x) \leq F_{X_{k,m}}(x) \leq \ell_2(x),\ \forall x\} \geq 1-\alpha. \tag{23}$$

This implies that if one wishes to draw a large enough sample to ensure that the deviation between the empirical distribution and the actual CDF is less than or equal to 10%, with 90% confidence, then for $\epsilon = 0.1$ in (21), a sample size of approximately $W = 150$ samples is needed.

4.5. Window size versus feedback reduction tradeoff

In the previous section, it has been shown that the performance difference between ranking-based user selection and max-rate scheduling is decreased for W increasing. In practical systems, the feedback channel shared by all users has a fixed bandwidth and thus the rate of reporting $\hat{r}_k(t)$ is finite and generally fixed. As a result, under a fixed feedback rate constraint of $B = \lceil \log_2 N \rceil$ bits, when W is increased, the accuracy of $\hat{r}_k(t)$ is decreased as the distortion of the quantizer $\mathcal{Q}(\cdot)$ is increased. This is evidently due to the fact that the dynamic range of the integer values $r_k(t) \in (0, W+1]$ to be quantized by B bits is increased. In order to guarantee

the same throughput performance for increasing W, the number of feedback bits B should scale accordingly so that the quantization error is fixed. This results in an interesting tradeoff between the following:

(i) the capacity performance,

(ii) the window size W,

(iii) the number of feedback bits B.

Consider that uniform scalar quantization is used to quantize a source R that is uniformly distributed over $[0, 1]$. The error variance (distortion) is given by

$$\sigma_Q^2 = \mathbb{E}\{(R - \mathcal{Q}(R))^2\} = \int_{-\infty}^{+\infty} (r - \mathcal{Q}(r))^2 f_R(r) dr = \frac{(r_{\max} - r_{\min})^2}{12N^2}, \tag{24}$$

where $f_R(r)$ is the PDF of the uniform source R, and $r_{\max}$ and $r_{\min}$ are the maximum and minimum value of ranking-based feedback, respectively. For fixed variance of the quantization error $\sigma_Q^2 = \delta^2$, $r_{\min} = 1$ and $r_{\max} = W + 1$, the number of bits B should scale proportionally to $B \sim (\log_2(W/\delta) - 1.8)$ bits. This feedback requirement can be decreased if nonuniform quantization (e.g., optimal entropy-constrained [36]) is employed. The problem of optimum quantization design for ranking-based feedback is beyond the scope of this paper.

5. EXTENSIONS TO CODEBOOK-BASED SDMA SCHEMES

The concept of ranking-based feedback, as presented above, is not restrictive to the random beamforming; it can be generalized to other downlink precoding configurations. The ranking-based concept can indeed be applied to any kind of feedback information of interest. In a MIMO broadcast channel, for instance, it can be additionally used to represent some kind of channel direction information (CDI) as a means to select near orthogonal user with large channel gains. Consider a system in which each user can report CDI feedback based on a predefined codebook in addition to the CQI value that can take on the form of channel norm or estimate of SINR [37, 38]. Consider a quantization codebook $\mathcal{V}_k = \{\mathbf{v}_{k1}, \mathbf{v}_{k2}, \ldots, \mathbf{v}_{kL}\}$ containing L unit norm vectors $\mathbf{v}_{ki} \in \mathbb{C}^M$, for $i = 1, \ldots, L$, which is assumed to be known to both the kth receiver and the transmitter. In each scheduling interval, each receiver k quantizes its channel to the codevector that maximizes the following inner product:

$$\widehat{\mathbf{h}}_k = \mathbf{v}_{kl} = \arg\max_{\mathbf{v}_{ki} \in \mathcal{V}_k} \left| \overline{\mathbf{h}}_k^H \mathbf{v}_{ki} \right|^2 = \arg\max_{\mathbf{v}_{ki} \in \mathcal{V}_k} \cos^2(\angle(\overline{\mathbf{h}}_k, \mathbf{v}_{ki})) \tag{25}$$

where the normalized channel vector $\overline{\mathbf{h}_k} = \mathbf{h}_k / \|\mathbf{h}_k\|$ corresponds to the channel direction, and we refer to $\widehat{\mathbf{h}}_k$ as the kth user channel quantization.

Denote $r_{g,k}$ the kth user ranking of its CQI among W past values, where CQI is given by the channel norm $\|\mathbf{h}_k\|$. Let $r_{d,k}$ be the ranking-based CDI given by alignment between the directions of the actual channel and the quantized one, that is, $\cos^2(\angle(\overline{\mathbf{h}}_k, \widehat{\mathbf{h}}_k)) = |\overline{\mathbf{h}}_k^H \widehat{\mathbf{h}}_k|^2$. The users report back to the transmitter both $r_{g,k}$ and $r_{d,k}$ and the scheduler selects the user set with minimum ranking values in both CQI and CDI, thus selects the users with high instantaneous channel gain and small quantization error. Alternatively to the previous centralized protocol, the set of scheduled users can be constructed using a decentralized approach. In such cases, only the subset $\mathcal{L}$ of users whose ranking values are below a threshold is allowed to report their CSIT to the BS. This pre-selection protocol is given by

$$\mathcal{L} = \{k \in \mathcal{K} : r_{g,k} \leq \tau_g \text{ and } r_{d,k} \leq \tau_d\}, \tag{26}$$

where $\mathcal{K}$ is the population of all users, and τ_g, τ_d the thresholds for the channel norm and channel alignment, respectively. The fact that $r_{g,k}$, $r_{d,k}$ are uniformly distributed facilitates the calculation of optimal threshold values.

6. SCHEDULING WITH HETEROGENEOUS USERS

Up to this point, we considered a system with statistically identical users and studied the system throughput when all users exhibit equal average signal-to-noise ratio (SNRs'). However, in a typical wireless network, user channels are not necessarily i.i.d. and mobile terminals experience unequal average signal-to-noise ratio (SNR's) due to different distances from the BS and the corresponding different path losses (near-far effects). Hence, if a max-rate scheduler is used as a means to exploit multiuser diversity, the sum rate will be maximized by transmitting to the users with the strongest channels. As the selected users are highly likely to be the ones closest to the BS, the issue of fairness arises. Restoring fairness requires considering a different scheduling policy that sacrifices capacity for the sake of equalizing the probability that a user is scheduled.

In heterogeneous system configurations, the sum rate is no longer an appropriate performance metric, as it cannot guarantee any fairness constraints and rate balancing among users with nonsymmetric average SNRs. We focus on the problem of maximizing the weighted sum rate in order to reflect the potential fairness issues that arise. Assume that the channel vector of each user can be written as $\mathbf{h}_k = \sqrt{\rho_k}\,\widetilde{\mathbf{h}}_k$, where ρ_k denotes the kth user average SNR and $\widetilde{\mathbf{h}}_k \sim \mathcal{CN}(0, 1)$. The equivalent channel model becomes

$$y_k = \sqrt{\rho_k}\,\widetilde{\mathbf{h}}_k \mathbf{x} + n_k, \quad k = 1, \ldots, K. \tag{27}$$

We consider a weighted sum-rate maximization criterion, which results in the optimization problem

$$\max_{\mathcal{S} \in \mathcal{G}} \sum_{k \in \mathcal{S}} w_k \mathcal{R}_k \quad \text{s.t.} \sum_{k \in \mathcal{S}} w_k = 1 \; w_k \geq 0 \quad \forall k, \tag{28}$$

where $\mathcal{R}_k$ and w_k are the rate and weighting factor of the kth user, respectively.

Let φ_k be the fraction of time slots allocated to user k, with $\sum_{k=1}^{K}\varphi_k = 1$. A general CCDF-based user selection policy on mth beam is defined as

$$k_m^* = \arg\min_{1\le k\le K}\left(1 - F_{X_{k,m}}(x_{k,m})\right)^{1/\varphi_k}. \tag{29}$$

In other words, using the minimum tail scheduler, user k can gain access to the channel with probability φ_k. In [29], it has been shown that this scheduling policy can guarantee equal access to the channel for heterogeneous users. This can also be achieved if ranking-based feedback is employed during the scheduling stage. More formally, let $\mathcal{A}_{k,m}$ be the event that user k is selected on beam m based on ranking-based feedback. If all users have the same time fraction, that is, $\varphi_k = 1/K$, then, following the proof in [29], we have

$$\begin{aligned}\Pr\{\mathcal{A}_{k,m}\} &= \int_0^\infty \Pr\{\mathcal{A}_{k,m} \mid X_{k,m} = x\} f_{X_{k,m}}(x)dx \\ &= -\int_0^\infty \left(1 - F_{X_{k,m}(x)}\right)^{(1-K)/K} dF_{X_{k,m}(x)} = 1/K.\end{aligned} \tag{30}$$

Interestingly, the probability that the kth user is selected $\Pr\{\mathcal{A}_{k,m} = 1\}$ does not depend on the distribution of the other users, even if the users' channels are independent but not necessarily identically distributed. The independence of the selection probability from the other users' statistics can be inferred from the fact that the ranking of each user's CQI follows a uniform distribution independently of the other users' fading characteristics. Thus, in addition to its feedback reduction merits, ranking-based metric can also restore temporal fairness by sharing the scheduling time slots in a fair manner among users.

The average user throughput of a heterogeneous network (non-i.i.d channel distributions) with $\mathcal{M} = 1$ and max-CDF scheduling is studied in [29]. In the appendix, we provide an additional proof of following result [29].

Proposition 5. *The average sum rate, $\mathcal{R}$, of a heterogeneous system in which ranking-based feedback is used for the purposes of user selection is given by*

$$\mathcal{R} = \sum_{m=1}^{\mathcal{M}} K \int_0^1 F_{X_{k,m}}^{-1}(z) z^{K-1} dz. \tag{31}$$

Proof. See Appendix C. □

7. NUMERICAL RESULTS

In this section, we compare the performance of the following schemes.

(i) Scheme I: RBF employing quantized ranking-based CQI for user selection in the scheduling stage.

(ii) Scheme II: RBF in which users are selected based on quantized SNR/SINR feedback in the scheduling stage.

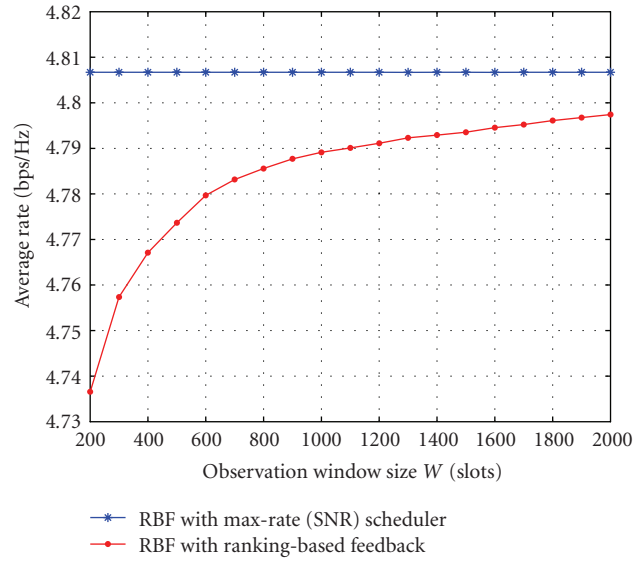

Figure 3: Average rate comparison as a function of window size W for single-beam RBF with $M = 2$ antennas, SNR = 10 dB and $K = 10$ active users. User selection based on ranking-based feedback converges to capacity-optimal max-rate (SNR) scheduling for $W \to \infty$.

As stated above, we consider a two-stage approach, thus the proposed CSIT representation is used solely for selecting the group of scheduled users. Thus, in both schemes under comparison, once the group of users (among all active K ones) is identified in the first stage, the BS requests the transmission rate of the $\mathcal{M}$ selected users in order to perform link adaptation.

In the first set of simulations, we consider single-beam RBF [17] as downlink transmission scheme with $M = 2$ transmit antennas and SNR = 10 dB. In Figure 3, the throughput difference between Scheme I and II is plotted as a function of observation window size W. Expectedly, for small values of W, ranking-based feedback cannot capture sufficiently the CQI distribution, failing to select the users that are on their highest quantile of their distribution. This results in a rate reduction penalty as the system does not exploit multiuser diversity and does not schedule users with large channel gains. As stated in Proposition 1, for W increasing, the performance of ranking-based system converges to that of max-rate scheduler (for $W \to \infty$).

Figures 4 and 5 show the effect of feedback quantization on the system throughput. In Figure 4, the signal-to-noise ratio (SNR) feedback metric is quantized with $B = 5$ bits using the optimal Max-Lloyd algorithm, whereas the ranking-based CQI is quantized using $B = 3$ bits. For different values of W, the proposed feedback representation is able to identify correctly the users with the highest instantaneous rate as compared to the quantized signal-to-noise ratio (SNR) feedback, resulting in capacity gain even with feedback load reduction of 40%. This is mainly due to the inherent digital form of ranking-based CQI and its dynamic range, which allows for efficient compression. In Figure 5, the performance of ranking-based

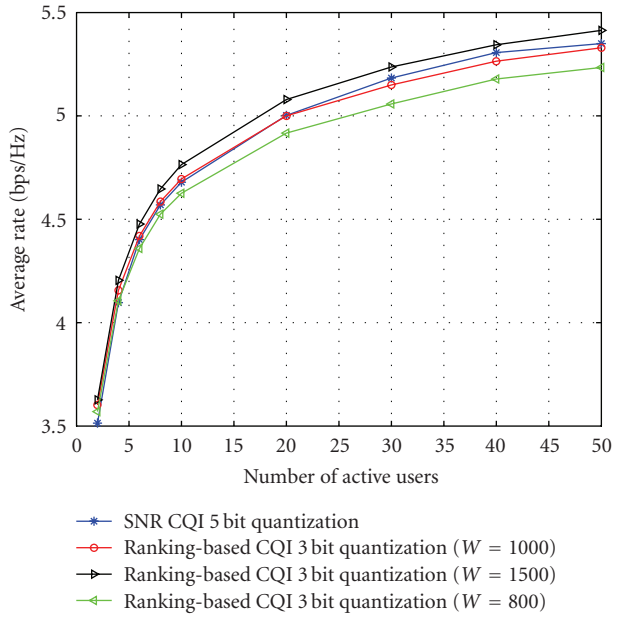

Figure 4: Average rate as a function of the number of users for single-beam RBF with $M = 2$ antennas, for SNR = 10 dB and different values of window size W. With proper choice of W, ranking-based user selection can reduce the feedback load as compared to signal-to-noise ratio (SNR)-based RBF with no throughput reduction.

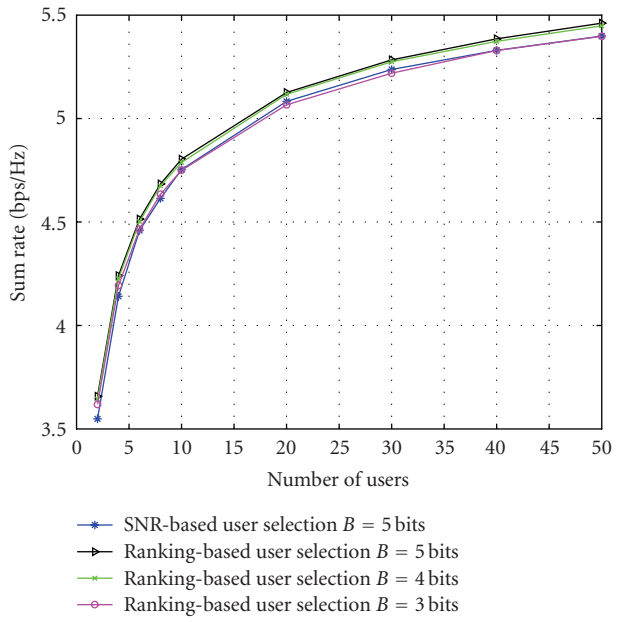

Figure 5: Average rate as a function of the number of users for single-beam RBF with $M = 2$ antennas, SNR = 10 dB, $W = 1000$ slots, and ranking-based CQI metric quantized with different number of bits. The required feedback load can be reduced with almost no expense on the system throughput achieved by quantized signal-to-noise ratio (SNR)-based user selection.

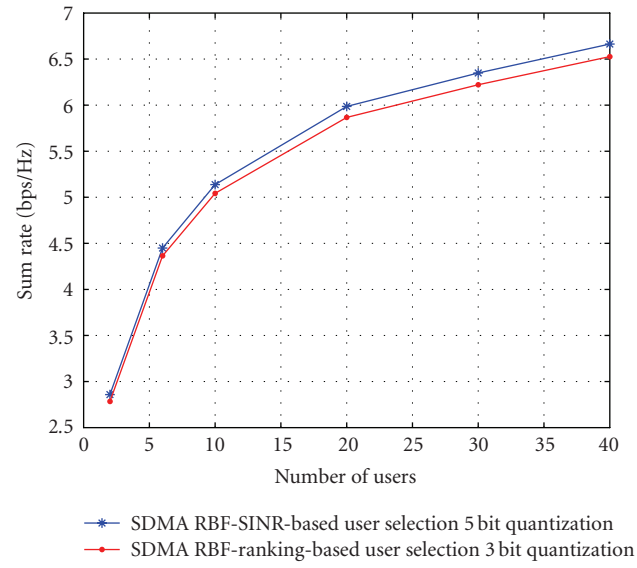

Figure 6: Sum rate as a function of the number of users for multibeam RBF with $M = 2$ antennas, SNR = 10 dB and $W = 1000$ slots. The SINR feedback is quantized using $B = 5$ bits, whereas only 3 bits are used for ranking-based feedback quantization. The feedback reduction advantage of ranking-based representation is preserved in an SDMA context.

user selection for different quantization rates is compared with that of signal-to-noise ratio (SNR)-based user selection for fixed observation window size. The feedback load can be reduced up to 40% with negligible capacity reduction (~0.1 bps/Hz).

In the second set of simulations, the multibeam variant of RBF [18] is used as transmission scheme. The SINR feedback is quantized using $B = 5$ bits, whereas only 3 bits are used for ranking-based CQI quantization. As shown in Figure 6, the proposed feedback representation in an SDMA downlink with $M = 2$ antennas provides similar results as in the single-beam case by representing more efficiently the user selection metric, thus reducing the uplink channel rate with no compromise on the system throughput. A heterogeneous network in which the users' average power are uniformly distributed from −10 to 30 dB is also considered for multibeam RBF with $M = 4$ antennas. The loss in sum rate observed in Figure 7 is expected since in the non-i.i.d. case, the ranking-based feedback does not necessarily select the users with the highest absolute instantaneous CQI values, but those whose instantaneous CQI values are near to a peak with respect to their own distribution. Nevertheless, cell-edge users that enjoy lower averagesignal-to-noise ratio (SNRs') have equal probability of being selected if their CQI values are on the highest quantile. Selecting users with higher pathloss (lower average SNR) results in system throughput reduction, however, temporal fairness is restored as the access time per user is equalized independently as shown in Figure 8.

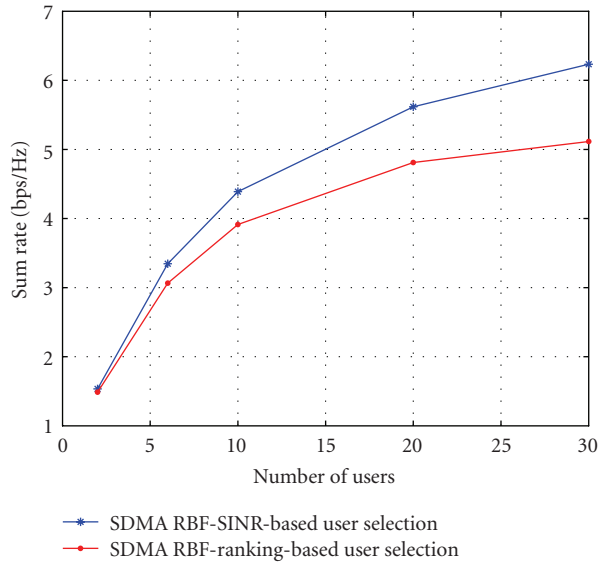

Figure 7: Sum rate as a function of users for multibeam RBF in a heterogeneous network in which users' averagesignal-to-noise ratio (SNRs') range from −10 dB to 30 dB, $M = 4$ antennas, and $W = 1000$ slots. The system throughput reduction is due to the fact that ranking-based feedback does not always select the high SINR users as a means to restore temporal fairness among users with different pathlosses.

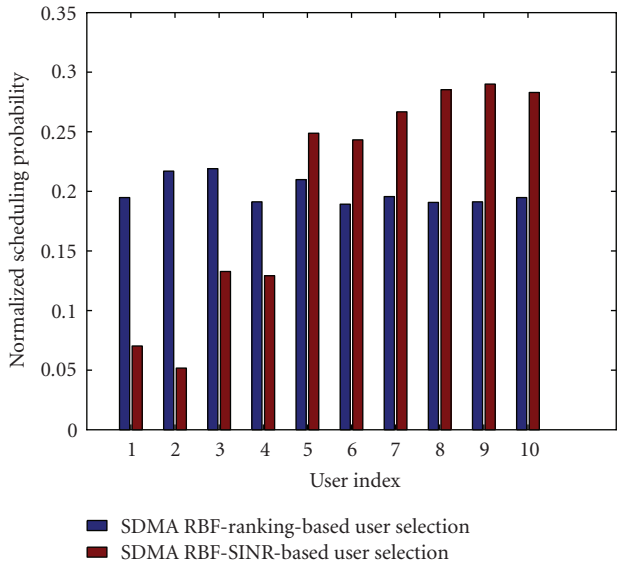

Figure 8: Normalized scheduling probability versus user index for multibeam RBF with $M = 4$ antennas and $K = 10$ users. The users are sorted from the lowest to the highest average signal-to-noise ratio (SNR) and the signal-to-noise ratio (SNR) range is from −10 dB to 30 dB. The probability of selection is equalized among users when ranking-based CQI instead of SINR feedback is employed.

8. CONCLUSION

We considered the problem of feedback reduction in a multiuser multiple-antenna downlink system with more users than transmit antennas, under partial channel knowledge at the transmitter due to limited rate feedback. A novel type of CSIT representation, coined as ranking-based feedback, has been proposed as a means to reduce the required feedback load in the scheduling stage. The performance of random opportunistic beamforming in which users are first selected based on ranking-based metric has been analyzed. When users have i.i.d. channels, it is shown that ranking-based user selection can reduce substantially the uplink feedback rate with negligible decrease in multiuser diversity gain and system throughput. In heterogeneous networks (non-i.i.d. channels), it is shown that temporal fairness is provided at little expense of throughput due to the fact that users have equal access in the channel medium, irrespective to the distribution of other users.

This work opens several interesting questions for future research in low-rate feedback schemes and CSIT representation. First, as ranking-based feedback is in digital form, design of efficient, low-complexity compression, and quantization schemes that can capture the multiuser diversity effects and provide near-optimal performance is of particular interest. Second, the nontrivial tradeoff among sum-rate performance, amount of feedback bits, and observation window size needs to be further explored as a means to provide useful design guidelines and quantify the actual benefits when feedback resources and complexity requirements are carefully accounted for. Another assumption made here is that the channel is instantaneous and error-free. The effect of feedback delay and CSI estimation errors on the performance requires further study, especially in large doppler spread channels where delays are more prominent. Finally, it still remains open to determine which form of channel knowledge representation is sufficient and/or necessary for the transmitter in order to select spatially separable users with large channel gains.

APPENDICES

A. PROOF OF PROPOSITION 1

The ranking $r_{k,m}(t)$, measured over W past samples, provides information about the empirical distribution of the rate process. More formally, $r_k(t)/W \approx 1 - \hat{F}^W_{X_{k,m}}(x)$. We want to show that the difference between $\hat{F}^W_{X_{k,m}}(x)$ and the actual CDF $F_{X_{k,m}}(x)$ vanishes to zero when $W \to \infty$. A measure of closeness of the two functionals, *called maximum discrepancy* (Kolmogorov-Smirnov statistic), is given by

$$D_W = \sup_{-\infty<x<\infty} \left| \hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x) \right| \tag{A.1}$$

whose probability density function is independent of $F(\cdot)$ provided that $F(\cdot)$ is continuous.

Proposition 1 is a direct consequence of the following theorem.

Theorem 2 (Glivenko-Cantelli [39]). *Let* $X_1, X_2, \ldots, X_W \sim F_{X_{k,m}}(x)$*, then the sample paths of* $\hat{F}^W_{X_{k,m}}$ *get uniformly closer to* $F_{X_{k,m}}$ *as* $W \to \infty$*, that is,*

$$\left\| \hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x) \right\|_\infty = \sup_x \left| \hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x) \right| \xrightarrow{as} 0. \tag{A.2}$$

The above theorem implies that for large W, the empirical distribution converges to the distribution function almost surely. Hence $\hat{F}^W_{X_{k,m}}$, which is observed over a window of size W, is almost surely a good approximation for $F_{X_{k,m}}$, and the approximation becomes better as the number of observations increases. In this case, user selection based on ranking-based CSIT becomes equivalent to max-CDF scheduling, which, in turn, is equivalent to max-rate scheduling for large W and i.i.d. channel distributions, that is,

$$\begin{aligned} k^*_m(t) &= \arg\min_{1\le k\le K} r_k(t) \\ &= \arg\min_{1\le k\le K} \left(1 - F_{X_{k,m}}(x_{k,m})(t)\right) \\ &= \arg\max_{1\le k\le K} x_{k,m}(t). \end{aligned} \tag{A.3}$$

B. PROOF OF PROPOSITION 3

Let $F_{X_{k^*_m}}(x) = \Pr\{X_{k^*_m}(t) \le x\}$ be the rate distribution of the selected user k over beam m and let $F_{w,m}(x)$ be the probability that in beam m, the w largest values among W are greater than x, then for a selected user k^* over beam m conditioning on $F_{w,m}(x)$, we have

$$\begin{aligned} \Pr\{X_{k^*_m}(t) \le x\} &= \sum_{w=0}^{W-1} \Pr\{r_{k^*_m}(t) > w\} F_{w,m}(x) \\ &= \sum_{w=0}^{W-1} \left(\frac{W-w}{W}\right)^K F_{w,m}(x), \end{aligned} \tag{B.1}$$

where $\Pr\{r_{k^*_m}(t) > w\} = \Pr\{\min_{1\le k\le K} r_{k,m}(t) > w\} = [1 - F_r(w)]^K = ((W-w)/W)^K$ as the ranking-based CSIT is uniformly distributed with CDF $F_r(w)$ over the set of W past values. Using results from order statistics [40], we have that

$$F_{w,m}(x) = \binom{W}{w} \left(F_{X_{k^*_m}}(x)\right)^{W-w} \left(1 - F_{X_{k^*_m}}(x)\right)^w. \tag{B.2}$$

Therefore, the expected sum rate $\mathcal{R}$ is given by

$$\begin{aligned} \mathcal{R} &= \sum_{m=1}^{\mathcal{M}} \int_0^\infty \Pr\{X_{k^*_m}(t) > x\} dx \\ &= \sum_{m=1}^{\mathcal{M}} \int_0^\infty \left(1 - \Pr\{X_{k^*_m}(t) \le x\}\right) dx \\ &= \sum_{m=1}^{\mathcal{M}} \int_0^\infty 1 - \sum_{w=0}^{W-1} \left(\frac{W-w}{W}\right)^K F_{w,m}(x) dx, \end{aligned} \tag{B.3}$$

which gives (16) as $F_{0,m}(z) = \left(F_{X_{k^*_m}}(x)\right)^W$.

C. PROOF OF PROPOSITION 5

Before proceeding to the proof, we state the following result.

Lemma 1. *The random variable* $U_{k,m} = F_{X_{k,m}}(X_{k,m})$ *is uniformly distributed on the interval* $[0, 1]$.

Proof. In the lines of [29], suppose that x is an arbitrary number and $u = F_{X_{k,m}}(x)$, with $0 \le u \le 1$. The distribution function (CDF) of $U_{k,m}$ is given as

$$\begin{aligned} F_{U_{k,m}}(u) &= \Pr\{U_{k,m} \le u\} \\ &= \Pr\{F_{X_{k,m}}(X_{k,m}) \le u\} \\ &= \Pr\left\{X_{k,m} \le F^{-1}_{X_{k,m}}(u)\right\} = u, \quad 0 \le u \le 1, \end{aligned} \tag{C.1}$$

which implies that $U_{k,m}$ is uniformly distributed on $[0, 1]$.

The average sum rate of multibeam random beamforming is given by

$$\mathcal{R} = \sum_{m=1}^{\mathcal{M}} \mathcal{R}_{k,m}, \tag{C.2}$$

where $\mathcal{R}_{k,m}$ is the average rate of the selected user k on beam m given by

$$\mathcal{R}_{k,m} = \mathbb{E}\left(X^{(K)}_{k,m}\right), \tag{C.3}$$

where $X^{(K)}_{k,m} = \max\{X^1_{k,m}, X^2_{k,m}, \ldots, X^K_{k,m}\}$ (maximum over K i.i.d. random variables) with $X^i_{k,m} \sim X_{k,m}$. As $\mathbb{E}(X^{(K)}_{k,m}) = \mathbb{E}(F^{-1}_{X_{k,m}}(U^{(K)}_{k,m}))$ with $U^{(K)}_{k,m} = \max\{U^1_{k,m}, U^2_{k,m}, \ldots, U^K_{k,m}\}$, from order statistics [40, eqation 3.1.1], we have that

$$\mathbb{E}\left(F^{-1}_{X_{k,m}}\left(U^{(K)}_{k,m}\right)\right) = K \int_0^1 F^{-1}_{X_{k,m}}(z) z^{K-1} dz. \tag{C.4}$$

Putting (C.4) into (C.2) results in (31). □

REFERENCES

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[2] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[3] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *Proceedings of the 38th Conference on Information Sciences and Systems (CISS '04)*, Princeton, NJ, USA, March 2004.

[4] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Transactions on Communications*, vol. 55, no. 1, pp. 11–15, January 2007.

[5] 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, March 2006.

[6] G. Dimić and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, October 2005.

[7] A. Lapidoth, S. Shamai, and M. Wigger, "On the capacity of fading MIMO broadcast channels with imperfect transmitter side-information," in *Proceedings of the 43rd Allerton Conference on Communication, Control and Computing*, Monticello, Ill, USA, September 2005.

[8] S. A. Jafar and A. J. Goldsmith, "Isotropic fading vector broadcast channels: the scalar upper bound and loss in degrees of freedom," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 848–857, March 2005.

[9] D. Gesbert, M. Kountouris, R.W. Heath Jr., C.-B. Chae, and T. Sälzer, "From single user to multiuser communications: shifting the MIMO paradigm," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 36–46, September 2007.

[10] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1423–1436, October 1998.

[11] 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, October 2003.

[12] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2562–2579, October 2003.

[13] S. Zhou, Z. Wang, and G. B. Giannakis, "Quantifying the power loss when transmit beamforming relies on finite-rate feedback," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1948–1957, July 2005.

[14] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, November 2006.

[15] P. Ding, D. J. Love, and M. D. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Transactions on Signal Processing*, vol. 55, no. 7 I, pp. 3417–3428, July 2007.

[16] N. Jindal, "A feedback reduction technique for MIMO broadcast channels," in *Proceedings of IEEE International Symposium on Information Theory (ISIT '06)*, pp. 2699–2703, Seattle, Wash, USA, July 2006.

[17] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[18] 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, Febreuary 2005.

[19] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of IEEE International Conference on Communications (ICC '95)*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[20] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?" in *Proceedings of IEEE International Conference on Communications (ICC '04)*, vol. 1, pp. 234–238, Paris, France, June 2004.

[21] V. Hassel, M.-S. Alouini, D. Gesbert, and G. E. Øien, "Exploiting multiuser diversity using multiple feedback thresholds," in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05)*, vol. 2, pp. 1302–1306, Stockholm, Sweden, May-June 2005.

[22] S. Sanayei and A. Nosratinia, "Opportunistic beamforming with limited feedback," in *Proceedings of the 39th Asilomar Conference on Signals, Systems and Computers*, pp. 648–652, Pacific Grove, CA, USA, October 2005.

[23] S. Sanayei and A. Nosratinia, "Exploiting multiuser diversity with only 1-bit feedback," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '05)*, vol. 2, pp. 978–983, New Orleans, La, USA, March 2005.

[24] J. Diaz, O. Simeone, and Y. Bar-Ness, "How many bits of feedback is multiuser diversity worth in MIMO downlink?" in *Proceedings of the 9th IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '06)*, pp. 505–509, Manaus, Brazil, August 2006.

[25] J. Diaz, O. Simeone, and Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," in *Proceedings of IEEE International Symposium on Information Theory (ISIT '06)*, pp. 1944–1948, Seattle, Wash, USA, July 2006.

[26] K. Huang, R. W. Heath Jr., and J. G. Andrews, "Space division multiple access with a sum feedback rate constraint," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3879–3891, July 2007.

[27] T. Tang and R. W. Heath Jr., "Opportunistic feedback for downlink multiuser diversity," *IEEE Communications Letters*, vol. 9, no. 10, pp. 948–950, October 2005.

[28] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," in *Proceedings of the 5th European Wireless Conference (EW '04)*, Barcelona, Spain, February 2004.

[29] D. Park, H. Seo, H. Kwon, and B. G. Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Transactions on Communications*, vol. 53, no. 11, pp. 1919–1929, November 2005.

[30] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks with heterogeneous users," in *Proceedings of the 38th Conference on Information Sciences and Systems (CISS '04)*, Princeton, NJ, USA, March 2004.

[31] S. Patil and G. de Veciana, "Measurement-based opportunistic scheduling for heterogeneous wireless systems," to appear in IEEE/ACM Transactions on Networking.

[32] M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in *Proceedings of the 6th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC '05)*, vol. 2005, pp. 975–979, New York, NY, USA, June 2005.

[33] R. Zakhour and D. Gesbert, "A two-stage approach to feedback design in multi-user MIMO channels with limited channel state information," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, Athens, Greece, September 2007.

[34] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 642–669, 1956.

[35] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990.

[36] A. György and T. Linder, "Optimal entropy-constrained scalar quantization of a uniform source," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2704–2711, November 2000.

[37] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-Antenna Downlink Channels with Limited Feedback and User Selection," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 25, no. 7, pp. 1478–1491, September 2007.

[38] 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 3, pp. 109–112, Honolulu, Hawaii, USA, April 2007.

[39] P. Billingsley, *Probability and Measure*, John Wiley & Sons, New York, NY, USA, 3rd edition, 1995.

[40] H. A. David and H. N. Nagaraja, *Order Statistics*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.