Exploiting Multiuser Diversity in MIMO Broadcast Channels with Limited Feedback

Marios Kountouris*, *Student Member, IEEE,*
Ruben de Francisco, *Student Member, IEEE,*
David Gesbert, *Senior Member, IEEE,* Dirk T. M. Slock, *Fellow, IEEE,* and
Thomas Sälzer

Abstract

We consider a multiple antenna broadcast channel in which a base station equipped with M transmit antennas communicates with $K \geq M$ single-antenna receivers. Each receiver has perfect channel state information (CSI), whereas the transmitter has partial channel knowledge obtained via a limited rate feedback channel. Built upon multiuser interference bounds, we propose scalar feedback metrics that incorporate information on the channel gain, the channel direction, and the quantization error, with the goal to provide an estimate of the received signal-to-noise plus interference ratio (SINR) at the transmitter. These metrics, combined with efficient user selection algorithms and zero-forcing beamforming on the quantized channel are shown to achieve a significant fraction of the capacity of the full CSIT case by exploiting multiuser diversity. A multi-mode scheme that allows us to switch from multiuser to single-user transmission is also proposed as a means to compensate for the capacity ceiling effect of quantization error and achieve linear sum-rate growth in the interference-limited region. The asymptotic sum-rate performance for large K, as well as in the high and low power regimes, is analyzed and numerical results demonstrate the performance and the advantages of the proposed metrics in different system configurations.

Index Terms

MIMO systems, Broadcast Channel, Multiuser Diversity, Limited Feedback, Scheduling, Channel State information (CSI).

M. Kountouris and T. Sälzer are with France Telecom Research and Development, Issy-les-Moulineaux, France (email: {marios.kountouris, thomas.salzer}@orange-ftgroup.com). Marios Kountouris is the corresponding author (Tel: +33 4 93 00 81 03, Fax: +33 4 93 00 82 00)

R. de Francisco, D. Gesbert and D. T. M. Slock are with Eurecom Institute, Sophia-Antipolis, France (email: {defranci, gesbert, slock}@eurecom.fr).

I. INTRODUCTION

In multiple-input multiple-output (MIMO) broadcast channels, the capacity can be boosted by exploiting the spatial multiplexing capability of transmit antennas and transmit to multiple users simultaneously, by means of Space Division Multiple Access (SDMA), rather than trying to maximize the capacity of a single-user link [1], [2]. As the capacity-achieving dirty paper coding (DPC) approach [3], [4] is difficult to implement, low complexity downlink schemes, such as downlink linear (zero-forcing) beamforming [5], have attracted particular attention, since they achieve a large fraction of DPC capacity while exhibiting reduced complexity [6]–[8]. Nevertheless, the capacity gain of multiuser MIMO systems seems to remain highly sensitive and dependent on the channel state information available at the transmitter (CSIT). If a base station (BS) with M transmit antennas communicating with K single-antenna receivers has perfect channel state information (CSI), a multiplexing gain of $\min(M, K)$ can be achieved. The approximation of close to perfect CSI at the receiver (CSIR) is often reasonable; however, this assumption is unrealistic at the transmitter. Recently, it was shown that if the BS has imperfect channel knowledge, the full multiplexing gain is reduced at high signal-to-noise ratio (SNR) [9], whereas if there is complete lack of CSI knowledge, the multiplexing gain collapses to one [10]. Hence, as the broadcast channel's capacity is sensitive to the accuracy of CSIT, it is of particular interest to identify what kind of partial CSIT can be conveyed to the BS in order to achieve capacity reasonably close to the optimum.

Several limited feedback approaches, imposing a bandwidth constraint on the feedback channel have been studied in point-to-point MIMO systems [11]–[14]. In this context, each user feeds back finite precision (quantized) CSIT 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 $K \leq M$ is made in [15], [16]. In [15] it was shown that the feedback load per user must increase approximately linearly with M and the transmit power (in dB) in order to achieve the full multiplexing gain. All the above schemes rely only on channel direction information (CDI), as no information on the channel magnitude is provided. Recently, MIMO broadcast channels with limited feedback and more users than transmit antennas (i.e. $K \geq M$) have attracted particular interest and several joint beamforming and scheduling schemes aiming to maximize the sum rate have been proposed. A

popular, very low-rate feedback technique, coined as random beamforming (ORBF), is proposed in [17], where M random orthonormal beamforming vectors are generated and the best user on each beam is scheduled. By exploiting multiuser diversity [18], this scheme is shown to yield the optimal capacity growth of $M \log\log K$ for $K \to \infty$. However, the sum rate of this scheme degrades quickly with decreasing number of users.

A different type of limited feedback approaches considers that each user reports CDI related to a codebook back to the transmitter, as well as some form of scalar channel quality information (CQI). As transmission strategy, several beamforming methods have been investigated, including orthogonal unitary beamforming [19], transmit matched-filtering [20], and zero-forcing beamforming [20]–[23]. Note that in the above contributions, the channel gain feedback is considered unquantized for analytical simplicity. Considering the more realistic constraint of finite CSI feedback rate, [24] studies the problem of optimal bit rate allocation between CQI and CDI, and the resulting multiuser diversity - multiplexing gain tradeoff in limited feedback MIMO systems, whereas [25] proposes a threshold-based feedback scheme for SDMA systems under a sum feedback rate constraint.

The benefit of using some form of signal-to-interference plus noise ratio (SINR) as CQI was shown in [26]. One challenge when designing feedback metrics is that the SINR measurement depends, among others, on the channel as well as on the number of other mobiles being simultaneously scheduled along with the user making the measurement. As user cooperation is not allowed, the number of simultaneous users and the available power for each of them will generally be unknown at the mobile. A principal drawback of previous works is that the proposed metrics assume a fixed number of scheduled SDMA users, being also based on non-achievable SINR upper bounds. However, schemes allowing adaptive transition between SDMA and time division multiple access (TDMA) modes, as well as more practical received SINR estimates are of interest. This problem is addressed in [22] and further investigated here.

In this work, we consider a limited rate feedback model for the case when $K \geq M$, where each user is allowed to feed back B-bit quantized information on its channel direction (CDI), complemented with additional instantaneous channel quality information (CQI), as a means to intelligently select M spatially separable users with large channel gains. We study the problem of sum-rate maximization with scheduling and linear precoding in the above setting. For analytical

simplicity, CQI is considered unquantized while the directional information is quantized, however the effect of CQI quantization is explored through simulations. The contributions of this paper are the following:

- We propose several scalar feedback metrics by using bounds on the multiuser interference, which encapsulate information on the channel gain, the channel direction, as well as on the quantization error. These metrics can be interpreted as estimates of the received SINR, providing in parallel awareness of the multiuser interference. This information is in principle not available to the individual users who only have knowledge on their own channels. Complementary to [26] where a CQI metric based on a bound of the expected interference is proposed, we derive bounds on the actual multiuser interference. Similarly to [26], a key observation of this paper is that by using multiuser interference expressions, simplifications arise that give the user the possibility of estimating the individual SINR on its signal as it is detected by the BS.
- We employ these metrics in a system employing linear (zero-forcing) beamforming on the quantized channel directions and greedy user selection. For that, we extend the scheduling algorithm of [8] for the limited feedback case. This algorithm has the advantages that it does not depend on any a priori defined system parameter (such as quantized channels' orthogonality [26]) and is able to switch from multiuser to single-user transmission.
- Using the above precoding setting, we derive upper bounds on the instantaneous multiuser interference that allows us to analytically predict the worst case interference and SINR in a system employing zero-forcing on the quantized channel directions.
- The sum rate of the proposed scheme is analyzed and its asymptotic optimality in terms of capacity growth (i.e. $M \log\log K$) is shown for $K \to \infty$. We obtain sum-rate bounds for the high and low SNR regimes.
- A metric suited for switching the transmission mode from multiuser to single-user is proposed, based on a refined feedback strategy. We show that expectedly single-user mode is preferred as the average SNR increases, whereas multiuser mode is favored when the number of users increases.

The remainder of this paper is organized as follows. The system model is described in Section II, including the considered feedback model and our joint scheduling and beamforming framework. In Section III, three scalar feedback metrics are proposed, and the considered user selection schemes are presented in Section IV. The system sum rate is analyzed in Section V. The performance of the proposed metrics is numerically evaluated in Section VI, and Section VII concludes the paper.

Notation: 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 Euclidean 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 $\log_2(\cdot)$.

II. System Model

We consider a multiple antenna broadcast channel with M antennas at the transmitter and K single-antenna receivers. The received signal y_k of the k-th user is mathematically described as

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

where $\mathbf{x} \in \mathbb{C}^{M \times 1}$ is the transmitted signal, $\mathbf{h}_k \in \mathbb{C}^{M \times 1}$ is the channel vector, and n_k 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 transmitted signal is subject to an average transmit power constraint P, i.e. $\mathbb{E}\{\|\mathbf{x}\|^2\} = P$. We consider i.i.d. flat Rayleigh fading and an homogeneous network where all users have the same average SNR. Due to the noise variance normalization to one, P takes on the meaning of average SNR. We assume that the number of mobiles is greater than or equal to the number of transmit antennas, i.e. $K \geq M$, and we wish to select M out of K users.

A. CDI Finite Rate Feedback Model

Consider a quantization codebook $\mathcal{V}_k = \{\mathbf{v}_{k1}, \mathbf{v}_{k2}, \ldots, \mathbf{v}_{kN}\}$ containing $N = 2^B$ unit norm vectors $\mathbf{v}_{ki} \in \mathbb{C}^M$, for $i = 1, \ldots, N$, which is assumed to be known to both the k-th receiver and

the transmitter.[1] At each time instant t, each receiver k, based on its current channel realization $\mathbf{h}_k$, determines its 'best' vector from the codebook, i.e. the codeword that optimizes a certain cost function. In this paper, we assume that each receiver quantizes its channel to the vector that maximizes the following inner product as done by several authors including [11]–[13]

$$\hat{\mathbf{h}}_k = \mathbf{v}_{kn} = \arg \max_{\mathbf{v}_{ki} \in \mathcal{V}_k} |\bar{\mathbf{h}}_k^H \mathbf{v}_{ki}|^2 = \arg \max_{\mathbf{v}_{ki} \in \mathcal{V}_k} \cos^2(\angle(\bar{\mathbf{h}}_k, \mathbf{v}_{ki})) \tag{2}$$

where the normalized channel vector $\bar{\mathbf{h}}_k = \mathbf{h}_k / \|\mathbf{h}_k\|$ corresponds to the channel direction, and we refer to $\hat{\mathbf{h}}_k$ as the k-th user channel quantization. The BS exploits the quantized channel information in order to design the downlink beams.

Evidently, a codebook designed by quantizing the actual (not normalized) channel $\mathbf{h}_k$ would be optimal; however, as an optimal vector quantizer is difficult to obtain and analyze, and as beamforming on the quantized spatial information is used by our proposed scheme, the vector $\bar{\mathbf{h}}_k$ that we need to quantize is constrained to be unit-norm; hence it lies on the unit hypersphere, whereas the channel instantiation $\mathbf{h}_k$ lies anywhere in the $\mathbb{C}^M$ space.

Each user sends the corresponding quantization index n back to the transmitter through an assumed error-free, and zero-delay feedback channel using $B = \lceil \log_2 N \rceil$ bits. The error-free assumption can be well approximated through the use of sufficiently powerful error correcting codes over the feedback link, whereas the zero-delay assumption is valid when the processing and feedback delays are small relative to the channel's coherence time. These are classical assumptions, although could be challenged in some situations. However we choose to emphasize the problem of feedback metric design rather than the effect of an imperfect feedback channel which is left for future work.

Additionally we decide not to focus on the particular subproblem of optimal codebook design. Evidently, the performance of a system relying on quantized CSI depends on the codebook choice. However, the problem of optimum codebook design is not yet fully solved and beyond the scope of our paper. See e.g. [15], [27] for the performance loss relative to optimum vector quantization, when suboptimal codebooks, such as random vector quantization (RVQ), are used.

[1]The complexity of generating a different codebook for each user can be reduced by generating a common, general codebook $\mathcal{V}_g$ known at both ends of the link, and afterwards each user obtains its specific codebook through random unitary rotation of $\mathcal{V}_g$. In that case, each code-vector is independent from user to user.

B. Joint Scheduling and Beamforming with limited feedback

We focus here on the case of linear beamforming schemes, where exactly M spatially separated users access the channel simultaneously on the downlink. In this case the joint scheduling and beamforming problem can be stated as follows. Let $\mathbf{w}_k$ and s_k be the (normalized) beamforming vector and data symbol of the k-th user, respectively. Define $\mathbf{H} \in \mathbb{C}^{K \times M}$ as the concatenation of all user channels, $\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2 \dots \mathbf{h}_K]^H$, where the k-th row is the channel of the k-th receiver ($\mathbf{h}_k^H$).

Let $\mathcal{G}$ be the set of all possible subsets of cardinality $|\mathcal{G}|$ of disjoint indices among the complete set of user indices $\{1, \cdots, K\}$. Let $\mathcal{S} \in \mathcal{G}$ be one such candidate group of $|\mathcal{S}| = \mathcal{M} \leq M$ users selected for transmission at a given time slot. Then $\mathbf{H}(\mathcal{S})$, $\mathbf{W}(\mathcal{S})$, $\mathbf{s}(\mathcal{S})$, $\mathbf{y}(\mathcal{S})$ are the concatenated channel vectors, normalized beamforming vectors, uncorrelated data symbols and received signals respectively for the set of scheduled users. The signal model is given by

$$\mathbf{y}(\mathcal{S}) = \mathbf{H}(\mathcal{S})\mathbf{W}(\mathcal{S})\sqrt{\mathbf{P}}\mathbf{s}(\mathcal{S}) + \mathbf{n} \tag{3}$$

where $\mathbf{P}$ is a diagonal power allocation matrix.

As here the base station relies on quantized CSI, we use zero-forcing beamforming on the quantized channel directions available at the transmitter as a multiuser transmission strategy. Although nonlinear precoding schemes can achieve a better sum rate than linear beamforming, they often exhibit more complexity and a lack of robustness with respect to imperfect CSIT. Zero-forcing is a linear beamformer that can be implemented with reduced complexity and is asymptotically optimal at high SNR or for large K [7], [8]. Apart from its simplicity and tractability, further motivation for choosing zero-forcing beamforming along the quantized channel directions comes from [28], where the optimality of beamforming with quantized feedback is derived under certain conditions.

The beamforming matrix is then given by

$$\mathbf{W}(\mathcal{S}) = \hat{\mathbf{H}}(\mathcal{S})^{\dagger} = \hat{\mathbf{H}}(\mathcal{S}) \left(\hat{\mathbf{H}}(\mathcal{S})^H \hat{\mathbf{H}}(\mathcal{S}) \right)^{-1} \mathbf{\Lambda} \tag{4}$$

where $\hat{\mathbf{H}}(\mathcal{S})$ is a matrix whose columns are the quantized channels $\hat{\mathbf{h}}_k$ (codevectors) of the users selected for transmission and $\mathbf{\Lambda}$ is a diagonal matrix that normalizes the columns of $\mathbf{W}(\mathcal{S})$. Note that in contrast to the perfect CSIT case, the finite precision on the available CSIT is such that the multiuser interference cannot be eliminated perfectly.

Assuming Gaussian input distribution, the sum rate is given by

$$\mathcal{R} = \mathbb{E}\left\{\sum_{k\in\mathcal{S}} \log\left(1 + SINR_k\right)\right\} \tag{5}$$

where the SINR at the k-th receiver is

$$SINR_k = \frac{P_k|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum\limits_{j\in\mathcal{S}-\{k\}} P_j|\mathbf{h}_k^H\mathbf{w}_j|^2 + 1} \tag{6}$$

with $\sum_{i\in\mathcal{S}} P_i = P$ in order to satisfy a power constraint on the transmitted signal. For simplicity, equal power allocation is considered and thus $P_i = \frac{P}{\mathcal{M}}, \forall i \in \mathcal{S}$.

III. Efficient CQI Metrics exploiting Multiuser Diversity

In the case where $K \geq M$, there is need for user selection. Ideally, the selection would be based on maximizing the sum rate. However, (6) depends on $\mathbf{h}_k$ that is unknown at the BS. Therefore, some kind of channel quality information is necessary to be available at the BS besides the quantization index of the normalized channel. Intuitively, this CQI feedback ought to enable us to select a set of M users with mutually orthogonal channels, large channel gains and small quantization errors. In other words, a good metric has to incorporate information on the channel norm and the quantized channel direction, as well as on the channel quantization error.

In this section, we consider the problem of an efficient design of channel quality feedback. Our objective is to seek scalar feedback metrics, denoted as ξ_k, that allow us to exploit the multiuser diversity and achieve close to optimum sum-rate performance. We first review an existing metric, and then proceed to expand over new metrics. The tradeoff between these metrics is clarified later.

A. Metric I: Bounding the expected multiuser interference

Let $\phi_k = \angle(\hat{\mathbf{h}}_k, \bar{\mathbf{h}}_k)$ denote the angle between the quantized channel direction and the normalized channel vector. We consider that each user provides information on its effective channel (SINR) by feeding back the following scalar metric

$$\xi_k^I = \frac{P\left\|\mathbf{h}_k\right\|^2 \cos^2\phi_k}{P\left\|\mathbf{h}_k\right\|^2 \sin^2\phi_k + M} \tag{7}$$

proposed previously in [26], [29], [30]. This type of CQI encapsulates information on the channel gain as well as the CDI quantization error, defined as $\sin^2\phi_k = 1 - \left|\hat{\mathbf{h}}_k^H \bar{\mathbf{h}}_k\right|^2$. The above metric results from bounding the expected SINR by using the expected value of multiuser interference due to quantization and an upper bound on the expected received signal power. Clearly, the above SINR cannot be achieved, but it can be interpreted as an upper bound (UB) on each user's received SINR in a system where equal power is allocated over M beamforming vectors. Remarkably, it offers a good estimate of the multiuser interference at the mobile side without any receiver cooperation, and in a way that depends only on the user of interest's channel. If the M beamforming vectors at the transmitter are perfectly orthogonal (i.e. the columns of $\hat{\mathbf{H}}(\mathcal{S})$ are orthogonal), then $\hat{\mathbf{H}}(\mathcal{S})$ is unitary and $\mathbf{W}(\mathcal{S}) = \hat{\mathbf{H}}(\mathcal{S})$, the upper bound becomes tight, yielding a received per-user SINR_k equal to the one predicted by (7).

B. Metric II: Bounding the actual multiuser interference

The previous metric presents the following two major limitations: a) it is appropriate for block fading channels with sufficiently long Gaussian codebooks in order to average the capacity over the fading distribution, b) the SINR values predicted by (7) are not achievable, since in general the beamforming vectors are not perfectly orthogonal, especially in networks with low to moderate number of users; thus a second step of feedback from the users scheduled for transmission is required for outage-free rate allocation.

In order to avoid the need for a second step of feedback and guarantee an outage-free transmission, we propose to use a *lower bound* on the k-th user actual SINR rather than an upper bound, which is based on an upper bound on the *actual* multiuser interference rather than on an upper bound of the expected interference. Furthermore, the SINR estimated by (7) does not take into account the fact that zero-forcing is used as transmission scheme. Here we derive interference bounds that also consider the misalignment between the actual channel and the zero-forcing beamformers, which results to additional power loss.

1) Multiuser Interference Bounds: Let $\mathbf{w}_k$ be the normalized zero-forcing beamforming vector intended for the k-th user, with channel alignment $\cos\theta_k = \left|\bar{\mathbf{h}}_k^H \mathbf{w}_k\right|$ and channel norm $\|\mathbf{h}_k\|^2$. Define the matrix $\mathbf{\Psi}_k(\mathcal{S}) = \sum_{j\in S, j\neq k} \mathbf{w}_j \mathbf{w}_j^H$, the operator $\lambda_{max}\{\cdot\}$, which returns the largest eigenvalue, and $\mathbf{U}_k \in \mathbb{C}^{M\times(M-1)}$ an orthonormal basis spanning the null space of $\mathbf{w}_k$.

Theorem 1: Given an arbitrary set of unit-norm beamforming vectors of the users in $\mathcal{S}$, an upper bound on the interference over the normalized channel, defined as $\bar{\mathrm{I}}_k(\mathcal{S}) = \sum_{j\in\mathcal{S},j\neq k} \left|\bar{\mathbf{h}}_k^H \mathbf{w}_j\right|^2$, experienced by the k-th user is given by

$$\bar{\mathrm{I}}_k(\mathcal{S}) \leq \cos^2\theta_k \alpha_k(\mathcal{S}) + \sin^2\theta_k \beta_k(\mathcal{S}) + 2\sin\theta_k\cos\theta_k \gamma_k(\mathcal{S}) \tag{8}$$

where

$$\begin{cases} \alpha_k(\mathcal{S}) = \mathbf{w}_k^H \boldsymbol{\Psi}_k(\mathcal{S}) \mathbf{w}_k \\ \beta_k(\mathcal{S}) = \lambda_{max}\{\mathbf{U}_k^H \boldsymbol{\Psi}_k(\mathcal{S}) \mathbf{U}_k\} \\ \gamma_k(\mathcal{S}) = \left\|\mathbf{U}_k^H \boldsymbol{\Psi}_k(\mathcal{S}) \mathbf{w}_k\right\| \end{cases} \tag{9}$$

Proof: See Appendix I.A.

Consider now that we impose an ϵ-orthogonality constraint between two quantized channel vectors $\hat{\mathbf{h}}_i$ and $\hat{\mathbf{h}}_j$, i.e. $\left|\hat{\mathbf{h}}_i^H \hat{\mathbf{h}}_j\right| \leq \epsilon$, and define the worst-case orthogonality under zero-forcing beamforming as $\epsilon_{ZF} = \max_{i,j\in\mathcal{S}} \left|\mathbf{w}_i^H \mathbf{w}_j\right|$. For notation simplification, the dependence on $\mathcal{S}$ can be dropped, expressing the worst interference received by the k-th user in terms of $\cos\theta_k$ and ϵ_{ZF}.

Lemma 1: The worst-case orthogonality of the set of M zero-forcing normalized beamforming vectors (ϵ_{ZF}) and alignment with the normalized channel ($\cos\theta_k$) are bounded as a function of $\cos\phi_k$ and $\epsilon < \frac{1}{M-1}$ as follows:

$$\epsilon_{ZF} \leq \vartheta \tag{10}$$

$$\cos\theta_k \geq \frac{\left|\cos\phi_k - \sqrt{\vartheta}\right|}{1+\vartheta} \tag{11}$$

$$\text{with } \vartheta = \frac{\epsilon}{1-(M-1)\epsilon}$$

Proof: See Appendix I.B.

The following result can now be obtained:

Theorem 2: Given a user set $\mathcal{S}$ of cardinality $|\mathcal{S}| = M$ and constrained to be ϵ-orthogonal, a system that performs zero-forcing beamforming can guarantee the following SINR for the k-th user

$$SINR_k^{ZF} \geq \frac{P\left\|\mathbf{h}_k\right\|^2 \cos^2\theta_k}{P\left\|\mathbf{h}_k\right\|^2 \overline{I}_{UB_k} + M} \tag{12}$$

where

$$\overline{I}_{UB_k} = (M-1)\left(\vartheta\cos\theta_k + \sin\theta_k\right)^2 - (M-2)(1-\vartheta)\sin^2\theta_k \tag{13}$$

with $\cos\theta_k = \frac{\left|\cos\phi_k - \sqrt{\vartheta}\right|}{1+\vartheta}$ and $\vartheta = \frac{\epsilon}{1-(M-1)\epsilon}$.

Proof: See Appendix I.C.

2) CQI feedback metric: Motivated by the above results, we propose that each user provides information on a lower bound on its SINR by reporting the following scalar metric

$$\xi_k^{II} = \frac{\frac{P}{(1+\vartheta)^2}\left\|\mathbf{h}_k\right\|^2(\cos\phi_k - \sqrt{\vartheta})^2}{P\left\|\mathbf{h}_k\right\|^2 \bar{I}_{UB_k} + M} \tag{14}$$

The basic difference between (7) and (14) is on the estimation of both the multiuser interference and the received signal. The normalized (channel direction) vector $\bar{\mathbf{h}}_k$ can be expressed as $\bar{\mathbf{h}}_k = \sqrt{1-\sin^2\phi_k}\hat{\mathbf{h}}_k + \sqrt{\sin^2\phi_k}\hat{\mathbf{h}}_k^{\perp}$, where $\hat{\mathbf{h}}_k^{\perp}$ is the normalized projection of $\bar{\mathbf{h}}_k$ onto the orthogonal complement of $\hat{\mathbf{h}}_k$. Note that the actual phase information in $\bar{\mathbf{h}}_k$ is omitted since it is not relevant for SINR computation. In (7) the interference is replaced by an upper bound on its average value, i.e. $\mathbb{E}\left\{\sum_{j\in\mathcal{S}\setminus\{k\}}\frac{P}{M}\left\|\mathbf{h}\right\|^2|\bar{\mathbf{h}}_k\mathbf{w}_j|^2\right\} = \frac{P(|\mathcal{S}|-1)}{M(M-1)}\left\|\mathbf{h}\right\|^2\sin^2\phi_k \leq \frac{P}{M}\left\|\mathbf{h}\right\|^2\sin^2\phi_k$. This bound is due to the fact that

$$|\bar{\mathbf{h}}_k^H\mathbf{w}_j|^2 = (1-\sin^2\phi_k)|\hat{\mathbf{h}}_k^H\mathbf{w}_j|^2 + \sin^2\phi_k|\hat{\mathbf{h}}_k^{\perp H}\mathbf{w}_j|^2 = \sin^2\phi_k|\hat{\mathbf{h}}_k^{\perp H}\mathbf{w}_j|^2, \quad \forall k\neq j$$

as the zero-forcing beamforming vector $\mathbf{w}_j$ is chosen orthogonal to the quantized channel vectors of all other users, i.e., $\hat{\mathbf{h}}_k^H\mathbf{w}_j = 0$ for all $k\neq j$, $k\in\mathcal{S}$, and $\mathbb{E}\left\{\sum_{j\in\mathcal{S}\setminus\{k\}}|\hat{\mathbf{h}}_k^{\perp H}\mathbf{w}_j|^2\right\} = \frac{|\mathcal{S}|-1}{M-1}$ [15]. In contrast to that, for metric II the upper bound on the actual multiuser interference (13) is used.

For the received signal, we have that $\cos^2(\angle(\bar{\mathbf{h}}_k,\mathbf{w}_k)) \geq \cos^2(\angle(\bar{\mathbf{h}}_k,\hat{\mathbf{h}}_k)+\angle(\hat{\mathbf{h}}_k,\mathbf{w}_k))$ as $\angle(\bar{\mathbf{h}}_k,\mathbf{w}_k) \leq \angle(\bar{\mathbf{h}}_k,\hat{\mathbf{h}}_k)+\angle(\hat{\mathbf{h}}_k,\mathbf{w}_k)$, and $\cos x$ is a monotonically decreasing function in x for the interval of interest. Metric I can be viewed as a SINR estimate assuming that the quantized channel $\hat{\mathbf{h}}_k$ and zero-forcing beamforming vector $\mathbf{w}_k$ coincide, i.e. $\angle(\hat{\mathbf{h}}_k,\mathbf{w}_k)=0$. This assumption becomes valid for large number of users K. Hence in metric I, the following approximation is used $\cos^2(\angle(\bar{\mathbf{h}}_k,\mathbf{w}_k)) \approx \cos^2(\angle(\bar{\mathbf{h}}_k,\hat{\mathbf{h}}_k))$, whereas in metric II the power loss introduced by the angle shift due to the misalignment of $\hat{\mathbf{h}}_k$ and $\mathbf{w}_k$ is bounded using lemma 1.

If $\epsilon = 0$ as in the case of metric I, we have $\cos^2\theta_k = \cos^2\phi_k$, thus $\bar{I}_{UB_k} = \sin^2\theta_k = \sin^2\phi_k$ and $\frac{(\cos\phi_k-\sqrt{\vartheta})^2}{(1+\vartheta)^2} = \cos^2\phi_k$, and (14) takes exactly the form of (7). Note that as metric I is derived under the ideal assumption $\epsilon = 0$ is not achievable. However metric II is a lower bound on the user SINR at each slot and can be used for rate adaptation, although it can be a pessimistic SINR estimate during some channel realizations.

A limitation of both metric I and II is that they provide an estimate on the SINR by assuming that $\mathcal{M} = M$ users are to be scheduled. However, in the case of limited CSIT, in the high SNR regime and/or for low number of users, it is often beneficial from a sum-rate point of view to transmit to $\mathcal{M} < M$ users. The system should then offer a highly desirable adaptivity between SDMA of various orders and even TDMA. This problem has not been addressed before. For that reason, a different form of CQI feedback that offers flexibility on estimating the user SINR for various values of $\mathcal{M}$ is of interest.

C. Soft transition from SDMA to TDMA

Here, we capitalize on the idea of [31] and decompose the CQI, letting each user feed back the following two scalar values (in addition to the quantized channel index): 1) the square of the alignment $\cos^2\phi_k$, and 2) the channel norm, $\|\mathbf{h}_k\|$. This feedback strategy enables the BS to calculate a more accurate SINR estimate for any set of scheduled users with cardinality $\mathcal{M} \leq M$ as a more accurate estimate on the multiuser interference can be calculated by having the CQI in the form of channel gain and quantization error. This strategy also enables a multi-mode scheme where the BS switches between single-user transmission mode (TDMA) and multiuser mode (SDMA). Note that under a certain finite feedback rate constraint each scalar value is quantized with reduced accuracy compared to the case of only a single scalar CQI metric (e.g. metric I and II). The effect of CQI quantization is studied through simulations in Section VI, where it can be seen that the reduced precision of the two scalar CQIs does not reduce the sum-rate performance compared to the one scalar CQI case. Based on Theorem 2, we propose that the scheduler selects the user based on the following lower bound on the received SINR, referred to as *metric III*:

$$\xi_k^{III} = \frac{P\left\|\mathbf{h}_k\right\|^2 \rho_k^2}{P\left\|\mathbf{h}_k\right\|^2 \overline{I}_{UBd_k} + \mathcal{M}} \tag{15}$$

where

$$\rho_k^2 = \cos^2(\phi_k + \angle(\hat{\mathbf{h}}_k, \mathbf{w}_k)) \tag{16}$$

and

$$\overline{I}_{UBd_k} = \rho_k^2 \alpha_k(\mathcal{S}) + \left(1-\rho_k^2\right)\beta_k(\mathcal{S}) + 2\rho_k\sqrt{1-\rho_k^2}\gamma_k(\mathcal{S}) \tag{17}$$

which can be explicitly calculated at the transmitter using (9). For $\epsilon \to 0$, we have $\bar{\mathrm{I}}_{UBd_k} \to \sin^2\phi_k$, and when $\epsilon = 0$ the following *metric IV*, interpreted as an upper bound on the received

SINR, can be used

$$\xi_k^{IV} = \frac{P \left\| \mathbf{h}_k \right\|^2 \rho_k^2}{P \left\| \mathbf{h}_k \right\|^2 \sin^2 \phi_k + \mathcal{M}} \tag{18}$$

Actually, setting ϵ to be inversely proportional to K, it can be seen from Lemma 1 that as $K \to \infty$, $\epsilon_{ZF} \to 0$, and $\cos\theta_k \to \cos\phi_k$. Thus, for $K \to \infty$, $\bar{\mathrm{I}}_{UB_k} = \sin^2\phi_k$ and hence (18) converges to (7) for $\mathcal{M} = M$.

As it can be seen, (15) provides a more accurate SINR estimate compared to (14) as $\rho_k^2 \geq \frac{(\cos\phi_k - \sqrt{\vartheta})^2}{(1+\vartheta)^2}$ and $\bar{I}_{UBd_k} \leq \bar{I}_{UB_k}$. Furthermore, as $\rho_k^2 \leq \cos^2\phi_k$, we have that $\xi_k^I \geq \xi_k^{IV} \geq \xi_k^{III} \geq \xi_k^{II}$. The difference is that ξ_k^{II} and ξ_k^{III} can always be supported by the system and can be used for outage-free rate allocation, whereas ξ_k^I and ξ_k^{IV} are upper bounds that are not achievable in general.

IV. User Scheduling Schemes

The metrics in Section III are combined with two user selection algorithms in a system employing zero-forcing beamforming. As our optimization objective is to maximize the system capacity, the optimum policy under max-sum-rate scheduling is to select $\mathcal{M} \leq M$ users among K users that maximize the sum rate through exhaustive search. As the complexity of such a combinatorial optimization problem is prohibitively high for large K, we resort to low-complexity scheduling strategies based on greedy user selection (see e.g. [7], [8], [26]).

A. Greedy-SUS algorithm

We first review a heuristic scheduling algorithm based on semi-orthogonal user selection (SUS) [7], [26]. Using ξ_k defined either as $(7), (14), (15), (18)$ and $\hat{\mathbf{h}}_k$, $k = 1, \ldots, K$, the BS performs user selection to support up to M out of K users at each time slot. The algorithm is outlined in Table I. The first user is selected from the set $\mathcal{Q}^0 = \{1, \ldots, K\}$ of cardinality $|\mathcal{Q}^0| = K$ as the one having the highest channel quality, i.e. $k_1 = \arg\max_{k \in \mathcal{Q}^0} \xi_k$. The $(i+1)$-th user, for $i = 1, \ldots, M-1$, is selected as $k_{i+1} = \arg\max_{k \in \mathcal{Q}^i} \xi_k$ among the user set $\mathcal{Q}^i$ with cardinality $|\mathcal{Q}^i| \leq K$, defined as $\mathcal{Q}^i = \left\{ k \in \mathcal{Q}^{i-1} \mid |\hat{\mathbf{h}}_k^H \hat{\mathbf{h}}_j| \leq \epsilon \ \forall j \in \mathcal{S} \right\}$. The system parameter ϵ defines the maximum allowed non-orthogonality (maximum correlation) between quantized channels and it is a parameter set in advance. Evidently, if ϵ is very large, the selected users may cause

significant multiuser interference, reducing the system sum rate. Conversely, if ϵ is too small, the scheduler cannot find enough semi-orthogonal users to transmit to.

B. Greedy-US algorithm

We generalize here the low-complexity greedy user selection (GUS) scheme [8] for the case of quantized CSIT. Each user feeds back its quantized channel direction based on a predetermined codebook $\mathcal{V}_k$ and scalar instantaneous feedback ξ_k, which are used to perform joint scheduling and beamforming with quantized CSIT. The algorithm is summarized in Table II, where $\mathcal{R}(\mathcal{S}_i) = \sum_{k\in\mathcal{S}_i} \log_2(1+\xi_k)$, with ξ_k being either: ξ_k^I, ξ_k^{II}, ξ_k^{III} or ξ_k^{IV}, and $\mathcal{S}_i$ the set of selected users up to the i-th step. The user with the highest rate (equivalently ξ_k metric) among K users is first selected, and at each iteration, a user is added only if the sum rate (based on the estimated SINR) is increased. At each step, it is important to re-process the set of previously selected users (thus, re-calculating the zero-forcing beamformers) once a user is added to the set $\mathcal{S}_i$.

A main advantage of Greedy-US algorithm compared to Greedy-SUS is that it does not necessarily require the use of the predetermined system parameter ϵ. The value of the orthogonality constraint ϵ affects the performance of the user selection algorithm. If ϵ is set too small, the multiuser diversity gain decreases, and the user set $\mathcal{Q}^i$ can be empty before M quasi-orthogonal users are found. The optimal value decreases with K, as the probability of finding M semi-orthogonal users among K is larger, however it is difficult to be optimized analytically. Furthermore, as the Greedy-US algorithm re-processes the set of already selected users under the same zero-forcing beamforming optimization when one user is added at each step, its performance can be better or equal to that of Greedy-SUS algorithm.

V. Performance Analysis

In this section, we analyze the sum-rate performance of a system using metric (14) in conjunction with Greedy-SUS algorithm (referred to as scheme II) for the case of $K \to \infty$, and at low and high SNR. We consider an approximate codebook design [21], [32], coined as Approximate Vector Quantization (AVQ), which yields a lower bound to the quantization error.

A. Asymptotic (in K) sum-rate analysis

We analyze the sum-rate performance of scheme II considering the asymptotic case of $K \to \infty$ and M fixed. As (14) is a lower bound on the user's SINR, the expected sum rate $\mathcal{R}$ of scheme

II is lower bounded as

$$\mathcal{R} \geq \mathbb{E}\left\{\sum_{i=1}^{M} \log_2\left(1+\xi_{k_i}^{II}\right)\right\} = \mathbb{E}\left\{\sum_{i=1}^{M} \log_2\left(1+\max_{k\in\mathcal{K}_i} \xi_k^{II}\right)\right\} \tag{19}$$

where $\mathcal{K}_i = K\kappa_{i-1}$ captures the multiuser diversity gain reduction due to greedy user selection. The above lower bound on the expected sum rate is due to the fact that the exact SINR is unknown at either the BS or receiver. The actual SINR for all metrics can be obtained through a second stage of link adaptation.

From the user selection procedure, we have that the CQI metric ξ_{k_i} of the selected user at the i-th step of the algorithm, k_i, is equal to the maximum of $\mathcal{K}_i = |\mathcal{Q}^{i-1}|$ i.i.d. random variables with common cumulative distribution function (CDF) $F_\xi(x)$. Obviously, the multiuser diversity gain of $\log|\mathcal{Q}^0| = \log K$ is experienced only from the first selected user and decreases with the user index. A bound on the cardinality of $|\mathcal{Q}^i|$ can be calculated through the probability that a user i in $\mathcal{Q}^i$ is ϵ-orthogonal to users in $\mathcal{Q}^{i-1}$, which is equal to $I_{\epsilon^2(i,M-i)}$, where $I_x(a,b)$ is the regularized incomplete beta function. The k_i-th user is the one that has the maximum SINR-like metric among $\mathcal{Q}^{i-1}$, whose cardinality converges to the following value (by using the law of large numbers) [33], [34]:

$$\left|\mathcal{Q}^{i-1}\right| \approx KPr\{\mathbf{v}\in\mathcal{Q}^{i-1}\} \geq KI_{\epsilon^2}(i-1, M-i+1)$$

with $|\mathcal{Q}^0| = K$.

For large number of users K and choosing $\epsilon = 1/\log K$, so that $\lim_{K\to\infty} KI_{\epsilon^2}(i-1,M-i+1) = \infty$ and $\lim_{K\to\infty} \epsilon = 0$, we have that $\xi_k^{II} \to \xi_k^I$. Denoting $\beta = \frac{1}{N}\cdot(P/M)^{M-1}$, the following theorem can be proved:

Theorem 3: The sum rate of the proposed scheme $\mathcal{R}$ converges to the optimum capacity of MIMO broadcast channel $\mathcal{R}_{opt}$, for $K\to\infty$, i.e.

$$\lim_{K\to\infty}(\mathcal{R}_{opt} - \mathcal{R}) = \lim_{K\to\infty}\left[M\log_2\frac{1+\frac{P}{M}\log K}{1+\frac{P}{M}\log\left(\frac{K}{\beta}\right)}\right] = 0 \tag{20}$$

with probability one.

Proof See Appendix II.

The above theorem implies that the optimal $M\log\log K$ capacity growth can be achieved for $K\to\infty$ by using the proposed metric (14) with greedy user selection algorithm and beamforming on the channel quantizations. Note also that this notion of sum rate convergence

is stronger than that capacity ratio convergence as the latter cannot guarantee that there is no infinite SINR gap between the two methods.

B. Sum-rate performance in the interference-limited region

Here, we characterize the sum-rate performance of scheme II in the high-power regime (interference-limited region). For $P \to \infty$, it can be shown that

Theorem 4: The sum rate of scheme II at high SNR with finite B and K is upper bounded by

$$\mathcal{R} \leq \frac{M}{M-1}\left(B + \frac{log_2 e}{\kappa_{max}} H_K\right) \tag{21}$$

where $H_K = \sum_{k=1}^{K} \frac{1}{k}$ is the harmonic number and $\kappa_{max} = \max_{i=1,\ldots,M} \kappa_{i-1}$.

Proof See Appendix III.

The above theorem implies that the system becomes interference-limited and its sum rate converges to a constant value at high SNR, even for arbitrary large but finite B and K. This behavior is inherited to all finite feedback-based MISO systems due to the quantization error, which results to loss of the multiplexing gain at high SNR. Furthermore, as $\partial\mathcal{R}/\partial M < 0$, the sum rate is a monotonically decreasing function with M, implying that at high SNR the sum rate is maximized by using $M = 1$ beams.

For large number of users ($K \to \infty$), the harmonic number can be asymptotically expanded as $H_K = \gamma + \log K + \frac{1}{2K} + O(\frac{1}{12K^2})$, resulting to $\lim_{K\to\infty} H_K = \log K + \gamma$, where γ is the Euler-Mascheroni constant. Thus, the sum rate at high SNR and $K \to \infty$ exhibits logarithmic growth with K due to the multiuser diversity gain. In other words, for fixed B, although only a fraction of the full multiplexing gain is achieved ($r = \frac{M}{M-1}$), the sum rate scales as $\log K$, compensating for the loss in degrees of freedom and 'shifting' the interference-limited region to higher SNR values. Similar results on the asymptotic (in average SNR) sum-rate behavior are presented in [21].

C. Sum-rate performance in the low-power regime

The low-power regime corresponds to the noise-limited region ($P \to 0$), in which for the scheme II we have that $\xi_k^{II} = \frac{P}{M(1+\vartheta)^2} \|\mathbf{h}_k\|^2 (\cos\phi_k - \sqrt{\vartheta})^2 \leq \frac{P}{M} \|\mathbf{h}_k\|^2 \cos^2\phi_k$.

Lemma 2: The distribution of $X = \|\mathbf{h}_k\|^2 \cos^2\phi_k$ is given by

$$f_X(x) = (1-\delta)\sum_{k=0}^{\infty} \frac{\zeta_k x^{k+M-1} e^{-x/(1-\delta)}}{(1-\delta)^{k+M}\Gamma(k+M)} u(x) \tag{22}$$

where $u(\cdot)$ is the unit step function, and ζ_k is obtained recursively by

$$\begin{cases} \zeta_0 = 1 \\ \zeta_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} \left(\delta^i\right) \zeta_{k+1-i} \qquad k = 0, 1, 2, \ldots \end{cases} \tag{23}$$

Proof: Since the random variable $X = \|\mathbf{h}_k\|^2 \cos^2 \phi_k$ can be described as the sum of two independent Gamma distributed random variables with parameters $(1, 1)$ and $(M-1, 1-\delta)$ [21], equation (22) is derived, after some manipulations, by applying results from [35].

Based on the above lemma, upper bounding the coefficients ζ_k by one and using eq. (3.381.2) in [36], the following result can be shown:

Proposition 1: The random variable X stochastically dominates the random variable $\tilde{X}$, or $F_X(x) \leq F_{\tilde{X}}(x)$, whose CDF is given by

$$F_{\tilde{X}}(x) = \int_0^x (1 - e^{-t/(1-\delta)})^{M-1} dt \tag{24}$$

Proof: See Appendix IV.

Replacing the CDF of X by (24), we derive a sum-rate lower bound (for finite B) as

$$\mathcal{R} \geq \sum_{i=1}^{M} \int_0^{\infty} \log_2 \left(1 + \frac{P}{M} x\right) dF_{\tilde{X}}^{\mathcal{K}_i}(x) \approx \log_2 e \frac{P}{M} \sum_{i=1}^{M} \int_0^{\infty} x dF_{\tilde{X}}^{\mathcal{K}_i}(x) \tag{25}$$

where the approximation follows from considering the first-order terms of Taylor series expansion around $P = 0$.

VI. Numerical Results

In this section, we evaluate the performance of a system that performs zero-forcing beamforming using the proposed scalar feedback metrics and scheduling algorithms through simulations. We consider $M = 2$ transmit antennas, orthogonality constraint $\epsilon = 0.4$ and codebooks generated using random vector quantization [15], [27]. The achieved sum rate is compared with two alternative transmission techniques for the MIMO downlink, random beamforming [17] and zero-forcing beamforming with full CSI (and equal power allocation).

In Figure 1 we compare the sum-rate performances of the proposed CQI metrics as a function of the average SNR, for K=30 users and B = 4 bits per user for CDI quantization. Metric I and II achieve similar sum-rate, exhibiting however the same bounded behavior at high SNR, where the system capacity converges to a constant value. Given a fixed number of CDI bits B, the

system becomes unavoidably interference-limited at high SNR and the rate curves flatten out. This is due to the fact that the accuracy of knowledge of the quantization error remains constant while SNR increases, and also that the Greedy-SUS algorithm forces the system to schedule always M users. Nevertheless, the scheme using Metric III in (15) combined with Greedy-US algorithm provides more flexibility by transmitting to $\mathcal{M} \leq M$ users, thus keeping a linear sum-rate growth in the interference-limited region and converging to TDMA for $P \to \infty$, where $\mathcal{M} = 1$ is optimal.

In Figure 2 we plot the sum rate as a function of K for average SNR = 20dB and codebook of size B = 4 bits. It can be seen that all scalar metrics can efficiently benefit from the multiuser diversity gain. The gap with respect to the full CSIT case can be decreased by increasing the feedback load B. However, the slightly different scaling of metric IV is due to the fact that the user selection estimates the sum-rate and thus the regions where $\mathcal{M} < M$ beams ought to be used based on limited feedback. Thus, erroneous estimations can sometimes lead to sub-optimal decisions in terms of the number of users to be scheduled. Furthermore, in a system with fixed orthogonality factor ϵ, the accuracy of the lower bound (metric II) does not improve as K increases. On the other hand, the upper bound (metric I) becomes more realistic due to a higher probability of finding orthogonal quantized channels, hence yielding slightly better user selection.

In order to evaluate the effect of CQI quantization, we consider a system in which each user has in total 10 bits available for feedback. A sum-rate comparison as a function of the number of users for SNR = 20 dB is shown in Figure 3. We use B bits for feeding back the index of the quantized channel and the remaining $(10-B)$ bits for CQI quantization. For metric IV, 2 bits are used for quantization of the channel norm and 3 bits for the alignment. The random beamforming scheme uses $B = 1$ bits in order to specify the chosen transmitted beam ($B = \lceil \log_2 M \rceil$) and the remaining ($9$ bits) for SINR quantization. A simple quantization technique has been used that minimizes the mean squared distortion (max Lloyd algorithm). For this amount of available feedback, it can be seen that for the simulated range of K, 6 bits are enough to capture a large portion of multiuser diversity and preserve the scaling (case $B = 4$). Note also that the performance is similar to that of Figure 2, in which the CQI metrics are considered unquantized.

VII. Conclusion

We studied a multiple antenna broadcast channel with more users than transmit antennas, in which partial CSIT is conveyed via a limited rate feedback channel. We proposed scalar feedback metrics which, combined with efficient joint scheduling and zero-forcing beamforming, can achieve a significant fraction of the capacity of the full CSI case by means of multiuser diversity. These metrics are built upon multiuser interference bounds and incorporates information on both channel gain and quantization error. A scheme that combines these metrics with zero-forcing beamforming and efficient user selection algorithms is considered and its sum-rate performance is investigated. Furthermore, an adaptive scheme, switching from multiuser to single-user transmission mode is also proposed, exhibiting linear sum-rate growth in the interference-limited region. The merits of the proposed schemes in terms of sum rate are studied both analytically and through numerical results.

Appendix I
Multiuser Interference Bounds

A. Proof of Theorem 1

Before proceeding to the proof of Theorem 1, we first state the following result.

Lemma 3: Let $\mathbf{U}_k \in \mathbb{C}^{M\times(M-1)}$ be an orthonormal basis spanning the null space of $\mathbf{w}_k$. Then,

$$\left\|\bar{\mathbf{h}}_k^H \mathbf{U}_k\right\|^2 = 1 - \cos^2\theta_k \tag{26}$$

Proof: Define the orthonormal basis $\mathbf{Z}_k$ of $\mathbb{C}^M$ obtained by stacking the column vectors of $\mathbf{U}_k$ and $\mathbf{w}_k$: $\mathbf{Z}_k = [\mathbf{U}_k \mathbf{w}_k]$. Since $\mathbf{Z}_k\mathbf{Z}_k^H = \mathbf{I}$ and $\bar{\mathbf{h}}_k$ has unit power

$$\left\|\bar{\mathbf{h}}_k^H \mathbf{Z}_k\right\|^2 = \bar{\mathbf{h}}_k^H \mathbf{Z}_k\mathbf{Z}_k^H \bar{\mathbf{h}}_k = \bar{\mathbf{h}}_k^H \bar{\mathbf{h}}_k = 1 \tag{27}$$

Then, by definition of $\mathbf{Z}_k$ we can separate the power of $\bar{\mathbf{h}}_k$ as follows

$$\left\|\bar{\mathbf{h}}_k^H \mathbf{Z}_k\right\|^2 = \left\|\bar{\mathbf{h}}_k^H \left[\mathbf{U}_k \mathbf{w}_k\right]\right\|^2 = \left\|\bar{\mathbf{h}}_k^H \mathbf{U}_k\right\|^2 + \left|\bar{\mathbf{h}}_k^H \mathbf{w}_k\right|^2 = 1 \tag{28}$$

Setting $\left|\bar{\mathbf{h}}_k^H \mathbf{w}_k\right|^2 = \cos^2\theta_k$ and solving the above equation for $\left\|\bar{\mathbf{h}}_k^H \mathbf{U}_k\right\|^2$ we obtain the desired result.

Now we can proceed to the proof of Theorem 1. Using the definition of $\mathbf{\Psi}_k(\mathcal{S})$ and defining

$\pi_k^2 = \cos^2\theta_k$, the interference over the normalized channel for user k and index set $\mathcal{S}$, denoted as $\overline{I}_k(\mathcal{S})$, can be expressed as

$$\overline{I}_k(\mathcal{S}) = \sum_{i\in\mathcal{S}, i\neq k} \left|\overline{\mathbf{h}}_k^H \mathbf{w}_i\right|^2 = \sum_{i\in\mathcal{S}, i\neq k} \overline{\mathbf{h}}_k^H \mathbf{w}_i \mathbf{w}_i^H \overline{\mathbf{h}}_k = \overline{\mathbf{h}}_k^H \mathbf{\Psi}_k(\mathcal{S}) \overline{\mathbf{h}}_k \tag{29}$$

The normalized channel $\overline{\mathbf{h}}_k$ can be expressed as a linear combination of orthonormal basis vectors. Using *Lemma 3*, all possible unit-norm $\overline{\mathbf{h}}_k$ vectors with $\left|\overline{\mathbf{h}}_k^H \mathbf{w}_k\right| = \pi_k$ can be written as follows

$$\overline{\mathbf{h}}_k = \pi_k e^{j\alpha_k} \mathbf{w}_k + \sqrt{1-\pi_k^2}\mathbf{U}_k \mathbf{B}_k \mathbf{e}_k \tag{30}$$

where $\mathbf{B}_k$ is a diagonal matrix with entries $e^{j\beta_i}, i = 1, \ldots, M-1$ and $\mathbf{e}_k$ is an arbitrary unit-norm vector in $\mathbb{C}^{M-1}$. The complex phases β_i and α_k are unknown and lie in $[0, 2\pi]$. Substituting (30) into (29) we get

$$\begin{aligned} \overline{I}_k(\mathcal{S}) &= \pi_k^2\ \mathbf{w}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{w}_k \\ (a) \qquad &+ (1-\pi_k^2)\, \mathbf{e}_k^H \mathbf{B}_k^H \mathbf{U}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{U}_k \mathbf{B}_k \mathbf{e}_k \\ (b) \qquad &+ \pi_k \sqrt{1-\pi_k^2}\, [e^{-j\alpha_k} \mathbf{w}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{U}_k \mathbf{B}_k \mathbf{e}_k \\ &\qquad + \mathbf{e}_k^H \mathbf{B}_k^H \mathbf{U}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{w}_k e^{j\alpha_k}] \end{aligned} \tag{31}$$

Since the first term in (31) is perfectly known, the upper bound on $\overline{I}_k(\mathcal{S})$ is found by joint maximization of the summands (a) and (b) with respect to α_k, $\mathbf{B}_k$ and $\mathbf{e}_k$. We use a simpler optimization method, which consists of bounding separately each term.

(a) Defining $\mathbf{A}_k(\mathcal{S}) = \mathbf{U}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{U}_k$ for clarity of exposition, the second term can be bounded as follows

$$\begin{aligned} &\max_{\mathbf{B}_k, \mathbf{e}_k} \left(1-\pi_k^2\right) \mathbf{e}_k^H \mathbf{B}_k^H \mathbf{A}_k(\mathcal{S}) \mathbf{B}_k \mathbf{e}_k = \left(1-\pi_k^2\right) \lambda_{max}\{\mathbf{A}_k(\mathcal{S})\} \\ &s.t.\ \|\mathbf{e}_k\| = 1 \end{aligned} \tag{32}$$

where the operator $\lambda_{max}\{\cdot\}$ returns the largest eigenvalue. The maximum in (32) is obtained when the vector $\mathbf{B}_k \mathbf{e}_k$ equals the principal eigenvector of the matrix $\mathbf{A}_k(\mathcal{S})$.

(b) Defining $\mathbf{q}_k = \mathbf{B}_k^H \mathbf{U}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{w}_k e^{j\alpha_k}$ and noting that the matrix $\mathbf{\Psi}_k(\mathcal{S})$ is Hermitian by construction, the bound on the third term in (31) can be written as follows

$$\begin{aligned} &\max_{\mathbf{q}_k, \mathbf{e}_k} \pi_k \sqrt{1-\pi_k^2} \left[\mathbf{q}_k^H \mathbf{e}_k + \mathbf{e}_k^H \mathbf{q}_k\right] = \max_{\mathbf{q}_k} 2\pi_k \sqrt{1-\pi_k^2}\, \|\mathbf{q}_k\| \\ &s.t.\ \|\mathbf{e}_k\| = 1 \end{aligned} \tag{33}$$

The left hand side is maximized for $\mathbf{e}_k = \frac{\mathbf{q}_k}{\|\mathbf{q}_k\|}$, which satisfies the unit-norm constraint, yielding the modified bound in (33). The solution is given by

$$\max_{\mathbf{q}_k} 2\pi_k\sqrt{1-\pi_k^2}\,\|\mathbf{q}_k\| = \max_{\mathbf{B}_k,\alpha_k} 2\pi_k\sqrt{1-\pi_k^2}\,\left\|\mathbf{B}_k^H\mathbf{U}_k^H\mathbf{\Psi}_k(S)\mathbf{w}_k e^{j\alpha_k}\right\| = 2\pi_k\sqrt{1-\pi_k^2}\,\left\|\mathbf{U}_k^H\mathbf{\Psi}_k(S)\mathbf{w}_k\right\| \tag{34}$$

Finally, incorporating into (31) the bounds obtained in (32) and (34) we obtain the desired bound.

B. Proof of Lemma 1

By noting that ϵ_{ZF} corresponds to the maximum possible amplitude of the off-diagonal terms of $\left(\hat{\mathbf{H}}_k^H\hat{\mathbf{H}}_k\right)^{-1}$, and under the (not restrictive) assumption $\epsilon < \frac{1}{M-1}$, the bound on ϵ_{ZF} is found by bounding the amplitude of the off-diagonal terms in the Neumann series $\sum_{n=1}^{\infty} \mathit{offdiag}\left(\hat{\mathbf{H}}_k^H\hat{\mathbf{H}}_k\right)^n$, where $\mathit{offdiag}(\cdot)$ takes the off-diagonal part setting the elements in the diagonal to zero. By representing the non-normalized zero-forcing beamforming vectors as the sum of $\hat{\mathbf{h}}_k$ and its orthogonal complement $\tilde{\mathbf{w}}_k$, i.e. $\mathbf{w}_k = \hat{\mathbf{h}}_k + \tilde{\mathbf{w}}_k$ and bounding the amplitude of the diagonal terms of $\mathbf{I} + \sum_{n=1}^{\infty} \mathit{offdiag}\left(\hat{\mathbf{H}}_k^H\hat{\mathbf{H}}_k\right)^n$, we obtain the desired bound on $\cos\theta_k$.

C. Proof of Theorem 2

By using the definition of each user's $SINR_k$, $\cos\theta_k$ and equal power allocation, we have the following equalities

$$SINR_k^{ZF} = \frac{P\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2}{\sum_{j\in\mathcal{S},j\neq k} P|\mathbf{h}_k^H\mathbf{w}_j|^2 + M} = \frac{P\left\|\mathbf{h}_k\right\|^2\cos^2\theta_k}{P\left\|\mathbf{h}_k\right\|^2\sum_{j\in\mathcal{S},j\neq k}\left|\overline{\mathbf{h}}_k^H\overline{\mathbf{w}}_j\right|^2 + M} \tag{35}$$

We aim to find an upper bound on the multiuser interference given by *Theorem 1* that takes into account the worst-case orthogonality ϵ_{ZF}. Hence, the dependence on $\mathcal{S}$ can be dropped, expressing the worst interference received by the k-th user in terms of $\cos\theta_k$ and ϵ_{ZF}. In this case, the following bounds can be easily derived for equation (9)

$$\begin{cases} \alpha_k \leq (M-1)\epsilon_{ZF}^2 \\ \beta_k \leq 1 + (M-2)\epsilon_{ZF} \\ \gamma_k \leq (M-1)\epsilon_{ZF} \end{cases} \tag{36}$$

Hence, by substituting these values in equation (8), we obtain the upper bound $\overline{I}_k = \cos^2\theta_k\ (M-1)\epsilon_{ZF}^2 + \sin^2\theta_k\ [1+(M-2)\epsilon_{ZF}] + 2\sin\theta_k\cos\theta_k(M-1)\epsilon_{ZF} \leq \sin^2\theta_k$. By substituting $\epsilon_{ZF} = \vartheta$

and $\cos\theta_k = \frac{\left|\cos\phi_k - \sqrt{\vartheta}\right|}{1+\vartheta}$ (i.e. inequalities (10) and (11), respectively become equalities), where $\vartheta = \frac{\epsilon}{1-(M-1)\epsilon}$ in the previous expression, we have the upper bound given by (13). Using this bound on the SINR_k expression derived in (35), we obtain the SINR bound in equation (12).

APPENDIX II
PROOF OF THEOREM 3

Let $\xi_{k_i}^I$ denote the upper bound on the achieved SINR of user k_i (i.e. the user selected at the i-th iteration, for $i = 1, 2, \ldots, M$. From Theorem 1 in [21], we have that

$$\Pr\left\{u_{\mathcal{K}_1} - \frac{P}{M}\log\log\sqrt{K} \leq \xi_{k_1}^I \leq u_{\mathcal{K}_1} + \frac{P}{M}\log\log\sqrt{K}\right\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

with $u_{\mathcal{K}_1} = \frac{P}{M}\log(\frac{K}{\beta}) - \frac{P(M-1)}{M}\log\log(\frac{K}{\beta})$.

For $i = 2, \ldots, M$, we obtain

$$\Pr\left\{u_{\mathcal{K}_i} - \frac{P}{M}\log\log\sqrt{K} \leq \xi_{k_i}^I \leq u_{\mathcal{K}_i} + \frac{P}{M}\log\log\sqrt{K}\right\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

with $u_{\mathcal{K}_i} = \frac{P}{M}\log(\frac{\mathcal{K}_i}{\beta}) - \frac{P(M-1)}{M}\log\log(\frac{\mathcal{K}_i}{\beta})$.

From the user selection procedure, we have that $\xi_{k_1}^I \geq \xi_{k_2}^I \geq \ldots \geq \xi_{k_M}^I$, and after some manipulations it can be shown that for large K, we have

$$\Pr\left\{u_{\mathcal{K}_i} - \frac{P}{M}\log\log\sqrt{K} \leq \xi_{k_i}^I \leq u_{\mathcal{K}_1} + \frac{P}{M}\log\log\sqrt{K}\right\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

Since $\log(\cdot)$ is an increasing function, we have that

$$\Pr\left\{\log_2\left(1 + u_{\mathcal{K}_i} - \tfrac{P}{M}\log\log\sqrt{K}\right) \leq \log_2\left(1+\xi_{k_i}^I\right) \leq \log_2\left(1 + u_{\mathcal{K}_1} + \tfrac{P}{M}\log\log\sqrt{K}\right)\right\} \geq 1 - O\left(\tfrac{1}{\log K}\right)$$

Hence,

$$\lim_{K\to\infty}\Pr\left\{\frac{\log_2\left(1 + u_{\mathcal{K}_i} - \frac{P}{M}\log\log\sqrt{K}\right)}{\log_2(\frac{P}{M}\log K)} \leq \frac{\log_2\left(1+\xi_{k_i}^I\right)}{\log_2(\frac{P}{M}\log K)} \leq \frac{\log_2\left(1 + u_{\mathcal{K}_1} + \frac{P}{M}\log\log\sqrt{K}\right)}{\log_2(\frac{P}{M}\log K)}\right\} \geq 1 - O\left(\frac{1}{\log K}\right) \tag{37}$$

By substituting $u_{\mathcal{K}_1}$ and $u_{\mathcal{K}_i}$ in the above equation, we conclude that the LHS and the RHS of the inequalities both converge to one as $K \to \infty$, therefore

$$\lim_{K\to\infty}\frac{\mathcal{R}}{\log_2\left(\frac{P}{M}\log K\right)} = 1 \tag{38}$$

with probability one. Assuming equal power allocation and that M perfectly orthogonal users can be found, as $\Pr\{|\mathcal{S}| = M\} \overset{K\to\infty}{\to} 1$, we have that the proposed scheme achieves a sum rate of $M \log_2\left(\frac{P}{M}\log K\right)$.

An upper bound on $\mathcal{R}_{opt}$ is given in [6], where

$$\Pr\left\{\frac{\mathcal{R}_{opt}}{M} \leq \log_2\left(1+\frac{P}{M}\left(\log K + O(\log\log K)\right)\right)\right\} \geq 1 - O\left(\frac{1}{\log^2 K}\right)$$

Thus,

$$\begin{aligned}&\Pr\Big\{\log_2(1+\xi_{k_i}^I) - \tfrac{\mathcal{R}_{opt}}{M} \geq \\ &\log_2\left(1+u_{\mathcal{K}_i} - \tfrac{P}{M}\log\log\sqrt{K}\right) - \log_2\left(1+\tfrac{P}{M}\left(\log K + O(\log\log K)\right)\right)\Big\} \\ &\geq 1 - O\left(\tfrac{1}{\log K}\right) - O\left(\tfrac{1}{\log^2 K}\right)\end{aligned}$$

where the RHS of the inequality inside the Pr goes to zero for $K \to \infty$. As a result, for large K, we have that

$$0 \leq \log_2(1+\xi_{k_i}^I) - \frac{\mathcal{R}_{opt}}{M}, \quad i = 1, \ldots, M$$

with probability one, which results to (20) for $K \to \infty$, as $\mathcal{R}_{opt}$ is an upper bound on the sum rate of our proposed scheme.

APPENDIX III
PROOF OF THEOREM 4

For $P \to \infty$, we have

$$\xi_k^{II} = \lim_{P\to\infty} \frac{\frac{P}{(1+\vartheta)^2}\left\|\mathbf{h}_k\right\|^2(\cos\phi_k - \sqrt{\vartheta})^2}{P\left\|\mathbf{h}_k\right\|^2 \overline{I}_{UB_k} + M} = \frac{(\cos\phi_k - \sqrt{\vartheta})^2}{(1+\vartheta)^2 \overline{I}_{UB_k}} \leq \cot^2\phi_k \tag{39}$$

whose probability density function (PDF) is given by $f_{\cot^2\phi}(x) = \frac{(M-1)N}{(1+x)^M}$, for $x \geq (1-\delta)/\delta$ and zero elsewhere [21].

The expected sum rate for a user set $\mathcal{S}$ (of cardinality M) is given by

$$\mathcal{R} \leq \mathbb{E}\left\{\sum_{i=1}^{M}\log_2(1+\max_{k_i\in\mathcal{K}_i}\cot^2\phi_{k_i})\right\} = \sum_{i=1}^{M}\int_0^\infty \log_2(1+x)dF^{\mathcal{K}_i}_{\cot^2\phi}(x)dx =$$

$$= \sum_{i=1}^{M}\mathcal{K}_i\int_{\frac{1-\delta}{\delta}}^\infty \log_2(1+x)\frac{2^B(M-1)}{(1+x)^M}\left(1-\frac{2^B}{(1+x)^{M-1}}\right)^{\mathcal{K}_i-1}dx$$

$$\overset{(a)}{=} 2^B(M-1)\sum_{i=1}^{M}\mathcal{K}_i\sum_{k=0}^{\mathcal{K}_i-1}\binom{\mathcal{K}_i-1}{k}(-1)^k\int_{\frac{1-\delta}{\delta}}^\infty \log_2(1+x)\frac{2^{Bk}}{(1+x)^{k(M-1)+M}}dx$$

$$= \frac{\log_2 e}{M-1} \sum_{i=1}^{M} \mathcal{K}_i \sum_{k=0}^{\mathcal{K}_i - 1} \binom{\mathcal{K}_i - 1}{k} (-1)^k \left[\frac{B \log 2}{k+1} + \frac{1}{(k+1)^2} \right] \overset{(b)}{=} \frac{\log_2 e}{M-1} \sum_{i=1}^{M} (B \log 2 + H_{\mathcal{K}_i}) \tag{40}$$

where (a) follows from binomial expansion and to get (b) the Nörlund-Rice integral representation is applied [37]. Combining (40) with $\mathcal{K}_i \leq K\kappa_{max}$, we get (21).

APPENDIX IV
PROOF OF PROPOSITION 1

For the CDF of X we have

$$F_X(x) = \int_0^x f_X(t)dt \overset{(a)}{\leq} \int_0^x \frac{\gamma(M-1, \frac{t}{1-\delta})}{\Gamma(M-1)} dt \overset{(b)}{<} \int_0^x (1 - e^{-t/(1-\delta)})^{M-1} dt \tag{41}$$

where $\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt$ is the lower incomplete gamma function. Note that (a) follows by upper bounding the coefficients ζ_k in (23) by one and (b) holds from Alzer's inequality [38].

REFERENCES

[1] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

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

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

[4] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. of 38th Conf. Inform. Sciences and Systems (CISS'04)*, Princeton, NJ, Mar. 2004.

[5] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Sig. Processing*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

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

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

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

[9] A. Lapidoth and S. Shamai (Shitz), "Collapse of degrees of freedom in MIMO Broadcast with finite precision CSI," in *Proc. of 43rd Allerton Conf. on Commun., Control and Comput.*, Monticello, Illinois, USA, Sept. 2005.

[10] S. A. Jafar and A. Goldsmith, "Isotropic fading vector broadcast channels: The scalar upper bound and loss in degrees of freedom," *IEEE Trans. Inform. Theory*, vol. 51, no. 3, pp. 848–857, Mar. 2005.

[11] 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 Sel. Areas in Commun. (JSAC)*, vol. 16, no. 8, pp. 1423–1436, Oct. 1998.

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

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

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

[15] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[16] P. Ding, D. Love, and M. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. Sig. Processing*, vol. 55, no. 7, pp. 3417–3428, July 2007.

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

[18] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, Seattle, June 1995, pp. 331–335.

[19] K. Huang, J.G. Andrews and R.W. Heath, Jr., "Orthogonal beamforming for SDMA downlink with limited feedback," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP'07)*, Hawaii, USA, Apr. 2007.

[20] C. Swannack, G. Wornell, and E. Uysal-Biyikoglu, "MIMO Broadcast Scheduling with Quantized Channel State Information," in *Proc. of IEEE Int. Symp. Inform. Theory (ISIT'06)*, Seattle, Washington, USA, July 2006, pp. 1788–1792.

[21] T.Yoo, N. Jindal, and A. Goldsmith, "Multi-Antenna Broadcast Channels with Limited Feedback and User Selection," *to appear in IEEE Jour. Sel. Areas in Commun. (JSAC)*, 2007.

[22] 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 *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP'07)*, Hawaii, USA, Apr. 2007.

[23] M. Trivellato, F. Boccardi, and F. Tosato, "User selection schemes for MIMO broadcast channels with limited feedback," in *Proc. IEEE Vehic. Tech. Conf. (VTC'07 - Spring)*, Dublin, Ireland, Apr. 2007.

[24] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Sälzer, "Multiuser diversity - Multiplexing tradeoff in MIMO broadcast channels with limited feedback," in *Proc. of 40th IEEE Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA (invited paper), Oct. 2006.

[25] K. Huang, R.W. Heath, Jr. and J.G. Andrews, "Space Division Multiple Access with a Sum Feedback Rate Constraint," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP'07)*, Hawaii, USA, Apr. 2007.

[26] T. Yoo, N. Jindal, and A. Goldsmith, "Finite-rate feedback MIMO broadcast channels with a large number of users," in *Proc. of IEEE Int. Symp. Inform. Theory (ISIT'06)*, Seattle, Washington, USA, July 2006, pp. 1214–1218.

[27] C. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 458–462, Feb. 2007.

[28] S. Srinivasa and S. Jafar, "Vector channel capacity with quantized feedback," in *Proc. IEEE Int. Conf. on Communications (ICC'05)*, Seoul, S.Korea, May 2005, pp. 2674–2678.

[29] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Sälzer, "Efficient metric for Scheduling in MIMO Broadcast Channels with Limited Feedback," FT060303 - *France Telecom R&D internal report*, Mar. 2006.

[30] N. Jindal, "Finite Rate Feedback MIMO Broadcast Channels," in *Workshop on Inform. Theory and its Applications (ITA)*, UC San Diego, USA (invited paper), Feb. 2006.

[31] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Sälzer, "Low complexity scheduling and beamforming for multiuser MIMO systems," in *Proc. IEEE Sig. Proc. Adv. on Wir. Commun. (SPAWC'06)*, Cannes, France, July 2006.

[32] J. C. Roh and B. D. Rao, "Transmit beamforming in multiple-antenna systems with finite rate feedback: A VQ-based approach," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1101–1112, Mar. 2006.

[33] T. Yoo and A. Goldsmith, "Sum-rate optimal multi-antenna downlink beamforming strategy based on clique search," in *Proc. IEEE Global Tel. Conf. (GLOBECOM'05)*, St. Louis, MO, Dec. 2005, pp. 1510–1514.

[34] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "Finding NEMO: Near Mutually Orthogonal sets and applications to MIMO broadcast scheduling," in *Proc. Int. Conf. on Wir. Networks, Commun. and Mob. Computing*, June 2005, pp. 1035 – 1040.

[35] P. G. Moschopoulos, "The distribution of the sum of independent gamma random variables," *Mathematics of Computation*, vol. 66, pp. 541–544, 1997.

[36] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic Press Inc., 1994.

[37] N. E. Nörlund, *Vorlesungen über Differenzenrechnung*. New York: Chelsea Publishing Company, 1954.

[38] H. Alzer, "On some inequalities for the incomplete gamma function," *Mathematics of Computation*, vol. 66, no. 218, pp. 771–778, Apr. 1997.

TABLE I

OUTLINE OF GREEDY-SUS ALGORITHM

Step 0 set $\mathcal{S} = \emptyset$, $\mathcal{Q}^0 = 1, \ldots, K$

For $i = 1, 2, \ldots, M$ repeat

Step 1 $k_i = \arg \max_{k \in \mathcal{Q}^{i-1}} \xi_k$

Step 2 $\mathcal{S} = \mathcal{S} \cup k_i$

Step 3 $\mathcal{Q}^i = \left\{ k \in \mathcal{Q}^{i-1} \mid |\hat{\mathbf{h}}_k^H \hat{\mathbf{h}}_{k_i}| \leq \epsilon \right\}$

TABLE II

OUTLINE OF GREEDY-US ALGORITHM

Step 0 Initialization: Set $\mathcal{S}_0 = \emptyset$, $\mathcal{R}(\mathcal{S}_0) = 0$, and $\mathcal{Q}^0 = 1, \ldots, K$

Step 1 $k_1 = \arg \max_{k \in \mathcal{Q}^0} \xi_k$

Set $\mathcal{S}_1 = \mathcal{S}_0 \cup \{k_1\}$

While $i < M$ repeat

$i \leftarrow i + 1$

Step 2 $k_i = \arg \max_{k \in (\mathcal{Q}^0 - \mathcal{S}_{i-1})} \mathcal{R}(\mathcal{S}_{i-1} \cup \{k\})$

Step 3 Set $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{k_i\}$

if $\mathcal{R}(\mathcal{S}_i) \leq \mathcal{R}(\mathcal{S}_{i-1})$

Step 4 finish algorithm and $i \leftarrow i - 1$

Step 5 Set $\mathcal{S} = \mathcal{S}_i$ and $\mathcal{M} = i$

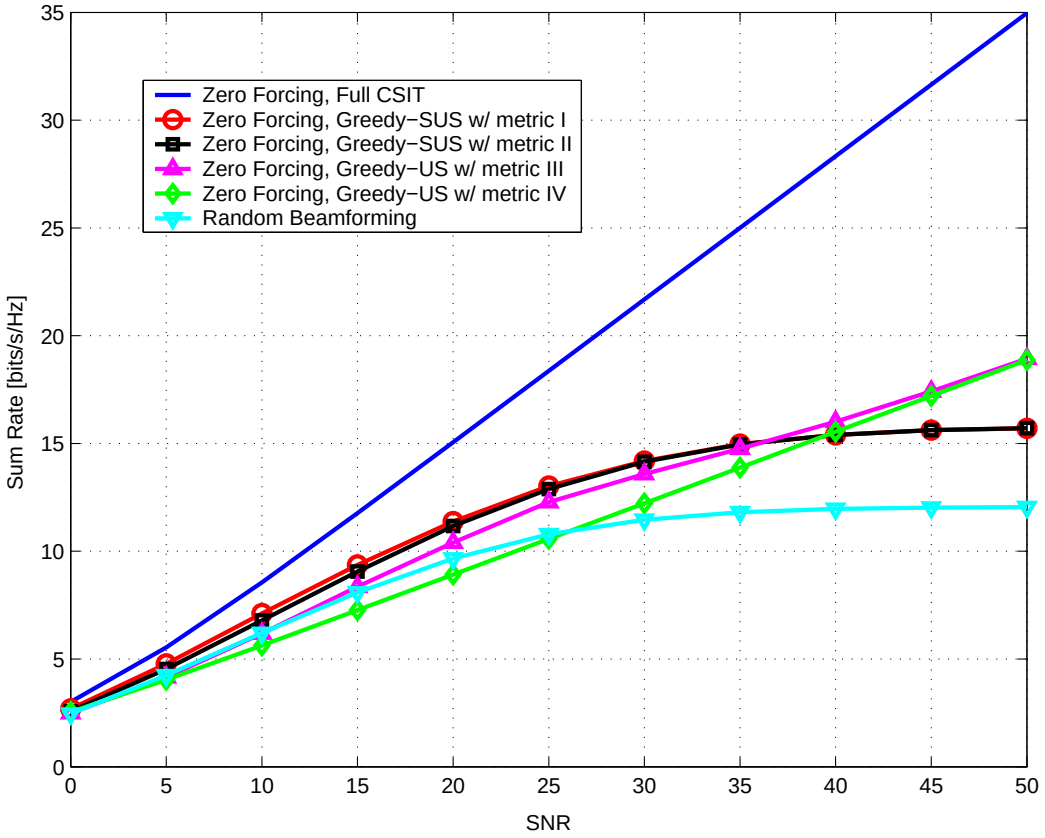

Fig. 1. Sum rate versus the average SNR for $B = 4$ bits, $M = 2$ transmit antennas and $K = 30$ users. Metric III combined with Greedy-US exhibits a linear sum-rate growth in the interference-limited region in contrast to Metric I and II, whose sum rate saturates at high SNR.

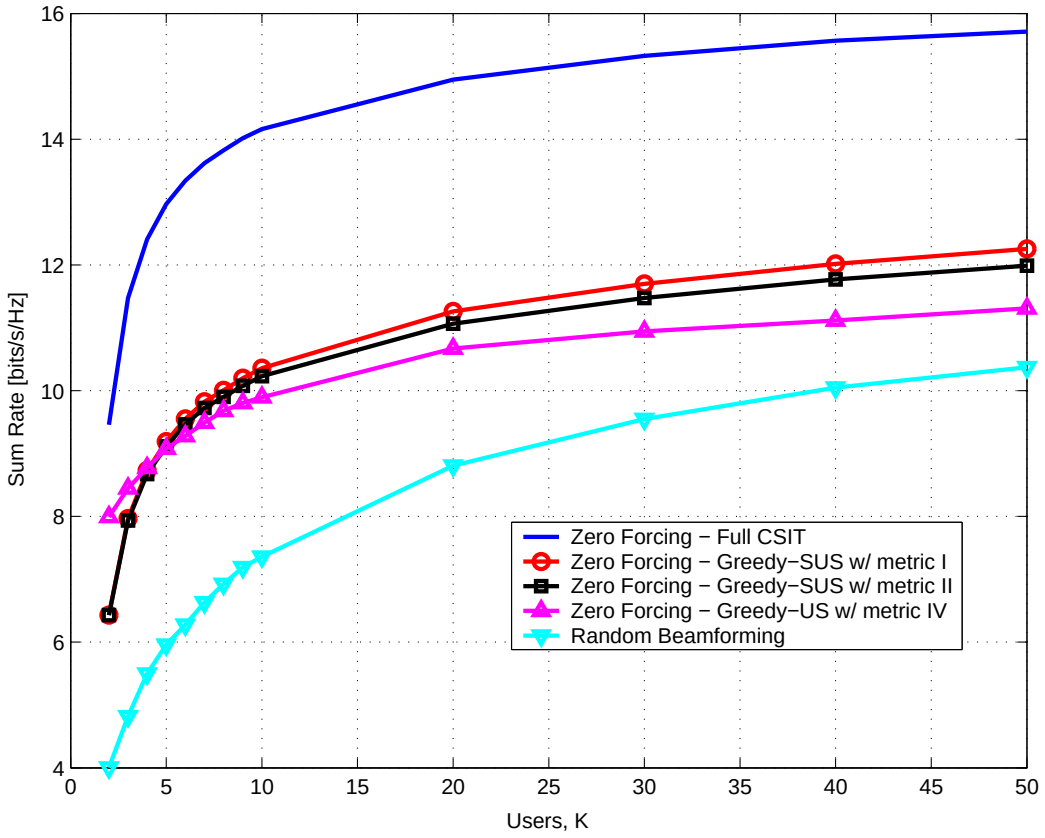

Fig. 2. Sum rate as a function of the number of users for $B = 4$ bits, $M = 2$ transmit antennas and $SNR = 20$ dB. All CQI scalar metrics can exploit the multiuser diversity gain, achieving optimal capacity growth.

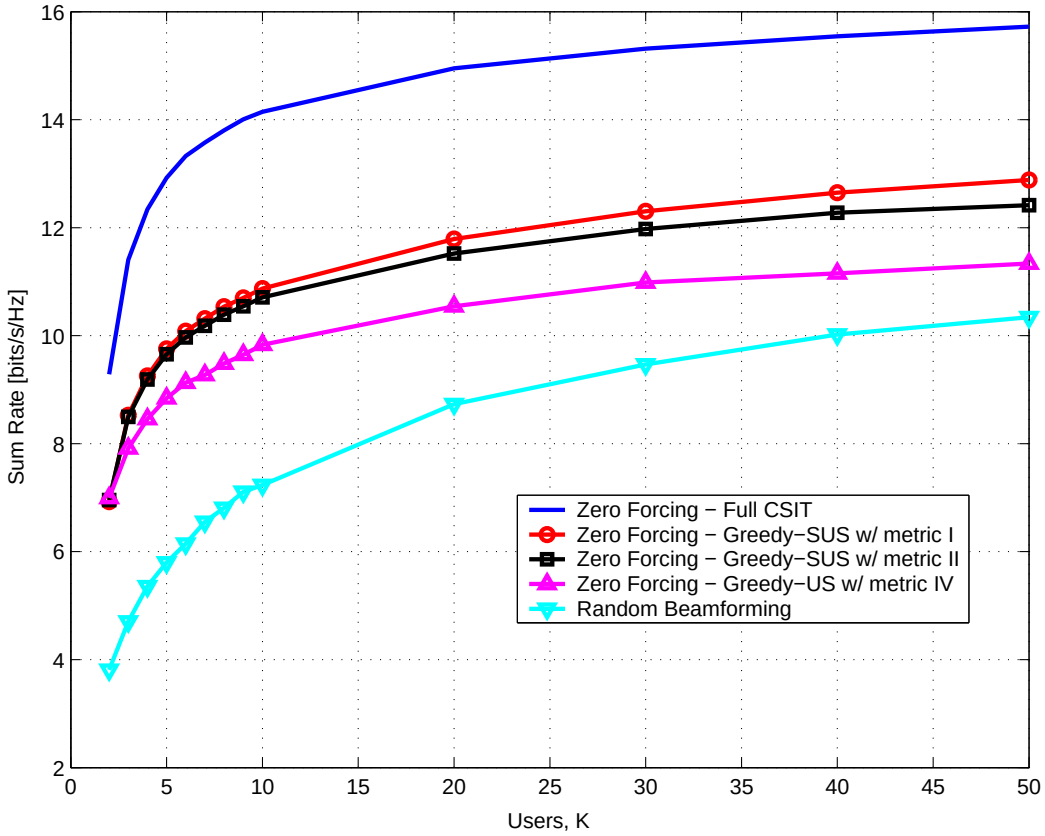

Fig. 3. Sum rate versus the number of users for different approaches with $SNR = 20$ dB, $M = 2$ transmit antennas and limited 10-bit total feedback bits where $B = 5$ bits are used for codebook indexing and $(10 - B$ bits) for CQI quantization. For metric IV, 2 bits are used for quantization of the channel norm and 3 bits for the alignment.