ON THE INTERPLAY OF ORTHOGONAL RANDOM BEAMFORMING AND CSI QUANTIZATION - STRATEGIES AND ROBUSTNESS ANALYSIS

Carles Antón-Haro

Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
Av. Canal Olímpic s/n. 08860 Castelldefels (Barcelona). SPAIN
carles.anton@cttc.es

ABSTRACT

In this paper, we assess the impact of CSI quantization on the performance of the Orthogonal Random Beamforming scheme. By resorting to a dynamic programming formulation, one can identify the *optimal* (i.e. pdf-matched) quantizer for which the sum-rate distortion is minimized. The robustness of such quantization scheme to uncertainty in the knowledge of specific system parameters along with the benefits resulting from inclusion of a second (and more refined) quantization step are analyzed as well. The performance, in terms of the resulting sum-rate, for this optimal quantizer is also assessed by means of computer simulations. Throughout this paper, a uniform quantizer is used as a benchmark.

Index Terms: multi-user diversity, scheduling, quantization, random beamforming.

1. INTRODUCTION

In a context of Multiple-Input Multiple-Output (MIMO) Broadcast Channels, Dirty Paper Coding (DPC) is known to be the capacity-achieving strategy [1]. However, DPC is computationally intensive and requires full channel state information at the transmitter (CSIT). In terms of sum-rate, however, there exist other precoding schemes such as Transmit Zero-Forcing (TxZF) beamforming [2] or Orthogonal Random Beamforming (ORB) [3] which asymptotically (in the number of active users) yield the same growth rate as DPC and, still, have lower computational complexity or merely require partial CSIT (i.e. SINR measurements). Multi-user diversity [4] can be efficiently exploited by the aforementioned transmit schemes but, to that aim, the centralized scheduler must be provided with at least partial CSI in order to make sure that the users experiencing the most favorable channel conditions are scheduled in each time instant.

In the literature, one can find a number of approaches aimed at minimizing the amount of CSI to be fedback by the (potentially high number of) terminals. In [5], for instance, the BS repeatedly polls the active users with a set of decreasing feedback thresholds; only those users which are above one of the thresholds (ideally a single user) convey their measured SINRs to the BS. The authors in [6] instead, propose a two-step feedback scheme: first, the scheduler decides on the active user subset on the basis of partial CSI (the SINRs resulting from a set of orthogonal random beams) which is conveyed by all the K users in the cell; next, full CSI is requested from the subset of scheduled users (M, with $M << K$) in order to construct a set of MMSE transmit beamformers. This scheme is particularly well-suited for sparse networks (i.e. with a moderate number of users) where ORB by itself would have difficulties in identifying a subset of quasi-orthogonal users. However, the above mentioned schemes assume that CSI is made available in *analog* form (i.e. infinite precision). In practice, CSI (either partial or full) must be quantized before being conveyed over a feedback channel. In this direction, [7] explores how to allocate L feedback bits (per user) in such a way that both multi-user diversity gains (resulting from roughly quantized channel quality information) and beamforming gains (resulting from the use of one beamvector out of those in a pre-designed codebook). In an ORB context, the limiting case of one-bit quantizers for the SINRs was addressed in e.g. [8], where the quantizer was designed in such a way that sum-rate of the quantized system retains the same growth rate as that of the *analog* one for a (very) large number of users. For a practical number of users, though, such design criterion may result in severe performance degradation.

Continuing our work in [9], we focus here on optimal (i.e. pdf-matched) quantization schemes with an arbitrary number of quantization levels (rather than only two as in [8]) for ORB. We complement [10] by addressing the multi-user (i.e. multi-beam) case instead of the single-user case and, unlike [10], we resort to numerical methods [11] which always converge to the optimal quantizer. As long as the pdf-matched quantizer depends a number of system parameters like the number of active users or the average SNR (which could be time-varying or be estimated with limited accuracy), we assess the robustness of the proposed quantization scheme to uncertainty in the knowledge of those parameters. In addition, we analyze the impact of a two-step quantization scheme by which the scheduling process is conducted on the basis of roughly quantized (1 bit) CSI, whereas the rate of the scheduled users is finely adjusted by using refined CSI information sent later by those users *only*.

2. SIGNAL MODEL

Consider the downlink of a wireless system with one Base Station (BS) equipped with M antennas, and K single-antenna terminals. In order to serve multiple users, we generate a pre-coding matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M]$[3], the columns of which, $\mathbf{w}_i \in \mathbb{C}^{M\times 1}$, $i = 1..M$, are isotropically-distributed random orthonormal vectors . Each of those vectors is then used to transmit data to the users experiencing the highest SINRs. The received signal at the k-th terminal when using beamformer i at the BS can be expressed as (time index has been dropped for the ease of notation):

$$r_{k,i} = \mathbf{h}_k^T \mathbf{w}_i s_i + \sum_{\substack{j=1 \\ j\neq i}}^{M} \mathbf{h}_k^T \mathbf{w}_j s_j + n_k \qquad (1)$$

where s_j stands for the symbol transmitted with beam j, $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ is the channel vector gain between the BS and the k-th terminal $\mathbf{h}_k \sim \mathcal{CN}(0, \mathbf{I}_M)$ (block Rayleigh fading, assumed to be i.i.d.

This work is partially supported by the IST projects COOPCOM (033533), ACE2 (26957) and NEWCOM++ (216715)

over users), and $n_k \in \mathbb{C}$ denotes additive Gaussian noise (AWGN) with zero mean and variance σ^2. Concerning CSI, we assume perfect knowledge at the terminals and the availability of a low-rate error- and delay-free feedback channel to convey partial CSI to the transmitter. Finally, the total transmit power, P_t, is constant and evenly distributed among the active beams, i.e., $\mathbb{E}\{\mathbf{s}^H\mathbf{s}\} = P_t$ and, hence, we can define $\rho = \frac{P_t}{\sigma^2}$ as the average SNR.

The last two terms in (1) clearly account for the interference-plus-noise contribution and, hence, the corresponding SINR measured at the terminal reads:

$$\gamma_{k,i} = \frac{|\mathbf{h}_k^T \mathbf{w}_i|^2}{M/\rho + \sum_{\substack{j=1 \\ j\neq i}}^{M} |\mathbf{h}_k^T \mathbf{w}_j|^2} = \frac{z}{M/\rho + y} \tag{2}$$

The scheduler in the BS operates in a slot-by-slot basis following a *max*-SINR (*greedy*) rule. That is, for beam i, the scheduler selects the active user k_i^* satisfying

$$k_i^* = \arg \max_{k=1..K} \{\gamma_{k,i}\} \qquad i = 1 \dots M$$

which experiences a *post-scheduling* SINR given by:

$$\gamma_i^* = \max_{k=1..K} \{\gamma_{k,i}\} \qquad i = 1 \dots M$$

As shown in [3], the pdf function of such random variable reads

$$f_{\mathsf{SINR}^*}(\gamma) = K \frac{e^{-\frac{\gamma M}{\rho}}}{(1+\gamma)^M} \quad \frac{M}{\rho}(1+\gamma) + M - 1 \times \quad 1 - \frac{e^{-\frac{\gamma M}{\rho}}}{(1+\gamma)^{M-1}} \Bigg)^{K-1}. \tag{3}$$

Finally, one can readily express the sum-rate R in terms of the pdf above as:

$$R \approx \mathbb{E}_{\gamma^*} \left[\sum_{i=1}^{M} \log_2 \quad 1 + \max_{1\leq k\leq K} \gamma_{k,i} \right] = M \int_0^\infty \log_2 (1+\gamma)\, f_{\mathsf{SINR}^*}(\gamma) d\gamma \tag{4}$$

In a realistic scenario, the BS is constrained to schedule users on the basis of a *quantized* version of the measured SINRs, $Q(\gamma_{k,i})$, rather than with the *analog* SINRs in the expressions above. This issue will be addressed in the next section where we introduce the *optimal* quantization strategy.

3. REVIEW OF OPTIMAL QUANTIZATION STRATEGIES

Let $\Gamma_d = \{\gamma_{d_0} < \gamma_{d_1} < \ldots < \gamma_{d_{N_q}}\}$ be the input decision levels and let $\Gamma_q = \{\gamma_{q_0} < \gamma_{q_1} < \ldots < \gamma_{q_{N_q-1}}\}$ be the output representative levels of an $N_q = 2^{L_q}$-level quantizer $Q(\cdot)$ defined as:

$$Q(\gamma) = \gamma_{q_j} \qquad \text{if } \gamma_{d_j} \leq \gamma < \gamma_{d_{j+1}}. \tag{5}$$

Hence, the (quantized) *post-scheduling* SINR on beam i reads :

$$\max_{k=1..K} \{Q(\gamma_{k,i})\} \qquad i = 1 \dots M$$

or, equivalently, by exchanging the max and Q operators[1]

$$Q \quad \max_{k=1..K} \{\gamma_{k,i}\} \quad = Q(\gamma_i^*) \qquad i = 1 \dots M$$

[1] Although the resulting quantized SINR is identical, a different selection of users may result from this change.

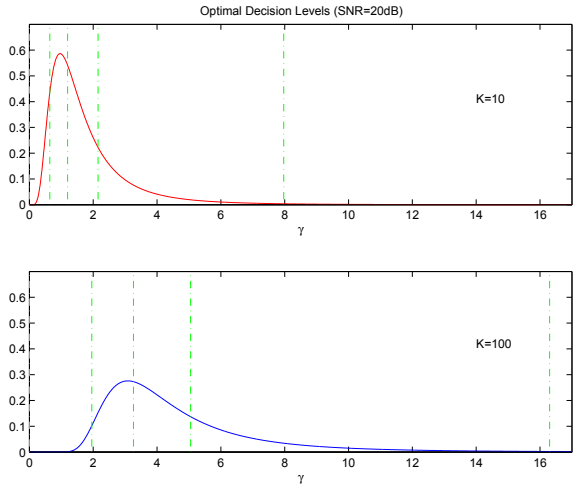

Fig. 1. Optimal decision levels of the *post-scheduling* SINR.

Thus, the problem to solve is that of identifying an optimal (i.e. pdf-matched) set of decision and representative levels $\{\Gamma_d^*, \Gamma_q^*\}$, such that the average distortion introduced by the N_q-level quantizer:

$$D_{N_q} = \mathbb{E}_{\gamma^*}\left[e\left(\gamma^*, Q(\gamma^*)\right)\right] = \sum_{i=0}^{N-1} \int_{\gamma_{d_i}}^{\gamma_{d_{i+1}}} e\left(\gamma, \gamma_{q_i}\right) f_{\mathsf{SINR}^*}(\gamma) d\gamma \tag{6}$$

is minimized, with $e(\cdot,\cdot)$ standing for an error weighting function of choice. Since the optimally-quantized SINRs should minimize the *sum-rate* distortion (4), we define the error function as

$$e(\gamma, \gamma_{q_i}) = \log_2(1+\gamma) - \log_2(1+\gamma_{q_i}) = \log_2 \quad \frac{1+\gamma}{1+\gamma_{q_i}}$$

In general, this problem cannot be solved analytically. Alternatively, one can resort to a dynamic program formulation [11] and obtain a numerical solution, which can be proved to be the global optimum[11]. In this context, we define two functions $D_1(\alpha,\beta)$ and $D_n(\alpha,\beta)$ in the following way: first, let $D_1(\alpha,\beta)$ denote the minimum value of the distortion measure when placing just one output level in the range (α,β), a subrange of $(\gamma_{d_0}, \gamma_{d_{N_q}})$:,

$$D_1(\alpha,\beta) \triangleq \min_y \int_\alpha^\beta e\left(\gamma, y\right) f_{\mathsf{SINR}^*}(\gamma) d\gamma. \tag{7}$$

Next, $D_n(\alpha,\beta)$ is defined as the minimum distortion when n levels are place in (α,β), for $n \geq 2$. Such distortion term can be conveniently expressed in terms of (7) as follows:

$$D_n(\alpha,\beta) = \min_{\substack{\gamma_{d_1},\ldots,\gamma_{d_{n-1}} \\ (\alpha<\gamma_{d_1}\ldots<\gamma_{d_{n-1}}<\beta)}} \sum_{i=0}^{n-1} D_1(\gamma_{d_i}, \gamma_{d_{i+1}}) \tag{8}$$

where $\gamma_{d_0} = \alpha$ and $\gamma_{d_n} = \beta$. Notice that γ_{d_i} and γ_{q_i} denote the interim search variables whereas their optimal counterparts will be labeled with the subscript *. The search algorithm consists of the following five steps:

1. **Initialization**: Compute and store the values of $D_1(\alpha,\beta)$ for all discrete α and β in $(\gamma_{d_0}, \gamma_{d_{N_q}})$. To do so, we assume N_q,γ_{d_0} and $\gamma_{d_{N_q}}$ to be set in advance.

2. **Insertion of decision levels**: For each n from two to N_q and all discrete γ in $(\gamma_{d_0}, \gamma_{d_{N_q}})$, compute and store both

$$D_n(\gamma_{d_0}, \gamma) = \min_{\gamma_{d_0} < \alpha < \gamma} [D_{n-1}(\gamma_{d_0}, \alpha) + D_1(\alpha, \gamma)]$$

and $\gamma_{d_n}(\gamma_{d_0}, \gamma)$, which denotes the optimum value of α for which $D_n(\gamma_{d_0}, \gamma)$ is minimized. By doing so, we identify the best point to insert an additional decision level in (γ_{d_0}, γ).

3. **Computation of the optimal decision levels**: For each n from N_q to two, set

$$\gamma^*_{d_{n-1}} = \gamma_{d_{n-1}}(\gamma^*_{d_0}, \gamma^*_{d_n}) \tag{9}$$

with $\gamma^*_{d_{N_q}} = \gamma_{d_{N_q}}$ and $\gamma^*_{d_0} = \gamma_{d_0}$.

4. **Computation of the optimal representative levels**: In general, one should compute for each n from zero to $N_q - 1$, the optimal $\gamma^*_{q_n}$ for which $D_1(\gamma^*_{d_n}, \gamma^*_{d_{n+1}})$ is the minimum value as given in (7). However, we impose here $\gamma_{q_i} = \gamma_{d_i}; i = 0 \ldots N_q - 1$. In this way, we ensure that the quantized SINRs (the ones used to adjust the constellation size and the coding scheme at the transmitter) are never above the actual SINR value. Otherwise, the estimated data rate would exceed that which can be reliably supported and an outage would result.

5. **End of algorithm**.

In Fig.1, we show the optimal decision levels associated to the *post-scheduling* pdf in Eq.3. Qualitatively, the optimal decision thresholds should be placed where the pdf takes non-zero values. For a growing number of users, this results in an overall shift towards high SINR values, which are more likely to occur.To conclude this section, notice that we are facing a cross-layer design in the sense that the optimal quantizer in the *physical* layer is tightly coupled with *system-level* parameters (introduced through f_{SINR^*} in Eq.6), such as the number of admitted users (K) or the number of antennas in the base station. Such system-level parameters should then be broadcasted to the user terminals in the cell through common signalling channels. As for the optimal decision levels, they should be precomputed and stored (for a limited set of input parameters) at the terminal in order to avoid costly on-line computations.

3.1. A two-step quantization strategy

The overall performance could be improved if we introduce a refinement in the quantization step. In particular, we propose to improve the process by asking the scheduled users *only* to subsequently provide the BS with a finely quantized version of the measured SINR. In this way, we restrict the impact of the *roughly* quantized CSI to the scheduling process, whereas the actual data rate in the downlink can be more accurately determined on the basis such *finely* quantized SINRs. In settings where only few terminals out of a relatively large population of users are selected for transmission, the increase in terms of signalling resulting from this dedicated feedback is potentially low (although this refinement would also imply some additional delay). The impact of this refinement will be assessed next.

4. COMPUTER SIMULATION RESULTS

We consider a system with $K = 1 \ldots 1000$ active users and one BS with $M = 4$ antennas. In Fig.2, we assess the performance of the optimal quantizer and compare it with that of (1) a uniform quantizer, and (2) a system where SINRs have *analog* precision (i.e. lower and upper bounds). First, we can observe that the performance exhibited by the optimal quantizer with $L_q = 3$ bits is very close to that of the analog system: $\frac{6.03}{6.71} = 90\%$ and $\frac{12.92}{13.61} = 95\%$ for the $K = 20$ and $K = 1000$ cases, respectively . With $L_q = 1$ quantization bits, the system still manages to retain up to $\frac{3.84}{6.74} = 57\%$ of the analog performance for $K = 20$ (and $\frac{10.19}{13.60} = 75\%$ for K=1000). In summary, the optimal quantizer adapts quite well to the different pdf shapes resulting from different values of K and ρ.

Besides, the pdf-matched quantizer clearly outperforms its uniform counterpart in all cases and scenarios. The performance loss of the uniform quantizer is more severe in the $L_q = 1$ bit case: up to $\frac{3.84-0.71}{3.84} = 81.5\%$ and $\frac{10.19-1.89}{10.19} = 81.4\%$ w.r.t. the optimal quantizer for the $K = 20$ and $K = 1000$ cases, respectively. The fact that performance relies on the value that the *single* decision level takes, makes it very sensitive to non-optimal designs. This can be observed in Fig.3 where we depict both the optimal and uniform decision levels for a varying number of active users. In the $L_q = 3$ case, there is a substantial overlap between both sets of decision thresholds and, hence, the uniform quantizer performs reasonably well. On the contrary, a severe performance loss results from the gap between the optimal and uniform decision thresholds for the $L_q = 1$ case. Still, the overlap between the optimal and uniform de-

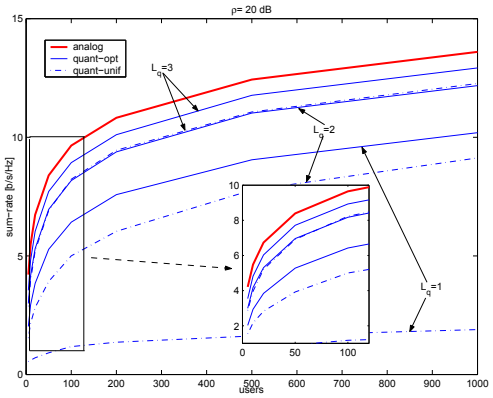

Fig. 2. Sum-rate vs. number of users with analog and quantized CSI. Interference-limited scenario (ρ=20 dB).

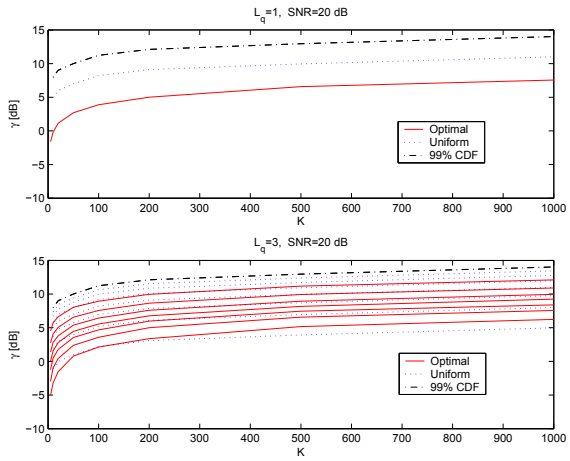

Fig. 3. Optimal and uniform decision thresholds vs. the number of active users. Dash-dotted curve: 99% percentile of dynamic range.

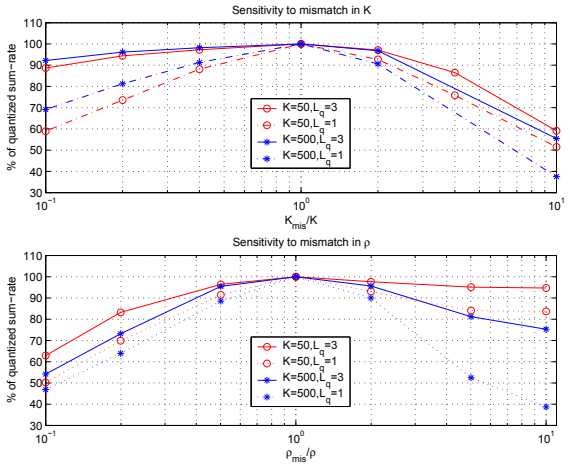

Fig. 4. Sensitivity of the optimal quantizer to parameter mismatch (ρ=10 dB). Results are normalized in both axes.

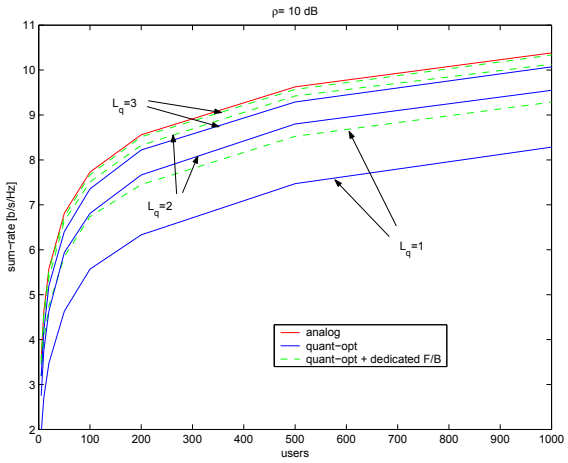

Fig. 5. Sum-rate vs. number of users with analog, quantized CSI and quantized information plus dedicated feedback.

cision levels strongly depends on parameters such as ρ or K and in some cases (e.g. $\rho = 0$, $K = 1000$) the gap between both curves is relatively small. However, only the optimal quantizer can guarantee close-to-analog performance in a general case. Next, in Fig. 4 (top) we illustrate the impact of a mismatch in the number of active users on the resulting sum-rate of the optimal quantizer. Clearly, the lower the number of decision levels, the faster performance degrades with an increasing mismatch (with only one threshold to rely on, any shift from the optimal value is critical). Besides, one can also observe that performance loss is negligible, less than 10%, when the algorithm is parameterized by K_{mis} within 0.5..2 times the number of actual users K. In other words, the quantization scheme is quite robust and, hence, there is no need for the BS to continuously broadcast updates in system-level parameters. As for the sensitivity to imperfect estimates of ρ, Fig. 4 (bottom) reveals that the quantization scheme is quite robust too: less than 10% performance loss within ± 3 dB. Interestingly enough, for higher values of ρ_{mis}/ρ performance still degrades quite gracefully. The explanation for this behavior can be found in equation (3): for a given number of active users increasing ρ_{mis} results into a wider pdf function but, still, the central parts (where the decision intervals lie) of the pdfs parameterized by ρ and ρ_{mis} notably overlap with each other. Conversely, by substantially increasing K (for a given ρ) we force the pdf curves to shift their central (i.e. non-zero values) far away from each other, this resulting into substantial performance losses in the high K_{mis}/K region. Finally, in Fig. 5, we compare the performance of the schemes with and without dedicated feedback for the scheduled users only. For the latter case, we assume that a sufficient number of bits to have close-to-analog quality are used in the refined quantization step (typically $L_q = 4$ bits suffice). Clearly, a substantial improvement results when in the rough quantization step we use 1 or 2 quantization bits ($\frac{9.29}{8.23} = 113\%$ and $\frac{10.1}{9.6} = 106\%$, respectively). However, the gain is much more moderate when $L_q = 3$ since, after all, both curves are already very close to that of the analog system.

5. CONCLUSIONS

In this paper, we have analyzed the impact of CSI quantization on the performance of orthogonal random beamforming. With as few as 3 or 4 bits, we have found that the optimal quantizer attains a sum-rate virtually identical to that of the analog system. With one quantization bit the optimal quantizer still retains on the order of 75% of the analog sum-rate. In those conditions, however, the uniform quantizer often suffers from substantial performance losses. In general, the optimal quantizer was shown to be quite robust to mismatches in the number of active users and the average SNR, with 90% of the ideal performance retained when parameterizing the quantization algorithm with values ± 3 dB the actual ones. Finally, we have found that by further refining the (optimally) quantized SINRs for the scheduled users only, the sum-rate performance can be substantially improved, in particular when very roughly quantized information was used during the scheduling stage.

6. REFERENCES

[1] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian MIMO broadcast channel," *Proc. CISS, Princeton*, 2004.

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

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

[4] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," *in Proc. IEEE ICC*, 1995.

[5] V. Hassel, D. Gesbert, M. Slim-Alouini, and G. Oien, "A threshold based channel state feedback algorithm for modern cellular systems," *IEEE Trans. on Wireless Comms.*, pp. –, 2007.

[6] M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in *Proc. IEEE Workshop on Signal Proc. Advances in Wireless Comm. (SPAWC)*, June 2005, pp. 975–979.

[7] S. Sanayei and A. Nosratinia, "Opportunistic beamforming with limited feedback," *IEEE Trans. on Wireless Communications*, vol. 6, no. 8, pp. 2765–2771, Aug. 2007.

[8] J. Diaz, O. Simeone, Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," in *In Proc. Int'l Symposium on Information Theory (ISIT), Seattle (WA)*, June 2006.

[9] C. Antón-Haro, "Optimal quantization schemes for orthogonal random beamforming - a cross-layer approach," in *In Proc. IEEE Int'l. Conf. on Acoustics, Speech and Signal Proc.*, vol. 3, April 2007, pp. 1234–1237.

[10] M. T. O. Ozdemir, "Performance of opportunistic beamforming with quantized feedback," in *In Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, Nov. 2006, pp. 1–5.

[11] D.K. Sharma, "Design of absolutely optimal quantizers for a wide class of distortion measures," *IEEE Trans. on Information Theory*, vol. 24, no. 6, pp. 693–702, Nov. 1978.