QUANTIZED FEEDBACK DESIGN FOR MIMO BROADCAST CHANNELS

Tùng T. Kim, Mats Bengtsson, and Mikael Skoglund

School of Electrical Engineering
Royal Institute of Technology, SE-10044 Stockholm, Sweden
e-mail: {tung.kim, mats.bengtsson, mikael.skoglund}@ee.kth.se

ABSTRACT

Low-rate feedback design for multiple-input multiple-output broadcast channels is studied under a vector quantization framework. Iterative algorithms are proposed to design the partial feedback link, the scheduler, and the linear precoding codebook. It is demonstrated that the gain due to multi-user diversity can be significant even with heavily quantized channel state information at the transmitter. Our results highlight the potential of multi-user diversity, even with simple schemes and extremely-low-rate feedback.

Index Terms— Fading channels, quantization, broadcast channels, feedback communications, information rates.

1. INTRODUCTION

Recent advances show that with perfect channel state information at the transmitter (CSIT), dirty-paper coding achieves the whole capacity region of the multiple-input multiple-output (MIMO) broadcast channels [1–4]. Such a non-linear coding approach is however complicated and furthermore, perfect CSIT is generally very difficult to obtain in practice. This motivates various work on broadcast channels with partial CSIT, for example simple yet efficient opportunistic approaches are proposed and analyzed in [5, 6].

Opportunistic communication schemes normally rely on feedback of the signal-to-noise ratio (SNR) or the signal-to-interference-noise ratio (SINR), implicitly requiring a possibly large amount of feedback. More explicit quantized feedback models are considered in e.g., [7]. Most previous work focused on the multiple-transmit single-receive antenna case, or on the asymptotic regime of a very large number of users.

In this work, we use tools in vector quantization to jointly design the partial feedback link, the scheduler and the precoding matrices in a broadcast channel. While vector quantization is widely used in the feedback design for the single-user case [8, 9], extending it to multi-user scenarios proved to be difficult due to its exponentially complexity as the number of users increases. We therefore restrict our attention to a class of schemes with *linear* precoding and *single-user* scheduling using heavily quantized CSIT. Our results are not asymptotic, and the presented design technique can be applied to any channel distribution.

The results demonstrate that significant gain compared to the single-user case can be achieved even with heavily quantized CSIT. The baseline scheme however requires exponential complexity in the number of users. We thus propose a reduced-complexity scheme, which only requires a codebook size that scales linearly as the number of users increases. Our results highlight the potential of multi-user diversity, even with simple schemes and extremely low feedback rate.

2. SYSTEM MODEL

Consider the discrete-time complex-baseband model of a multiple antenna broadcast channel with M users. The transmitter has N_{t} antennas. For simplicity of notation, assume that all users have the same number of receive antennas, N_{r}. Assume perfect channel state information at all receivers. Given a channel realization $\mathbf{H}_k$, user k employs an *index mapping* $\mathcal{I}_k(\mathbf{H}_k)$ from channel matrix to *feedback index*. Assume $i_k = \mathcal{I}_k(\mathbf{H}_k)$ takes values on $\{1, \ldots, K\}$ where K is a fixed integer, referred to as the *resolution* of the feedback link. All users have the same feedback resolution K. All feedback indices from the users are sent back to the transmitter via noiseless, zero-delay dedicated feedback links.

Upon receiving the set of indices $i_1^M \triangleq \{i_1, \ldots, i_M\}$, the transmitter employs a *scheduler* $\mathcal{J}(i_1^M)$, i.e., a mapping from the tuple i_1^M to an integer in $\{1, \ldots, M\}$ indicating which user is served. We exclusively focus on single-user scheduling. That is, at any channel use, only a single user is served. The symbols of user k, $\mathbf{s}_k$, are drawn from a "Gaussian codebook," with $\mathrm{E}[\mathbf{s}_k\mathbf{s}_k^{\mathrm{H}}] = \mathbf{I}_{N_{\mathrm{t}}}$ where $\mathbf{I}_{N_{\mathrm{t}}}$ is the identity matrix of size N_{t}, and then multiplied by a *precoding matrix* $\mathbf{W}_{i_1^M}$, which is only influenced by the indices. We refer to the set of all possible $\mathbf{W}_{i_1^M}$'s as the *precoding codebook*. In total, there are K^M such precoding matrices. Herein $[\cdot]^{\mathrm{H}}$ denotes complex conjugate and transpose.

The received signal of user k at time instant t, $t = 0, 1, \ldots$, can be written as

$$\mathbf{y}_k(t) = \mathbf{H}_k(t)\mathbf{W}_{i_1^M}\mathbf{s}_{\mathcal{J}\left(i_1^M\right)}(t) + \mathbf{n}(t), \qquad (1)$$

where the components of the noise vector $\mathbf{n}(t)$ are spatially

and temporally white complex Gaussian with zero mean and unit variance. A short-term power constraint is considered:

$$\operatorname{tr}\left[\mathbf{W}_{i_1^M}\mathbf{W}_{i_1^M}^{\mathrm{H}}\right] = \operatorname{tr}\mathbf{Q}_{i_1^M} \le P,\ \forall i_1^M \in \{1,\dots,K\}^M, \quad (2)$$

where $\operatorname{tr}[\cdot]$ denotes the trace of a matrix. That is, no temporal power control is employed.

We assume a memoryless ergodic model, i.e, for each user k, the corresponding channel $\mathbf{H}_k(t)$ changes independently from one time instant t to another. The channel matrices of different users are independent, but may follow different distributions. We are interested in optimizing the sum rate of the described scheme, with respect to the scheduler, the precoding codebook, and the index mappings:

$$\max_{\mathcal{J}\left(i_1^M\right),\{\mathbf{Q}_l\},\{\mathcal{I}_k(\mathbf{H}_k)\}} \mathrm{E}\log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_{\mathcal{J}\left(i_1^M\right)}\mathbf{Q}_l\mathbf{H}_{\mathcal{J}\left(i_1^M\right)}^{\mathrm{H}}\right), \quad (3)$$

where the expectation is over the randomness of the channels.

We do not impose any explicit constraint on the fairness of the system. It is however possible to address the fairness issue *indirectly* by replacing the sum rate criterion (3) with a *weighted* sum rate, which can be handled in a similar procedure to that presented in Sections 3 and 4. In practice, the weights can be adjusted to meet some fairness requirements. Such weighting parameters can be considered to be slowly varying, requiring a much smaller feedback rate compared to that required to quantize the short-term channel state.

3. JOINT SCHEDULING AND PRECODING DESIGN: A HIGH-COMPLEXITY APPROACH

We use the generalized Lloyd algorithm [8, 9] to design the scheduler, precoders and index mappers. More precisely, we first *approximate* the true joint distribution of the channel matrices by a sample (joint) distribution. That is $\mathbf{H}_k$'s are drawn from a set $\mathcal{H}_k$ with *finite* cardinality $|\mathcal{H}_k|$. Accordingly, the sum rate criterion in (3) is approximated as

$$\frac{1}{\prod_{k=1}^M |\mathcal{H}_k|} \times \sum_{\mathbf{H}_1\in\mathcal{H}_1}\cdots\sum_{\mathbf{H}_M\in\mathcal{H}_M} \log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_{\mathcal{J}(\mathcal{I}_1^M)}\mathbf{Q}_{\mathcal{I}_1^M}\mathbf{H}_{\mathcal{J}(\mathcal{I}_1^M)}^{\mathrm{H}}\right).$$

We then iteratively optimize the index mappings, the scheduler, and the precoding codebooks in a two-step procedure, described as follows. For convenience, define the *quantization region* i of user k as

$$\mathcal{R}_k^i \triangleq \{\mathbf{H}_k : \mathcal{I}_k(\mathbf{H}_k) = i\}.$$

Step 1: Fix the scheduler $\mathcal{J}\left(i_1^M\right)$, the codebooks $\mathbf{Q}_{i_1^M}$, and all other index mappings $\mathcal{I}_l(\mathbf{H}_l)$, $l \ne k$, find the optimal index mapping $\mathcal{I}_k(\mathbf{H}_k)$ for $k = 1,\dots,M$.

For clarity we present the solution of the optimization for $k = 1$. The index mappings of other users are optimized in a completely similar manner.

$$\begin{aligned}\mathcal{I}_1(\mathbf{H}_1) = &\max_{i_1\in\{1,\dots,K\}} \sum_{\mathbf{H}_2\in\mathcal{H}_2}\cdots\sum_{\mathbf{H}_M\in\mathcal{H}_M} \\ &\log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_{\mathcal{J}(i_1,\mathcal{I}_l(\mathbf{H}_l))}\mathbf{Q}_{i_1,\mathcal{I}_l(\mathbf{H}_l)}\mathbf{H}_{\mathcal{J}(i_1,\mathcal{I}_l(\mathbf{H}_l))}^{\mathrm{H}}\right) \\ = &\max_{i_1}\sum_{i_2=1}^K\cdots\sum_{i_M=1}^K \sum_{\mathbf{H}_2\in\mathcal{R}_2^{i_2}}\cdots\sum_{\mathbf{H}_2\in\mathcal{R}_M^{i_M}} \\ &\log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_{\mathcal{J}\left(i_1^M\right)}\mathbf{Q}_{i_1^M}\mathbf{H}_{\mathcal{J}\left(i_1^M\right)}^{\mathrm{H}}\right).\end{aligned}$$

Denoting

$$\begin{aligned}\bar{I}\left(i_1^M\right) \triangleq &\frac{1}{\left|\mathcal{R}_{\mathcal{J}\left(i_1^M\right)}^{i_{\mathcal{J}\left(i_1^M\right)}}\right|} \\ &\times \sum_{\mathbf{H}_{\mathcal{J}}\in\mathcal{R}_{\mathcal{J}}^{i_{\mathcal{J}}}} \log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_{\mathcal{J}\left(i_1^M\right)}\mathbf{Q}_{i_1^M}\mathbf{H}_{\mathcal{J}\left(i_1^M\right)}^{\mathrm{H}}\right),\end{aligned}$$

we finally obtain

$$\mathcal{I}_1(\mathbf{H}_1) = \max_{i_1\in\{1,\dots,K\}}\sum_{i_2}\cdots\sum_{i_M}\left(\prod_{l=2}^M |\mathcal{R}_l^{i_l}|\right) I\left(i_1^M\right) \quad (4)$$

where

$$I\left(i_1^M\right) = \begin{cases}\log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_1\mathbf{Q}_{i_1^M}\mathbf{H}_1^{\mathrm{H}}\right) & \text{if } \mathcal{J}\left(i_1^M\right) = 1, \\ \bar{I}\left(i_1^M\right) & \text{otherwise.}\end{cases}$$

Step 2: Fix index mappings (thus all the quantization regions), find the optimal codebooks $\{\mathbf{Q}_{i_1^M}\}$ and the optimal scheduler $\mathcal{J}\left(i_1^M\right)$.

To that end, for each $k \in \{1,\dots,M\}$, let

$$\begin{aligned}I_k^* = \max_{\mathbf{Q}\succeq 0} &\frac{1}{|\mathcal{R}_k^{i_k}|}\sum_{\mathbf{H}_k\in\mathcal{R}_k^{i_k}} \log\det\left(\mathbf{I}_{N_{\mathrm{r}}} + \mathbf{H}_k\mathbf{Q}\mathbf{H}_k^{\mathrm{H}}\right) \\ &\text{s.t. } \operatorname{tr}\mathbf{Q} \le P,\end{aligned}$$

which is the maximum expected achievable rate (conditioned on i_1^M) if user k is scheduled. This optimization problem is convex and thus its global optimum can be found with standard methods [10]. Let $\mathbf{Q}_k^*$ be the corresponding optimal transmit covariance matrix. Note that only a single user is served at a time, hence the optimal scheduler is given by

$$\mathcal{J}\left(i_1^M\right) = \arg\max_{k\in\{1,\dots,M\}} I_k^*, \quad (5)$$

and the optimal transmit covariance matrix corresponding to the tuple i_1^M is therefore given by

$$\mathbf{Q}_{i_1^M} = \mathbf{Q}_{\mathcal{J}\left(i_1^M\right)}^*. \quad (6)$$

The proposed iterative procedure guarantees convergence, but not necessarily to the global optimum due to the non-convexity of (3). We use a large number of randomly generated starting points to limit the effects of local convergence.

For implementation, since the values $\bar{I}\left(i_1^M\right)$ are not dependent on the channel realization, they can be, at least in principle, computed off-line and stored at the receivers. The robustness of the described scheme to mismatches in the assumed channel distributions and system configuration remains to be investigated.

4. JOINT SCHEDULING AND PRECODING DESIGN: A REDUCED-COMPLEXITY APPROACH

While the performance of the scheme described in Section 3 serves as a benchmark, such a scheme is too computationally demanding as the codebook size grows exponentially with the number of users. In this section we present a simpler alternative. The output of the scheduler and the index of the *scheduled user* are used to choose the precoding matrix. That is, given $k = \mathcal{J}\left(i_1^M\right)$, user k is selected and the precoding matrix in this case is denoted as $\mathbf{W}_{i_k}^k$. The codebook size is KM, i.e., linearly scaled with the number of users. Recall that there are K^M possible tuples i_1^M, meaning that some different tuples are mapped to the same precoding matrix. The design of the proposed scheme is also based on the idea of the Lloyd algorithm, and is divided into three steps, which are iterated until (possibly local) convergence.

Step 1: Fix the scheduler, codebooks, and all index mappings $\mathcal{I}_l(\mathbf{H}_l), l \neq k$, optimize $\mathcal{I}_k(\mathbf{H}_k)$. We again only present the optimization of $\mathcal{I}_1(\mathbf{H}_1)$. It can be verified that (4) still holds in this case. A subtle, but extremely important difference is that user 1 now needs only to compute K terms of the form $\log\det\left(\mathbf{I}_{N_\mathrm{r}} + \mathbf{H}_1\mathbf{Q}_{i_1}^1\mathbf{H}_1^{\mathrm{H}}\right)$, one for each possible value of i_1. In contrast, in Section 3, a user on the average has to compute $\frac{K^M}{M}$ terms of the form $\log\det\left(\mathbf{I}_{N_\mathrm{r}} + \mathbf{H}_1\mathbf{Q}_{i_1^M}\mathbf{H}_1^{\mathrm{H}}\right)$, which is not feasible for large numbers of users.

Step 2: Fix $\mathbf{Q}_{i_k}^k$, $k = 1, \ldots, M$ and all index mappings, optimize $\mathcal{J}\left(i_1^M\right)$. We readily have

$$\mathcal{J}\left(i_1^M\right) = \arg\max_{k\in\{1,\ldots,M\}} \frac{1}{\left|\mathcal{R}_k^{i_k}\right|}\sum_{\mathbf{H}_k\in\mathcal{R}_k^{i_k}} \log\det\left(\mathbf{I}_{N_\mathrm{r}} + \mathbf{H}_k\mathbf{Q}_{i_k}^k\mathbf{H}_k^{\mathrm{H}}\right). \quad (7)$$

Step 3: Fix $\mathcal{J}\left(i_1^M\right)$ and all the index mappings, optimize the codebook $\{\mathbf{Q}_{i_k}^k\}$. The optimization of $\mathbf{Q}_j^k$ is somewhat similar to the one in Section 3, but we have to consider all regions with $i_k = j$ and $\mathcal{J}(i_1^M) = k$ jointly. It can be shown that

$$\mathbf{Q}_j^k = \arg\max_{\mathbf{Q}\succeq 0, \mathrm{tr}\,\mathbf{Q}\leq P} \sum_{i_1^M, \mathcal{J}\left(i_1^M\right)=k, i_k=j} \left(\prod_{l=2, l\neq k}^{M}\left|\mathcal{R}_l^{i_l}\right|\right) \sum_{\mathbf{H}_k\in\mathcal{R}_k^{i_k}} \log\det\left(\mathbf{I}_{N_\mathrm{r}} + \mathbf{H}_k\mathbf{Q}\mathbf{H}_k^{\mathrm{H}}\right). \quad (8)$$

This is also a convex optimization problem, to which the optimal solution can be found efficiently [10].

The steps described above can be modified to obtain an even simpler scheme, which only uses limited feedback to *schedule* a user. In this case, given $k = \mathcal{J}\left(i_1^M\right)$, user k is selected and the precoding matrix is *always* $\mathbf{W}_k$. The codebook size is in this case *equals* to the number of users M.

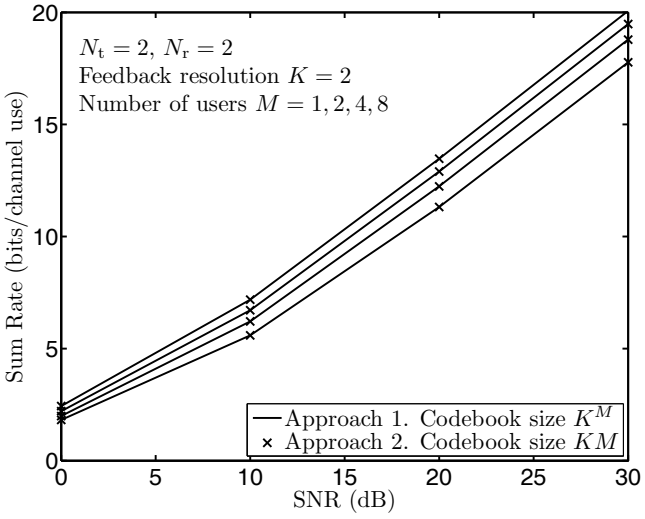

Fig. 1. Sum rate with different numbers of users M over 2×2 channels.

5. SIMULATION RESULTS

Figure 1 plots the sum rate achieved with the schemes in Sections 3 and 4 over 2×2 uncorrelated Rayleigh channels where the components of the channel matrices are i.i.d. complex Gaussian with zero mean and unit variance. Since the noise has unit variance, the SNR is defined as SNR $\triangleq P$. The feedback resolution is fixed at $K = 2$ (1 bit of feedback). The performance of the high-complexity approach ('Approach 1') and that of the reduced-complexity one ('Approach 2') are indistinguishable. This may be attributed to our assumption on the mutual independence of the users' channel matrices. If the users' channel matrices are correlated, the performance of the schemes may differ. As can be seen, increasing the number of users provides a significant power gain due to selection diversity. For example, at a sum rate of 10 bits per channel use,

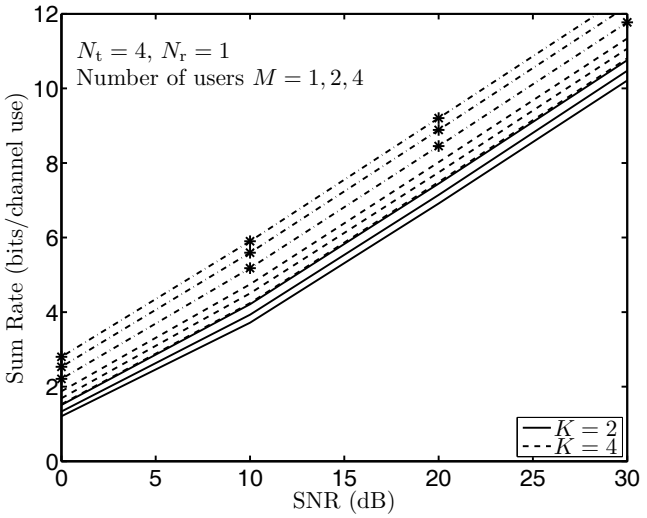

Fig. 2. Sum rates with different numbers of users M and feedback resolution K over 4×1 channels. The performance of a single-user scheduled system with perfect CSIT is marked with '*.'

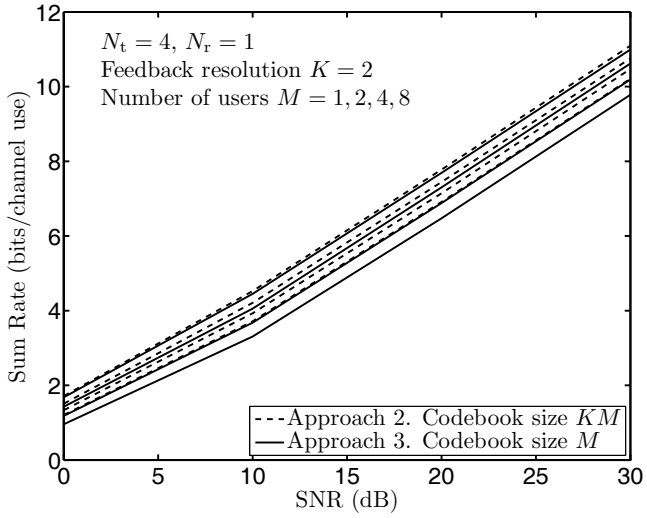

Fig. 3. Comparison of quantized scheduling and joint scheduling-precoding schemes over 4×1 channels.

a system with 8 users outperforms a point-to-point system by approximately 3 dB.

Figure 2 illustrates the performance of Approach 2 over 4×1 channels with different feedback resolution K. It is indicated that increasing K may be more efficient than increasing the number of users in certain cases. A significant portion of the perfect-CSIT gain (and single-user scheduling) is achieved even with such low-rate feedback schemes.

Figure 3 compares the joint scheduling precoding scheme (Approach 2) with that of the scheduling-only scheme ('Approach 3,' cf. Section 4) over 4×1 Rayleigh channels. There is a clear gap in performance between the two schemes, emphasizing the sub-optimality of Approach 3. Furthermore, the difference widens as the feedback resolution K increases (not plotted herein) since the joint scheme continues to benefit from an increasing number of precoding matrices to choose from. On the other hand, as K increases, the scheduling-only scheme merely benefits from a finer user selection.

6. REFERENCES

[1] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1691–1706, July 2003.

[2] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1912–1921, Aug. 2003.

[3] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.

[4] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inform. Theory*, vol. 52, pp. 3936–3964, Sept. 2006.

[5] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[6] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inform. Theory*, vol. 51, pp. 506–522, Feb. 2005.

[7] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. Inform. Theory*, vol. 52, pp. 5045–5060, Nov. 2006.

[8] 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 J. Select. Areas Commun.*, vol. 16, pp. 1423–1436, Oct. 1998.

[9] V. Lau, Y. Liu, and T.-A. Chen, "On the design of MIMO block-fading channels with feedback-link capacity constraint," *IEEE Trans. Commun.*, vol. 52, pp. 62–70, Jan. 2004.

[10] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 2, pp. 499–533, 1998.