Comparison of Quantization Techniques for Downlink Multi-User MIMO Channels with Limited Feedback

Qi Wang, Student Member, IEEE, Hao Feng, Student Member, IEEE, Leonard J. Cimini, Fellow, IEEE, Larry J. Greenstein, Life Fellow, IEEE, Douglas S. Chan, Senior Member, IEEE, and Ahmadreza Hedayat, Member, IEEE

Abstract—This letter evaluates three quantization techniques for downlink multiple-user multiple-input multiple-output (MU-MIMO) systems with limited feedback. The required feedback bits for a specified rate loss are quantified, as well as the complexity for each technique. Furthermore, the net capacity, which incorporates the effect of the overhead, is studied. The analysis and simulation results reveal the advantages and drawbacks of each quantization method and demonstrate under what conditions to use one of them rather than the other.

Index Terms—Multi-user, MIMO, limited feedback, overhead.

I. INTRODUCTION

MU-MIMO, which enables an access point (AP) or base station (BS) to communicate with more than one user simultaneously using the same spectrum, shows promise for increasing data rates [1]–[2]. Since the user terminal usually has a very limited number of antennas in commodity systems, MU-MIMO can better utilize resources compared with conventional single-user systems. However, the performance of MU-MIMO relies heavily on the accuracy of channel state information (CSI); perfect CSI at the AP/BS is required so that the interference among users can be cancelled through transmit beamforming. It is more likely, however, for the AP/BS to have only quantized observations, which could result in a severe performance degradation for MU-MIMO systems [3]–[5].

Several quantization techniques have been proposed and applied to different scenarios. Scalar quantization (SQ) is an inefficient but easy-to-implement approach and has been adopted in wireless local area networks (WLAN) standards [1]. In WLAN environments, the channel state usually remains constant for several transmission blocks. Thus, SQ is favored due to its low complexity. By contrast, vector quantization (VQ) improves the effectiveness of quantization at the cost of additional computational complexity. In cellular systems [2], the channel changes more rapidly due to user mobility and dynamic variations across a widespread environment. The quantization effectiveness is more important in this case and thus VQ is more desirable. Another quantization approach, sparse coding quantization (SCQ) proposed in [5], provides a tradeoff between these two extreme cases.

As we see, the accuracy and performance of different quantization approaches vary significantly. As a result, the relevance of each approach changes across applications. This letter evaluates three techniques, SQ, VQ and SCQ, in terms of complexity and net capacity, a metric we introduce here to include the impact of feedback overhead on throughput performance. We approximate the total number of feedback bits required to maintain a constant rate loss. In addition, the computational complexity and memory requirements for each quantization method are discussed. The analysis and simulation provide guidelines for determining the preferred technique, for a specific criterion and system configuration.

Notation: Italicized font denotes scalars, bold lower-case denotes vectors, and bold upper-case denotes matrices. The superscripts $(\cdot)^T$ and $(\cdot)^H$ represent the transpose and Hermitian operations, respectively. $I_N$ is the $N \times N$ identity matrix. All logarithm operations are base-2. $\Re(\cdot)$ and $\Im(\cdot)$ represent the real and imaginary parts of a complex number. $\lfloor \cdot \rfloor$ represents rounding the integer closest to the real number.

II. SYSTEM MODEL

A. Downlink

We focus on a downlink MU-MIMO system with $L$ users, each equipped with $N_r$ antennas. To satisfy the requirements of the beamforming technique applied, the number of transmit antennas $N_t$ at the AP\footnote{Note that this model is applicable to either cellular systems or wireless local area networks, and we will use AP to represent the transmitter from this point forward, without loss of generality.} is assumed to be equal to the aggregated number of receive antennas, i.e., $N_t = L N_r$. We assume that $N_r$ streams are sent from the AP to each user. Let $x_l$ be a vector (length $N_r$) of data symbols for the $l$th user, with the power constraint $E[|x_l|^2] = P_l$ and $\sum_{l=1}^{L} tr(P_l) \leq P$, where $P$ is the total transmit power. For simplicity, uniform power allocation among all data streams is assumed, i.e., $P_l = \frac{P}{L}, l = 1, 2, \cdots, L$; this is asymptotically optimal for high SNR [4]. At the AP, the data vector $x_l$ is beamformed by an $N_t \times N_r$ matrix $V_l$ and sent through $N_t$ antennas. We assume that $V_l^H V_j = I_{N_r}$ because of the power constraint. Then, the received signal at the $l$th user is $y_l = H_l^H V_l x_l + \sum_{j=1,j\neq l}^{L} H_l^H V_j x_j + n_l$, where $H_l$ is the $N_t \times N_r$ channel matrix and $n_l$ is the noise vector (length $N_r$) at the $l$th user. We assume that the channel experiences
independent and identically distributed (i.i.d.) Rayleigh fading, and all links have the same mean-square path gain, which has been normalized to 1. The entries of $H_l$ and $n_l$ are modelled by i.i.d. complex Gaussian random variables with zero mean and unit variance, i.e., $CN(0, 1)$.

We assume here that block diagonalization (BD) is applied to design the transmit beamforming matrix $V_j$. With perfect CSI at the AP, BD decomposes a downlink MU-MIMO channel into multiple parallel independent point-to-point MIMO links; thus each user can receive its own signals with no interference. This is only possible when $N_t \geq LN_r$ [6]. Here, we only consider the case with the maximum number of active users, i.e., $L = N_t/N_r$. We also assume that $N_r$ streams are sent from the AP to each user. In particular, $V_j$ is chosen to satisfy $H_j^H V_j = 0$, $\forall j \neq l$ so that the interference from other users is canceled. A closed-form expression for $V_j$ can be derived based on the singular-value decomposition of the aggregated channel matrix of the other users [6].

In practical communication systems, the estimated CSI at each user is represented by $B$ bits, which are fed back to the AP. The calculation of $V_j$ at the AP is based on the quantized version of the channel $\hat{H}_j$, $\forall j \neq l$. So, residual interference exists, and the system performance is interference-limited at high SNR. The relationship between the imperfect channel $\hat{H}_j$ and the true one $H_j$ depends on several factors, including estimation error, quantization, and delay. In this letter, we assume perfect CSI is available at the receiver, and, instead, focus on the impact of quantization methods applied at the user terminal.

### B. Performance Metric

Sum rate, $R_{\text{sum}}$, is a conventional metric to measure the performance of a communication system. However, the effect of feedback overhead is usually not taken into account. In practice, CSI feedback frames occupy the time or spectrum resources for data transmission and thereby degrade the actual data rate. In order to evaluate the implicit overhead-performance tradeoff, we use the net capacity [7, Chapter 8]

$$R_{\text{net}} = R_{\text{sum}}(1 - \theta),$$

where $\theta = T_f/\tau$ is the overhead ratio in time, $T_f = LB/R_{tb}$ represents the transmission time spent on feedback packets, $B$ is the number of feedback bits for each user, and $R_{tb}$ is the transmission rate for the feedback channel. We assume that CSI measurements are made at time intervals of length $\tau$, where $\tau$ is large compared to the symbol length. We also choose $\tau$ based on the autocorrelation function $\rho = J_0(2\pi f_d \tau)$, where $f_d$ is the Doppler frequency determined by the radio frequency wavelength, $\lambda$, the speed of the mobile user $v$ (i.e., $f_d = v/\lambda$), and $J_0(\cdot)$ is the zero-th order Bessel function of the first kind. The ergodic sum rate can be approximated as a function of the number of feedback bits [5]; accordingly, the approximate average net capacity can be derived.

Note that we ignore the preamble and header in the CSI feedback frame, and only the quantization bits are considered as overhead in our analysis. In a practical multicarrier system, each user needs to feed back the quantized channels for all subcarriers; this implies that quantization bits dominate the overall feedback bits.

### III. Quantization Methods

In this section, we study three quantization methods, namely scalar quantization (SQ), vector quantization (VQ), and sparse coding quantization (SCQ). We approximate and compare the required number of feedback bits for these quantization schemes, as well as their complexity. In the following, the subscripts (user index) in the channel and beamforming matrices are omitted because the quantization is a general process which can be applied for all users.

#### A. Scalar Quantization (SQ)

Since the AP only needs the spatial direction of the channel to eliminate the interference, the channel matrix $H$ is normalized and then quantized using SQ at each user. The real and imaginary parts of each complex element $h_{ij}$ in $H$ are quantized to $B_{SQ}$ bits, respectively. Note that one bit is reserved for the sign of each of the real and imaginary parts. In [1], the quantized version of $h_{ij}$ is $\tilde{h}_{ij} = \frac{1}{m} \left(\frac{Q_{B_{SQ}} - 1}{2}\right)$, where $m$ is a scaling ratio which guarantees the real/imaginary element in the normalized channel matrix is always less than or equal to one. In particular, $m$ can be chosen as the maximum value among all real and imaginary elements of the channel matrix $H$. Therefore, the number of feedback bits needed at each user is $B = 2N_tN_rB_{SQ}$, and the total amount of feedback overhead is $LB = 2N_t^2B_{SQ}$.

A closed-form expression for the relationship between the rate loss and the number of feedback bits is usually intractable. A feasible approach is to approximate the quantization error as a random variable with a given distribution, e.g., uniformly distributed in $[-2^{-B_{SQ}+1}, 2^{-B_{SQ}+1}]$. For the sake of analytical simplicity, we assume the quantization error is a Gaussian random variable with zero mean and variance $\sigma_r^2 = \frac{1}{12} 2^{-2B_{SQ}+2}$ [9]. As the sum rate is a function of the variance [5], by applying a similar approach as in Theorem 2 in [4], an approximation for the number of feedback bits required to maintain a performance gap of no more than 3 dB with respect to the case with perfect CSI is derived to be

$$B \approx 2N_tN_r \left( \frac{P_{db}}{3} - \frac{1}{2} \log \frac{N_t}{12} \left( \frac{5}{2^\frac{B_{SQ}}{} - 1} \right) \right),$$

where $P_{db} = 10 \log_{10} P$ is the normalized transmit power in units of dB.

The overall computation time for SQ is $O(N_t^2)$, which is independent of $B_{SQ}$. Furthermore, SQ does not require additional storage. Thus, increasing the resolution of the SQ quantizer does not affect the complexity. However, a large number of feedback bits might be required, especially when the number of antennas is large.

#### B. Vector Quantization (VQ)

With VQ, $\hat{H}$ is chosen from a codebook according to

$$\hat{H} = \underset{W \in c}{\arg\min} d^2(H, W),$$

This is based on the familiar Doppler spectrum due to Jakes [8], which we assume here for the sake of specificity.

$3$Although not shown in this letter, numerical results show that the Gaussian assumption provides an upper bound to characterize the quantization error.
where $C$ is a vector quantization codebook of size $2^{B_{VQ}}$. Each codeword $W \in C$ is an $N_t \times N_r$ unitary matrix, i.e., $W^H W = I_{N_r}$, and is independently and uniformly chosen from a unit sphere defined in an $N_t \times N_r$-dimensional complex space [10]. $d^2(H, W) = N_r - \text{tr}(H^H W W^H H)$ is the chordal distance, where $H$ is an orthonormal basis for the subspace spanned by the columns of $H$.

Each user chooses an appropriate codeword and sends back the index of the codeword. Thus, the number of feedback bits per user is $B = B_{VQ}$. The relation between the quantization error of VQ and $B_{VQ}$ has been extensively investigated. In general, the quantization error can be modeled as an additive Gaussian noise with zero mean and variance $\sigma^2_{VQ}$, which is related to the number of feedback bits. An approximation for $\sigma^2_{VQ}$ is given by [11],

$$\sigma^2_{VQ} \approx \frac{1}{T} \left( \frac{1}{T} \right)^2 C_{N_t, N_r} + 2^{-B_{VQ}},$$

(3)

where $T = N_r(N_t - N_r)$, $C_{N_t, N_r} = \frac{1}{T} \prod_{i=1}^{N_t} \left( \frac{N_r - i}{N_r - 1} \right)$, and $\Gamma(\cdot)$ represents the Gamma function. An approximation for the required number of bits for VQ to maintain the 3-dB performance gap is provided in [4]:

$$B \approx T \left( \frac{P_{dB}}{3} - \log N_r \right) - \log C_{N_t, N_r}.$$  

(4)

It has been shown that VQ requires significantly fewer feedback bits than SQ. However, the main drawback of VQ is the computational complexity and storage requirement, especially when the codebook is large. The overall time complexity of searching codewords and calculating the chordal distance is $O(L^2 B_{VQ})$. The storage requirement, on the other hand, is $O(N_t^2 2^{B_{VQ}})$. Using the simulation model in [12] as an example, 20-bit VQ is required to achieve acceptable performance. This indicates that each user must generate and store a codebook containing $2^{20}$ codewords and select the appropriate codeword from this large codebook.

C. Sparse Coding Quantization (SCQ)

Compared with conventional VQ, SCQ (proposed in [5]) achieves a much smaller codebook by exploiting a linear combination of several codewords to characterize the channel.

In particular, the quantized channel $\tilde{H}$ is assumed to be a linear combination of all codewords $\tilde{H} = \sum_{i=1}^{2^3} \alpha_i W_i$, where $W_i \in C$ is a codebook with $2^3$ codewords. The coefficient vector $\alpha = [\alpha_1, \ldots, \alpha_K]^T$ can be obtained by solving the following optimization problem

$$\min_{\alpha} d^2(\tilde{H}, H) \text{ s.t. } ||\alpha||_{\ell_0} \leq K$$

(5)

where $\ell_0$-norm $||\alpha||_{\ell_0}$ denotes the number of nonzero elements in $\alpha$, and the constraint $||\alpha||_{\ell_0} \leq K$ guarantees that no more than $K$ codewords are used to represent the true channel $H$. Each user sends back $K$ indices and a vector consisting of $K$ coefficients, which is quantized by another codebook of size $2^K$. Thus, the number of feedback bits per user is $B = K(\beta + \mu)$.

Based on (4) and the analysis in [5], the required codebook sizes for SCQ to maintain the 3-dB performance gap can be approximated as

$$\beta \approx T \left( \frac{P_{dB}}{3} - \log K N_r \right) - \log C_{N_t, N_r},$$

$$\mu \approx K - 1 \left[ \left( \frac{P_{dB}}{3} - \log N_r \right) - \log C_{N_t, N_r} \right].$$

(6)

In most scenarios, by setting $K = 2$, the benefit of SCQ can be achieved without incurring much overhead [5]. Accordingly, the required number of feedback bits is

$$B \approx \frac{2T^2 + 1}{T} \left( \frac{P_{dB}}{3} - \log N_r \right) - 2T - \frac{1 + T}{T} \log C_{N_t, N_r}.$$  

(7)

The time and space complexities of SCQ are $O(KL^2 2^3 + L2^3)$ and $O(N_t^2 2^3 + KL2^3)$, respectively. The analysis and simulation results in [5] illustrate that compared to SQ and VQ, SCQ provides an interesting option that allows us to trade off overhead for complexity.

IV. RESULTS

In the simulations, we consider a downlink MU-MIMO system which consists of a four-antenna AP and two users, each of which is equipped with two antennas. The results for SQ are averaged over $10^4$ channel realizations. For VQ and SCQ, $10^3$ independent realizations of the quantization codebooks are generated for each user. For each possible codebook, all simulation results are averaged over $10^4$ channel realizations. For all scenarios, uniform power allocation is assumed, and the feedback rate $R_{fb}$ is assumed to be 6.5 Mbps, which is the lowest data rate in the 802.11ac standard [1].

In Fig. 1, the required feedback overhead (in bits) per user for each quantization technique is illustrated; the sum rate performance is 3 dB less than that with perfect CSI at the AP. Here, the quantization error is assumed to be Gaussian, so the results are upper bounds on the required number of feedback bits. We can see that, as SNR increases, more feedback bits are needed so that stronger interference can be reduced to a specific level. Furthermore, SCQ incurs more overhead than VQ but requires a smaller number of feedback bits than SQ. Another observation from Fig. 1 is that VQ requires a 29-bit codebook for each user when SNR is 24 dB, implying that...
the computational complexity of VQ is $O(2^9)$. Meanwhile, the complexity of SQ, which is independent of the number of feedback bits, keeps constant ($O(1)$). According to (6), the complexity of SCQ is $O(2^{26})$, which indicates that SCQ provides an option for balancing complexity and overhead. Though the given example is for the scenario of high-rate feedback, our analysis is applicable to the low-rate feedback regime as well. From Fig. 1, we can see that only a small number of feedback bits is required for any quantization method at low SNR; in that range the analytical results in Sec. III reveal that SCQ also provides benefit in complexity compared to VQ.

In Fig. 2, the net capacity achieved by these three quantization methods is plotted as a function of the overhead ratio $\theta = B / R_{fb} \tau$, where $\rho$ is the autocorrelation of the random path gain for time separation $\tau$. For given $R_{fb}$ and $\tau$ (determined by $\rho$), the overhead ratio $\theta$ is determined by the total number of feedback bits, which varies from 16 to 128. We consider a high mobility environment: the velocity $v$ is assumed to be 25 m/s, i.e., around 90 km/h. The average received SNR is assumed to be 20 dB. The results indicate that, as $\theta$ increases, the net capacity first increases due to having more accurate CSI. However, the capacity eventually decreases because of the excessive amount of overhead. Although VQ outperforms SCQ in terms of net capacity, the performance gap is acceptable. In addition, SQ gives the worst performance as expected. When $\theta$ is sufficiently large, where the net capacity is overwhelmed by the overhead required, the performance of each technique reduces to the same level.

By choosing an appropriate number of feedback bits, the net capacity can be maximized for given system configurations. Fig. 3 illustrates the relation between the optimal net capacity and the velocity of the mobile users $v$. We can see that, as the mobility increases, the optimal net capacity decreases for any of these quantization techniques because the channel changes more rapidly. Another observation from Figs. 2 and 3 is that the impact of overhead becomes less important as $\rho$ decreases, indicating a preference for SQ due to its simplicity.

## V. Conclusion

In this letter, we have evaluated and compared three quantization methods in terms of the required number of feedback bits, the net capacity, and the complexity. The results illustrate the tradeoff among these schemes and show that SCQ is an option for balancing performance and complexity. SQ is more preferable when the overhead effect is negligible, e.g., when the channel condition changes very slowly.

## Acknowledgement

We thank the editor and the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

## References