

Noncoherent Trellis Coded Quantization: A Practical Limited Feedback Technique for Massive MIMO Systems

Junil Choi, Zachary Chance, David J. Love, and Upamanyu Madhow

Abstract—Accurate channel state information (CSI) is essential for attaining beamforming gains in single-user (SU) multiple-input multiple-output (MIMO) and multiplexing gains in multi-user (MU) MIMO wireless communication systems. State-of-the-art limited feedback schemes, which rely on pre-defined codebooks for channel quantization, are only appropriate for a small number of transmit antennas and low feedback overhead. In order to scale informed transmitter schemes to emerging massive MIMO systems with a large number of transmit antennas at the base station, one common approach is to employ time division duplexing (TDD) and to exploit the implicit feedback obtained from channel reciprocity. However, most existing cellular deployments are based on frequency division duplexing (FDD), hence it is of great interest to explore backwards compatible massive MIMO upgrades of such systems. For a fixed feedback rate per antenna, the number of codewords for quantizing the channel grows exponentially with the number of antennas, hence generating feedback based on look-up from a standard vector quantized codebook does not scale. In this paper, we propose noncoherent trellis-coded quantization (NTCQ), whose encoding complexity scales linearly with the number of antennas. The approach exploits the duality between source encoding in a Grassmannian manifold (for finding a vector in the codebook which maximizes beamforming gain) and noncoherent sequence detection (for maximum likelihood decoding subject to uncertainty in the channel gain). Furthermore, since noncoherent detection can be realized near-optimally using a bank of coherent detectors, we obtain a low-complexity implementation of NTCQ encoding using an off-the-shelf Viterbi algorithm applied to standard trellis coded quantization. We also develop advanced NTCQ schemes which utilize various channel properties such as temporal/spatial correlations. Monte Carlo simulation results show the proposed NTCQ and its extensions can achieve near-optimal performance with moderate complexity and feedback overhead.

Index Terms—Massive MIMO systems, limited feedback, trellis-coded quantization (TCQ), noncoherent TCQ.

Manuscript received May 21, 2013; revised September 17, 2013. The editor coordinating the review of this paper and approving it for publication was A. Ghayeb.

J. Choi and D. Love are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN (e-mail: {choi215, djlove}@purdue.edu).

Z. Chance is with MIT Lincoln Laboratory, Lexington, MA (e-mail: zac.chance@gmail.com).

U. Madhow is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA (e-mail: madhow@ece.ucsb.edu).

This paper was presented in part at the Information Theory and Application workshop, UCSD, 2013 [1] and the IEEE Conference on Information Sciences and Systems, Johns Hopkins University, 2013 [2].

Digital Object Identifier 10.1109/TCOMM.2013.111413.130379

I. INTRODUCTION

THE concept of wireless systems employing a large number of transmit antennas, often dubbed massive multiple-input multiple-output (MIMO) systems, has been evolving over the past few years. It was found in [3] that adding more antennas at the base station is always beneficial even with very noisy channel estimation because the base station can recover information even with a low signal-to-noise-ratio (SNR) once it has sufficiently many antennas. This motivates the concept of using a very large number of transmit antennas, where the number of antenna elements can be at least an order of magnitude more than the current cellular systems (10s-100s) [4]. Massive MIMO systems have the potential to revolutionize cellular deployments by accommodating a large number of users in the same time-frequency slot to boost the network capacity [5] and by increasing the range of transmission with improved power efficiency [6]. Recently, fundamental limits, optimal transmit precoding and receive strategies, and real channel measurement issues for massive MIMO systems were studied and summarized in [7] (see also the references therein).

When the transmitter has multiple antennas, channel state information (CSI) can provide significant performance gains, including beamforming gains in single-user (SU) multiple-input multiple-output (MIMO) systems and multiplexing gains in multi-user (MU) MIMO systems. Unlike conventional MU-MIMO systems with a small number of transmit antennas, massive MU-MIMO can be implemented with simple per-user beamforming such as matched beamforming due to the large number of degrees-of-freedom available in the user channels [4]. However, without accurate CSI, massive MU-MIMO systems would also experience a sum-rate saturation, which is known as a *ceiling effect*, even if the base station transmit power is unconstrained [8], [9].

The challenge, therefore, is to scale channel estimation and feedback strategies to effectively provide CSI. Most of the literature on massive MIMO sidesteps this challenge by focusing on time division duplexing (TDD), for which CSI can be extracted *implicitly* using reciprocity. However, since most cellular systems today employ frequency division duplexing (FDD), it is of great interest to explore effective approaches for obtaining CSI for massive MIMO upgrades of such systems. This motivates the work in this paper, which explores efficient approaches for quantizing high-dimensional channel vectors to generate CSI feedback.

There is a large body of literature devoted to accurate CSI

quantization for closed-loop MIMO FDD systems with a relatively small number of antennas [10]. Most approaches employ a common vector quantized (VQ) codebook at the transmitter and the receiver, and the explicit feedback sequence is simply the binary index of the codeword chosen in the codebook. Thus, the main focus has been on codebook design. For i.i.d. Rayleigh fading channel models, deterministic codebook techniques using Grassmannian line packing (GLP) were developed in [11]–[13], and the performance of random vector quantization (RVQ) codebooks was analyzed in [14], [15]. Limited feedback codebooks that adapt to spatially correlated channels were studied in [16]–[18], and temporal correlated channels were developed in [19]–[26].

It has been shown in [14] that an RVQ codebook is asymptotically optimal for i.i.d. Rayleigh fading channels when the number of transmit antennas gets large, assuming a fixed number of feedback bits per antenna. However, existing codebook-based techniques do not scale to approach the RVQ benchmark. In order to maintain the same level of channel quantization error, the feedback overhead must increase proportional to the number of transmit antennas [15], [27]. While the linear increase in feedback overhead with the number of antennas may be acceptable as we scale to massive MIMO, the corresponding exponential increase in codebook size makes a direct look-up approach for feedback generation infeasible.

In order to address this gap in source coding techniques, it is natural to turn to the duality between source and channel coding. Just as RVQ provides a benchmark for source coding, random coding produces information-theoretic benchmarks for channel coding. However, there are thousands of papers dedicated to practical channel code designs that aim to approach these benchmarks, with codes such as convolutional codes, Reed-Solomon codes, turbo codes, and LDPC codes implemented in practice [28]. While these ideas can and have been leveraged for source coding, the measures of distortion used have been the Hamming or Euclidean distortion. Our contribution in this paper is to establish and exploit the connection between source coding on the Grassmannian manifold (which is what is needed for the limited feedback application of interest to us) and channel coding for *noncoherent* communication. We coin the term *noncoherent trellis-coded quantization (NTCQ)* for the class of schemes that we propose and investigate. Our approach avoids the computational bottleneck of look-up based codebooks, with encoding complexity scaling linearly with the number of antennas, and its performance is near-optimal, approaching that of RVQ.

Approach: Our NTCQ approach relies on two key observations:

- (a) Quantization for beamforming requires finding a quantized vector, from among the available choices, that is best aligned with the true channel vector, in terms of maximizing the magnitude of their normalized inner product. This corresponds to a search on the Grassmann manifold rather than in Euclidean space. We point out, as have others before us, that this source coding problem maps to a channel coding problem of *noncoherent* sequence detection, where we try to find the most likely transmitted codeword subject to an unknown multiplicative complex-valued channel gain.
- (b) We know from prior work on noncoherent communication

that a noncoherent block demodulator can be implemented near-optimally using a bank of coherent demodulators, each with a different hypothesis on the unknown channel gain. Furthermore, signal designs and codes for coherent communication are optimal for noncoherent communication, as long as we adjust our encoding and decoding slightly to account for the ambiguity caused by the unknown channel gain.

The relationship between quantization based on a mean squared error cost function and channel coding for *coherent* communication over the AWGN channel has been exploited successfully in the design of trellis coded quantization (TCQ) [29], in which the code symbols take values from a standard finite constellation used for communication, such as phase shift keying (PSK) or quadrature amplitude modulation (QAM). The quantized code vector can then be found by using a Viterbi algorithm for trellis decoding. Our observation (b) allows us to immediately extend this strategy to the noncoherent setting. The code vectors for NTCQ can be exactly the same as in standard TCQ, but the encoder now consists of several Viterbi algorithms (in practice, a very small number) running in parallel, with a rule for choosing the best output. Thus, while approximating a beamforming vector on the Grassmann manifold as in (a) appears to be difficult, it can be easily solved by using several parallel searches in Euclidean space. Furthermore, just as noncoherent channel codes inherit the good performance of the coherent codes they were constructed from, NTCQ inherits the good quantization performance of TCQ.

Contributions: Our contributions are summarized as follows:

- We show that channel codes, and by analogy, source codes developed in a coherent setting can be effectively leveraged in the noncoherent setting of interest in CSI generation for beamforming. As shown through both analysis and simulations, the resulting NTCQ strategy provides near-optimal beamforming gain, and has encoding complexity which is linear in the channel dimension.
- We also develop adaptive NTCQ techniques that are optimized for spatial and temporal correlations. A differential version of NTCQ utilizes the temporal correlation of the channel to successively refine the quantized channel to decrease the quantization error. A spatially adaptive version of NTCQ exploits the spatial correlation of the channel so that it only quantizes the local area of the dominant direction of the spatial correlation matrix. Utilization of channel statistics using such advanced schemes can significantly improve the performance or decrease the feedback overhead by utilizing channel statistics.

An important feature of NTCQ is its flexibility, which makes it an attractive candidate for potentially providing a common channel quantization approach for heterogeneous fifth generation (5G) wireless communication systems, which could involve a mix of advanced network entities such as massive MIMO, coordinated multipoint (CoMP) transmission, relay, distributed antenna systems (DAS), and femto/pico cells. For example, massive MIMO systems could be implemented using a two-dimensional (2D) planar antenna array at the base station to reduce the size of antenna array [30]. Depending on the channel quality, the base station could turn on and off the

rows/columns of this 2D array to achieve better performance. The same situation could be encountered in CoMP and DAS because the number of coordinating transmit stations may vary over time. NTCQ can easily adjust to such scenarios, since it can adapt to different numbers of transmit antennas (or more generally, space-time channel dimension) by changing the number of code symbols, and can adapt CSI accuracy and feedback overhead by changing the constellation size and the coded modulation scheme.

Related work: We have already mentioned conventional look-up based quantization approaches and discussed why they do not scale. Trellis-based quantizers for CSI generation have been proposed previously in [31]–[34], but the path metrics used for the trellis search are *ad hoc*. On the other hand, the mapping to noncoherent sequence detection, similar to NTCQ, has been pointed out in [35]. Depending on the number of constellation points used for the candidate codewords, the proposed algorithms in [35] are dubbed as PSK & QAM singular vector quantization (SVQ). Although PSK/QAM-SVQ adopt similar codeword search methods as NTCQ, they do not consider coding. The use of nontrivial trellis codes as proposed here significantly enhances performance compared to PSK/QAM-SVQ with the same amount of feedback overhead. Furthermore, [35] employs optimal noncoherent block demodulation, derived in [36], [37], for quantization, incurring complexity $O(M_t^3)$ for QAM-SVQ and $O(M_t \log M_t)$ for PSK-SVQ, where M_t denotes the number of antennas. Our NTCQ scheme exhibits better complexity scaling: near-optimal demodulation in $O(M_t)$ complexity by running a small number of coherent decoders in parallel, as proposed in [38], suffices for providing near-optimal quantization performance.

The remainder of this paper is organized as follows. In Section II, we describe the system model and fundamentals underlying NTCQ. A detailed description of the NTCQ algorithm and its variation is provided in Section III. Advanced NTCQ schemes that exploit temporal and spatial correlation of channels are explained in Section IV. In Section V, simulation results are presented, and conclusions follow in Section VI.

II. SYSTEM MODEL AND THEORY

A. System Setup

We consider a block fading multiple-input single-output (MISO) communications system with M_t transmit antennas at the transmitter as in Fig. 1. The received signal, $y_\ell[k] \in \mathbb{C}$, for a channel use index ℓ in the k th fading block can be written as¹

$$y_\ell[k] = \mathbf{h}^H[k] \mathbf{f}[k] s_\ell[k] + z_\ell[k],$$

where $\mathbf{h}[k] \in \mathbb{C}^{M_t}$ is the MISO channel vector, $\mathbf{f}[k] \in \mathbb{C}^{M_t}$ is the beamforming vector with $\|\mathbf{f}[k]\|_2^2 = 1$, $s_\ell[k] \in \mathbb{C}$ is the message signal with $E[s_\ell[k]] = 0$ and $E[|s_\ell[k]|^2] = \rho$, and $z_\ell[k] \in \mathbb{C}$ is additive complex Gaussian noise such that $z_\ell[k] \sim \mathcal{CN}(0, \sigma^2)$. A number of different models for $\mathbf{h}[k]$

will be considered in the design and performance evaluation of quantization schemes, but for now, we allow it to be arbitrary. The receiver quantizes its estimate of $\mathbf{h}[k]$ into a B_{tot} -dimensional binary vector $\mathbf{b}[k]$, which is sent over a limited rate feedback channel. The transmitter uses this feedback to construct a beamforming vector $\mathbf{f}[k]$. In order to focus attention on channel quantization, we do not model channel estimation errors at the receiver or errors over the feedback channel.

Since we do not consider temporal correlation in $\{\mathbf{h}[k]\}$ for quantizer design in this section, we drop the time index k for the remainder of this section. Assuming an average power constraint at the transmitter, we wish to choose \mathbf{f} so as to maximize the *normalized beamforming gain* that is defined as

$$J(\mathbf{f}, \mathbf{h}) = \frac{|\mathbf{h}^H \mathbf{f}|^2}{\|\mathbf{h}\|_2^2 \|\mathbf{f}\|_2^2}. \quad (1)$$

Although $\|\mathbf{f}\|_2 = 1$, we still normalize with $\|\mathbf{f}\|_2$ in (1) to maintain notational generality. An equivalent approach is to minimize the *chordal distance* between \mathbf{f} and \mathbf{h} , defined as

$$d_c^2(\mathbf{f}, \mathbf{h}) = 1 - J(\mathbf{f}, \mathbf{h}) = 1 - \frac{|\mathbf{h}^H \mathbf{f}|^2}{\|\mathbf{h}\|_2^2 \|\mathbf{f}\|_2^2}.$$

These performance measures require searching for codewords on the Grassmann manifold, a projective space in which vectors are mapped to one-dimensional complex subspaces.

Conventional VQ codebook-based channel quantization typically employs exhaustive search to select a codeword from an unstructured and fixed B_{tot} -bit codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{2^{B_{\text{tot}}}}\}$ according to

$$\mathbf{c}_{\text{opt}} = \underset{\mathbf{c} \in \mathcal{C}}{\text{argmax}} J(\mathbf{c}, \mathbf{h}) = \underset{\mathbf{c} \in \mathcal{C}}{\text{argmin}} d_c^2(\mathbf{c}, \mathbf{h}), \quad (2)$$

and the binary sequence $\mathbf{b} = \text{bin}(\text{opt})$ is fed back to the transmitter where $\text{bin}(\cdot)$ converts an integer to its binary representation. Then the beamforming vector is reconstructed at the transmitter as

$$\mathbf{f} = \frac{\mathbf{c}_{\text{int}(\mathbf{b})}}{\|\mathbf{c}_{\text{int}(\mathbf{b})}\|_2}$$

where $\text{int}(\cdot)$ converts a binary string into an integer. Exhaustive search, which does not require geometric interpretation of the performance metric, incurs computational complexity $O(M_t 2^{B_{\text{tot}}})$, which is exponential in the number of bits. We shall see that utilizing the geometry of the Grassmann manifold, and in particular, relating it to Euclidean geometry, is key to more efficient quantization procedures.

Since our performance criterion is independent of the codeword norm, one could, without loss of generality, normalize the codewords to unit norm up front (i.e., set $\|\mathbf{c}\|_2 \equiv 1$). However, for the code constructions and quantizer designs of interest to us, it is useful to allow codewords to have different norms (the performance criterion, of course, remains independent of codeword scaling).

B. Feedback Overhead

The relation between the feedback overhead B_{tot} (or codebook size $2^{B_{\text{tot}}}$) and the performance of MIMO systems has been thoroughly investigated for i.i.d. Rayleigh fading

¹Lower- and upper-case bold symbols denote vectors and matrices, respectively. The two-norm of a vector \mathbf{x} is denoted as $\|\mathbf{x}\|_2$. The transpose and Hermitian transpose of a vector \mathbf{x} are denoted by \mathbf{x}^T , \mathbf{x}^H respectively. The expectation operator is denoted by $E[\cdot]$, and $X \sim \mathcal{CN}(m, \sigma^2)$ indicates that X is a complex Gaussian random variable with mean m and variance σ^2 .

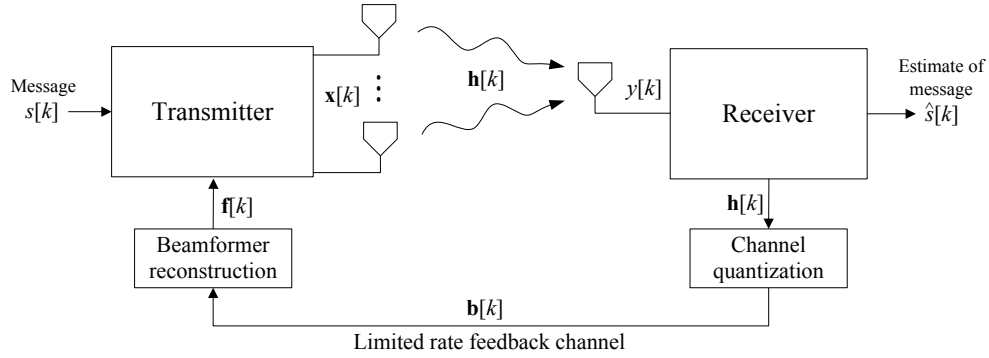


Fig. 1. Multiple-input, single-output communications system with feedback.

channels. In single user (SU) MISO channels with the B_{tot} bits RVQ codebook, the loss in normalized beamforming gain is given as [15]

$$E \left[1 - \max_{\mathbf{f} \in \mathcal{F}_{\text{RVQ}}} J(\mathbf{f}, \mathbf{h}) \right] = 2^{B_{\text{tot}}} \beta \left(2^{B_{\text{tot}}}, \frac{M_t}{M_t - 1} \right) \approx 2^{-\frac{B_{\text{tot}}}{M_t - 1}} \quad (3)$$

where \mathcal{F}_{RVQ} is an RVQ codebook, $\beta(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the Beta function, $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the Gamma function, and expectation is taken over \mathbf{h} and \mathcal{F}_{RVQ} . The expression in (3) indicates that the feedback overhead needs to be increased proportional to M_t to maintain the loss in normalized beamforming gain at a certain level.

For MU-MIMO zero-forcing beamforming (ZFBF), a similar conclusion is drawn in [8], [9]: in order to achieve the full multiplexing gain of M_t , the number of feedback bits per user, B_{user} , must scale linearly with SNR (in dB) and M_t as

$$B_{\text{user}} = (M_t - 1) \log_2 \rho \approx \frac{M_t - 1}{3} \rho_{\text{dB}}.$$

We therefore assume that at each channel use, the receiver sends back a binary feedback sequence of length

$$B_{\text{tot}} \triangleq B M_t + q$$

where B is the number of quantization bits used per transmit antenna and q is a small, fixed number of auxiliary feedback bits, which does not scale with M_t .

While linear scaling of feedback bits with the number of transmit elements is typically acceptable in terms of overhead, a VQ codebook-based limited feedback is computationally infeasible for massive MIMO systems with large M_t because of the exponential growth of codeword search complexity with M_t as $O(M_t 2^{B M_t})$. Thus, we need to develop new techniques to quantize CSI for large M_t .

In order to develop an efficient CSI quantization method for massive MIMO systems, we draw an analogy between searching for a candidate beamforming vector to maximize beamforming gain as in (2) and noncoherent sequence detection (e.g., [31], [35]). We then employ prior work relating noncoherent and coherent detection to map quantization on the Grassmann manifold to quantization in Euclidean space, which can be accomplished far more efficiently. This line of reasoning, which corresponds to the *process* of quantization, has been previously established in [35], but we

provide a self-contained derivation in Section II-C pointing to a low-complexity, near-optimal source encoding strategy. We then show, in Section II-D that structured quantization codebooks for Euclidean metrics are effective for quantization on the Grassmann manifold. This leads to a CSI quantization framework which is efficient in terms of both overhead and computation.

C. Efficient Grassmannian Encoding using Euclidean Metrics

Consider a single antenna noncoherent, block fading, additive white Gaussian noise (AWGN) channel with received vector

$$\mathbf{y} = \beta \mathbf{x} + \mathbf{n},$$

where $\beta \in \mathbb{C}$ is an unknown complex channel gain, $\mathbf{x} \in \mathbb{C}^N$ is a vector of N transmitted symbols, $\mathbf{n} \in \mathbb{C}^N$ is complex Gaussian noise, and $\mathbf{y} \in \mathbb{C}^N$ is the received signal. Using the generalized likelihood ratio test (GLRT) as in [35], [38], the estimate of the transmitted vector, $\hat{\mathbf{x}}$, is given by

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{C}^N}{\operatorname{argmin}} \min_{\beta \in \mathbb{C}} \|\mathbf{y} - \beta \mathbf{x}\|_2^2 \quad (4)$$

$$= \underset{\mathbf{x} \in \mathbb{C}^N}{\operatorname{argmin}} \min_{\alpha \in \mathbb{R}^+, \theta \in [0, 2\pi)} \|\mathbf{y}\|_2^2 + \alpha^2 \|e^{j\theta} \mathbf{x}\|_2^2 - 2\alpha \operatorname{Re}(e^{j\theta} \mathbf{y}^H \mathbf{x}) \quad (5)$$

$$= \underset{\mathbf{x} \in \mathbb{C}^N}{\operatorname{argmin}} \min_{\alpha \in \mathbb{R}^+} \|\mathbf{y}\|_2^2 + \alpha^2 \|\mathbf{x}\|_2^2 - 2\alpha |\mathbf{y}^H \mathbf{x}| \quad (6)$$

$$= \underset{\mathbf{x} \in \mathbb{C}^N}{\operatorname{argmax}} \frac{|\mathbf{y}^H \mathbf{x}|^2}{\|\mathbf{x}\|_2^2}, \quad (7)$$

where we decomposed the entire complex plain $\beta = \alpha e^{j\theta}$ with $\alpha \in \mathbb{R}^+$ and $\theta \in [0, 2\pi)$ in (5), and (6) comes from

$$\min_{\theta \in [0, 2\pi)} \{-\operatorname{Re}(e^{j\theta} \mathbf{y}^H \mathbf{x})\} = -|\mathbf{y}^H \mathbf{x}|.$$

To derive (7), we differentiate (6) with respect to α and set to 0 which gives $\alpha^* = \frac{|\mathbf{y}^H \mathbf{x}|}{\|\mathbf{x}\|_2^2}$. Note that α^* is the global minimizer of (6) because (6) is a quadratic function of α . We can derive (7) after plugging α^* into (6) and some basic algebra.

We can easily check from (2) and (7) that finding the optimal codeword for a MISO beamforming system and the noncoherent sequence detection problems are equivalent (although this relation is already shown in [35], we proved the duality of (4) and (2) more explicitly than [35]). Therefore, we can

find \mathbf{c}_{opt} for a MISO beamforming system with a Euclidean distance quantizer (or noncoherent block demodulator)

$$\min_{\alpha \in \mathbb{R}^+} \min_{\theta \in [0, 2\pi)} \min_{\mathbf{c}_i \in \mathcal{C}} \|\bar{\mathbf{h}} - \alpha e^{j\theta} \mathbf{c}_i\|_2^2. \quad (8)$$

where $\bar{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|_2}$ is the normalized channel direction.

Moreover, instead of searching over the entire complex plane by having $\alpha \in \mathbb{R}^+$ and $\theta \in [0, 2\pi)$, we know from prior work on noncoherent communication [38] that the noncoherent block demodulator in (8) can be implemented near-optimally using a bank of coherent demodulators over the optimized discrete sets of $\alpha \in \mathbb{A} = \{\alpha_1, \alpha_2, \dots, \alpha_{K_\alpha}\}$ and $\theta \in \Theta = \{\theta_1, \theta_2, \dots, \theta_{K_\theta}\}$. While *optimal* noncoherent detection can be accomplished with quadratic complexity in M_t [35], as we show through our numerical results, a small number of parallel coherent demodulators (which incurs complexity linear in M_t) is all that is required for excellent quantization performance.

The preceding development tells us that we can apply coherent demodulation, which maps to quantization using Euclidean metrics, to noncoherent demodulation, which maps to quantization on the Grassmann manifold. However, we must still determine how to choose the quantization codebook. Next, we present results indicating that we can simply use codes optimized for Euclidean metrics for this purpose.

D. Efficient Grassmannian Codebooks based on Euclidean Metrics

We begin with an asymptotic result for i.i.d. Rayleigh fading coefficients, which relies on the well-known rate-distortion theory for i.i.d. Gaussian sources.

Theorem 1. *If we quantize an $M_t \times 1$ i.i.d. Rayleigh fading MISO channel $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \sigma_h^2 \mathbf{I})$ with a Euclidean distance quantizer using B bits per entry (which corresponds to $\frac{B}{2}$ bits per each of real and imaginary dimension) as*

$$\mathbf{g}_{\text{ED}} = \min_{\mathbf{g}_i \in \mathcal{G}} \|\mathbf{h} - \mathbf{g}_i\|_2^2 \quad (9)$$

where $\mathcal{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_{2^{B_{\text{tot}}}}\}$, $B_{\text{tot}} = BM_t$, $\mathbf{g}_i \sim \mathcal{CN}(\mathbf{0}, (\sigma_h^2 - 2D)\mathbf{I})$ for all i , and $D = \frac{1}{2}\sigma_h^2 2^{-B}$, then the asymptotic loss in normalized beamforming gain, or chordal distance, is given by

$$d_c^2(\mathbf{h}, \mathbf{g}_{\text{ED}}) \xrightarrow{M_t \rightarrow \infty} 2^{-B}. \quad (10)$$

Proof: By expanding $\|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2$, we have

$$\begin{aligned} \|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2 &= \sum_{t=1}^{M_t} \left[\{\text{Re}(h_t) - \text{Re}(g_{\text{ED},t})\}^2 \right. \\ &\quad \left. + \{\text{Im}(h_t) - \text{Im}(g_{\text{ED},t})\}^2 \right] \end{aligned}$$

where h_t and $g_{\text{ED},t}$ are the t^{th} entry of \mathbf{h} and \mathbf{g}_{ED} , respectively. Note that $\text{Re}(h_t)$ and $\text{Im}(h_t)$ are from the same distribution $\mathcal{N}(0, \frac{1}{2}\sigma_h^2)$, and $\text{Re}(g_{\text{ED},t})$ and $\text{Im}(g_{\text{ED},t})$ are from the distribution $\mathcal{N}(0, \frac{1}{2}\sigma_h^2 - D)$. Assuming $\frac{B}{2}$ bits are used to quantize each of $\text{Re}(h_t)$ and $\text{Im}(h_t)$ for all t , by

rate-distortion theory for i.i.d. Gaussian sources [39], we can achieve the rate-distortion bound

$$\begin{aligned} E \left[\{\text{Re}(h_t) - \text{Re}(g_{\text{ED},t})\}^2 \right] &= E \left[\{\text{Im}(h_t) - \text{Im}(g_{\text{ED},t})\}^2 \right] \\ &= D \end{aligned}$$

as $M_t \rightarrow \infty$. Thus, by the weak law of large numbers, the following convergences hold²

$$\begin{aligned} \frac{1}{M_t} \|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2 &\xrightarrow{P} 2E \left[\{\text{Re}(h_t) - \text{Re}(g_{\text{ED},t})\}^2 \right] = 2D, \\ \frac{1}{M_t} \|\mathbf{h}\|_2^2 &\xrightarrow{P} 2E \left[\{\text{Re}(h_t)\}^2 \right] = \sigma_h^2, \\ \frac{1}{M_t} \|\mathbf{g}_{\text{ED}}\|_2^2 &\xrightarrow{P} 2E \left[\{\text{Re}(g_{\text{ED},t})\}^2 \right] = \sigma_h^2 - 2D \end{aligned}$$

as $M_t \rightarrow \infty$. Moreover, $\left| \frac{\mathbf{h}^H \mathbf{g}_{\text{ED}}}{M_t} \right|^2$ can be lower bounded as

$$\begin{aligned} \left| \frac{\mathbf{h}^H \mathbf{g}_{\text{ED}}}{M_t} \right|^2 &\geq \left(\frac{\text{Re}(\mathbf{h}^H \mathbf{g}_{\text{ED}})}{M_t} \right)^2 \\ &= \left(\frac{\|\mathbf{h}\|_2^2 + \|\mathbf{g}_{\text{ED}}\|_2^2 - \|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2}{2M_t} \right)^2 \\ &\xrightarrow{P} (\sigma_h^2 - 2D)^2. \end{aligned}$$

Then, the normalized beamforming gain loss relative to the unquantized beamforming case is bounded as

$$\begin{aligned} d_c^2(\mathbf{h}, \mathbf{g}_{\text{ED}}) &= 1 - \frac{|\mathbf{h}^H \mathbf{g}_{\text{ED}}|^2}{\|\mathbf{h}\|_2^2 \|\mathbf{g}_{\text{ED}}\|_2^2} \leq \frac{2D}{\sigma_h^2} = 2^{-B}, \\ d_c^2(\mathbf{h}, \mathbf{g}_{\text{ED}}) &\stackrel{(a)}{\geq} 2^{-\frac{BM_t}{M_t-1}} \end{aligned}$$

where (a) follows from the optimality of the RVQ codebook in large asymptotic regime [14]. As $M_t \rightarrow \infty$, the lower bound of $d_c^2(\mathbf{h}, \mathbf{g}_{\text{ED}})$ converges to the upper bound 2^{-B} , which finishes the proof. ■

Note that the loss in (10) is asymptotically the same as that of the RVQ codebook in (3). Since the RVQ codebook is known to be asymptotically optimal as $M_t \rightarrow \infty$ (fixing the number of bits per antenna) [14], we conclude that coherent Euclidean distance quantization as in (9) with a rich, rotationally invariant constellation such as a Gaussian codebook \mathcal{G} , is also an asymptotically optimal way to quantize the channel vector \mathbf{h} . Of course, in practice, for finite constellations and number of antennas, we must “align” the codewords \mathbf{g}_i with the channel \mathbf{h} , using parallel branches with different amplitude scaling α and phase rotations θ as in (8), prior to computing the Euclidean metric, in order to maximize the beamforming gain.

We also note that the use of nontrivial codes is implicit in Theorem 1, hence the uncoded constellations employed in [35] do not achieve optimal quantization performance. The constellation expansion employed in the NTCQ schemes considered here is required to approach optimal performance.

We now provide a *non-asymptotic* result regarding the chordal distances associated with Grassmannian line packing

²Let $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ and $\mu = E[X_i]$ for all i . We say \bar{X}_n converges to μ in probability as $\bar{X}_n \xrightarrow{P} \mu$ for $n \rightarrow \infty$ when $\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0$ for any $\epsilon > 0$.

(GLP) attained by codebooks optimized using Euclidean metrics. Let $N = 2^{B_{\text{tot}}}$ and $\mathcal{U}_{M_t}^N \in \mathbb{C}^{M_t \times N}$ denote the set of $M_t \times N$ complex matrices with unit vector columns. To minimize the average quantization error of (8) or (9) in Euclidean space with a fixed codebook \mathcal{C} , we have to maximize the minimum Euclidean distance between all possible codeword pairs

$$d_{E,\min}^2(\mathcal{C}) \triangleq \min_{1 \leq k < l \leq N} d_E^2(\mathbf{c}_k, \mathbf{c}_l)$$

where $d_E(\mathbf{x}, \mathbf{y}) \triangleq \|\mathbf{x} - \mathbf{y}\|_2$, and $\{\mathbf{c}_i\}_{i=1}^N$ are column vectors of \mathcal{C} . Let \mathcal{C}_{ED} denote an optimized Euclidean distance (ED) codebook that maximizes the minimum Euclidean distance as

$$\mathcal{C}_{\text{ED}} = \operatorname{argmax}_{\mathcal{C} \in \mathcal{U}_{M_t}^N} d_{E,\min}^2(\mathcal{C}).$$

On the other hand, beamforming codebooks are ideally designed for i.i.d. Rayleigh fading channels to maximize the minimum chordal distance between codewords as

$$d_{c,\min}^2(\mathcal{C}) \triangleq \min_{1 \leq k < l \leq N} d_c^2(\mathbf{c}_k, \mathbf{c}_l),$$

and a GLP codebook is given as [11], [13]

$$\mathcal{C}_{\text{GLP}} = \operatorname{argmax}_{\mathcal{C} \in \mathcal{U}_{M_t}^N} d_{c,\min}^2(\mathcal{C}).$$

Note that the optimization metrics of \mathcal{C}_{GLP} and \mathcal{C}_{ED} are different, the former is the chordal distance and the latter is the Euclidean distance. The following lemma shows the relation of the two metrics.

Lemma 1. *For any two unit vectors \mathbf{x} and \mathbf{y} , the squared chordal distance between \mathbf{x} and \mathbf{y} is upper bounded by a function of their Euclidean distance as*

$$\begin{aligned} d_c^2(\mathbf{x}, \mathbf{y}) &\leq 1 - \left(1 - \frac{1}{2}d_E^2(\mathbf{x}, \mathbf{y})\right)^2 \\ &= d_E^2(\mathbf{x}, \mathbf{y}) - \frac{1}{4}d_E^4(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Proof: Let us define $d_\theta^2(\mathbf{x}, \mathbf{y})$ as

$$\begin{aligned} d_\theta^2(\mathbf{x}, \mathbf{y}) &\triangleq \min_{\theta \in [0, 2\pi)} d_E^2(\mathbf{x}, e^{j\theta}\mathbf{y}) \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2 \max_{\theta \in [0, 2\pi)} \operatorname{Re}\{e^{j\theta}\mathbf{x}^H\mathbf{y}\} \\ &= 2 - 2|\mathbf{x}^H\mathbf{y}| \leq d_E^2(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Then, the squared chordal distance of \mathbf{x} and \mathbf{y} is upper bounded as

$$\begin{aligned} d_c^2(\mathbf{x}, \mathbf{y}) &= 1 - |\mathbf{x}^H\mathbf{y}|^2 \\ &= 1 - \left(1 - \frac{1}{2}d_\theta^2(\mathbf{x}, \mathbf{y})\right)^2 \\ &\leq 1 - \left(1 - \frac{1}{2}d_E^2(\mathbf{x}, \mathbf{y})\right)^2, \end{aligned}$$

which finishes the proof. \blacksquare

Moreover, Lemma 1 can be directly extended to the following corollary.

Corollary 1. *The minimum chordal distance of \mathcal{C}_{ED} , $d_{c,\min}^2(\mathcal{C}_{\text{ED}})$, is upper bounded by the minimum Euclidean distance of \mathcal{C}_{ED} , $d_{E,\min}^2(\mathcal{C}_{\text{ED}})$ as*

$$d_{c,\min}^2(\mathcal{C}_{\text{ED}}) \leq d_{E,\min}^2(\mathcal{C}_{\text{ED}}).$$

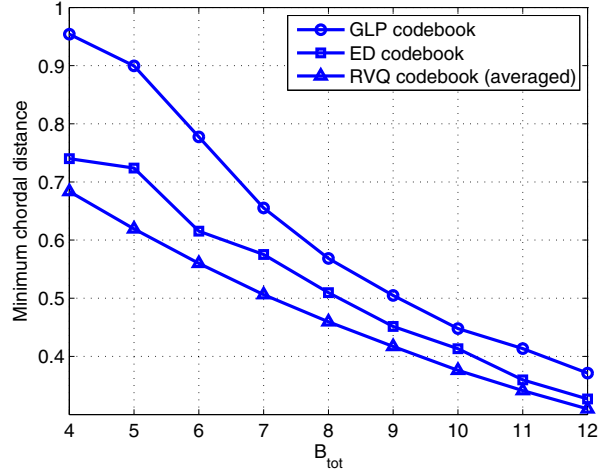


Fig. 2. The minimum chordal distances of different codebooks with $M_t = 8$. GLP and Euclidean distance (ED) codebook are numerically optimized according to their metrics, while the minimum distance of RVQ codebook is averaged over 1000 different RVQ codebooks.

Although Corollary 1 does not say that \mathcal{C}_{ED} maximizes the minimum chordal distance between its codewords, \mathcal{C}_{ED} is expected to have a *good* chordal distance property. We verify this by simulation with numerically optimized \mathcal{C}_{GLP} and \mathcal{C}_{ED} in Fig. 2. It is shown that the minimum chordal distance of \mathcal{C}_{ED} is larger than the (averaged) minimum chordal distance of the RVQ codebook for all B_{tot} values.

III. NONCOHERENT TRELLIS-CODED QUANTIZATION (NTCQ)

A. Euclidean Distance Codebook Design

The observations in the preceding section provide the following practical guidelines for quantization on the Grassmann manifold: (a) find a good codebook in Euclidean space whose structure permits efficient encoding (or, equivalently, find a good, efficiently decodable channel code); (b) use parallel versions of the Euclidean encoder with different amplitude scalings and phase rotations, and choose the best output (or, equivalently, implement block noncoherent decoding efficiently with a number of parallel coherent decoders). The proposed NTCQ emerges naturally from application of these guidelines.

NTCQ relies on trellis-coded quantization (TCQ) which was originally proposed in [29], exploiting the functional duality between source coding and channel coding to leverage the well-known trellis-coded modulation (TCM) channel codes designed for coherent communication over AWGN channels [40]. TCM integrates the design of convolutional codes with modulation to maximize the minimum Euclidean distance between modulated codewords. This is done by coding over partitions of the source constellation. Let \mathcal{C}_{TCM} denote a fixed codebook with N codewords generated by a TCM channel code. Then \mathcal{C}_{TCM} can be mathematically expressed as

$$\mathcal{C}_{\text{TCM}} = \operatorname{argmax}_{\mathcal{C} \in \mathcal{V}_{M_t}^N} d_{E,\min}^2(\mathcal{C})$$

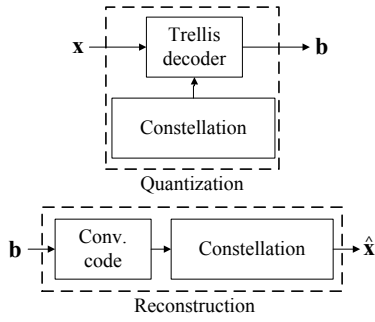


Fig. 3. Quantization and reconstruction processes for a Euclidean distance quantizer using trellis-coded quantization (TCQ).

where $\mathcal{V}_{M_t}^N \subset \mathcal{U}_{M_t}^N$ is the set of $M_t \times N$ complex matrices generated by a given trellis structure with a finite number of constellation points of interest for entries of the matrix. Note that \mathcal{C}_{TCM} is a Euclidean distance codebook within a given set $\mathcal{V}_{M_t}^N$. Thus, \mathcal{C}_{TCM} is expected to have a *good* chordal distance property as well.

In TCQ, the decoder and encoder of TCM are used to quantize and reconstruct a given source, respectively. From Fig. 3, we see that the TCQ system consists of a source constellation, a trellis-based decoder (for source quantization), and a convolutional encoder (for source reconstruction). Quantization is performed by passing a source vector $\mathbf{x} \in \mathbb{C}^N$ through a trellis-based optimization whose goal is to minimize a mean square error distortion between the quantized output and the source message input. The additive structure of the square of Euclidean distance implies that the Viterbi algorithm can be employed to efficiently search for a codebook vector that minimizes the Euclidean distance from a given source vector as

$$\mathbf{c}_{\text{opt}} = \underset{\mathbf{c}_i \in \mathcal{C}_{\text{TCM}}}{\text{argmin}} \|\mathbf{x} - \mathbf{c}_i\|_2^2, \quad (11)$$

which is then mapped to a binary sequence $\mathbf{b} = \text{bin}(\text{opt})$. The quantized source vector $\hat{\mathbf{x}}$ is reconstructed by passing the binary sequence \mathbf{b} into the convolutional encoder and mapping the binary output of the convolutional encoder to points on the source constellation (as if modulating the signal). Due to the linearity of the convolutional code, each unique binary sequence \mathbf{b} represents a unique quantized vector $\hat{\mathbf{x}}$.

NTCQ adopts TCQ to quantize CSI. Note that (11) is the same optimization problem as (8) with a given $\alpha \in \mathbb{A} = \{\alpha_1, \alpha_2, \dots, \alpha_{K_\alpha}\}$ and $\theta \in \Theta = \{\theta_1, \theta_2, \dots, \theta_{K_\theta}\}$. Thus, the minimization (8) can be performed using $K_\alpha \cdot K_\theta$ parallel instances of the Viterbi algorithm. This is the same paradigm proposed as in TCQ except for the search over α and θ parameters; due to the presence of these terms, the process is coined *noncoherent trellis-coded quantization*. Note that with PSK constellations, we can set $\alpha = 1$ because all the candidate beamforming vectors \mathbf{c}_i 's have the same norm.

We explain the implementation of NTCQ with 8PSK and 16QAM constellations next (we also report results for QPSK, but do not describe the corresponding NTCQ procedure, since it is similar to that for 8PSK). Before explaining the actual implementation, it should be pointed out that, because of the inherited TCM structure, the number of constellation points is

TABLE I
MAPPING OF QUANTIZING BITS/ENTRY (B) AND CONSTELLATIONS.

B	1 bit/entry	2 bits/entry	3 bits/entry
Constellation	QPSK	8PSK	16QAM

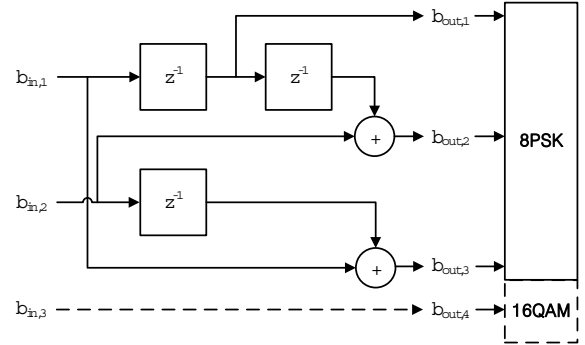


Fig. 4. This rate 2/3 convolutional code corresponds to the trellis in Fig. 6. In the figure, the smaller the index the less significant the bit, e.g., $b_{\text{in},1}$ is the least significant input bit and $b_{\text{in},3}$ is the most significant input bit.

larger than 2^B in NTCQ where B is the number of quantization bits per channel entry. We explicitly list the relationship between B and the constellations in Table I. This issue will become clear as we explain the 8PSK implementation.

B. NTCQ with 8PSK (2 bits/entry)

We adopt the rate 2/3 convolutional code in [40], as shown in Fig. 4. The source constellation is assumed to be 8PSK as in Fig. 5. Note that all constellation points are normalized with the number of transmit antennas M_t .

The construction of the feedback sequence is done using a trellis decoder. As is done in traditional decoding of convolutional codes, the encoding process is represented using a trellis showing the relationship between states of the encoder along with input and output transitions. The trellis with input/output state transitions corresponding to the convolutional code in Fig. 4 is shown in Fig. 6.

We select candidate beamforming vectors using an M_t -stage trellis where each stage selects an entry in each of the candidate vectors. Thus, each path through the trellis corresponds to a unique candidate beamforming vector. It is important to note that there are only four state-transitions from any of the eight states in Fig. 6. Each transition is mapped to one point of the 8PSK constellation. Therefore, even though the source constellation is 8PSK, each element of $\hat{\mathbf{h}}$ is quantized with one of the QPSK subconstellations marked by black or white circles in Fig. 5, which results in 2 bits quantization per entry as shown in Table I.

The path choices are enumerated with binary labels, and each path also corresponds to a unique binary sequence. The candidate vector or path that is chosen for output is the one that optimizes the given path metric. The path metric is chosen to reflect the desired Euclidean distance minimization regarding codeword \mathbf{c}_i in (8) for a given α and θ . The output of the quantization is the binary sequence corresponding to the best candidate path.

Each transition from each state at the t^{th} stage, $s_t \in \{1, 2, \dots, S\}$, in the trellis to a state at the $(t+1)^{\text{th}}$ stage,

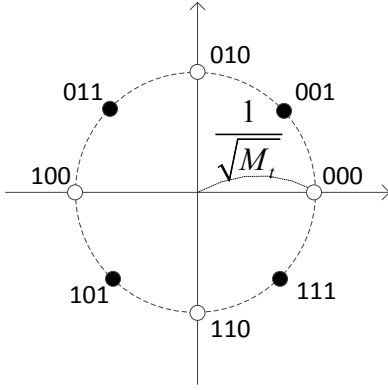


Fig. 5. 8PSK constellation points used in NTCQ are labeled with binary sequences.

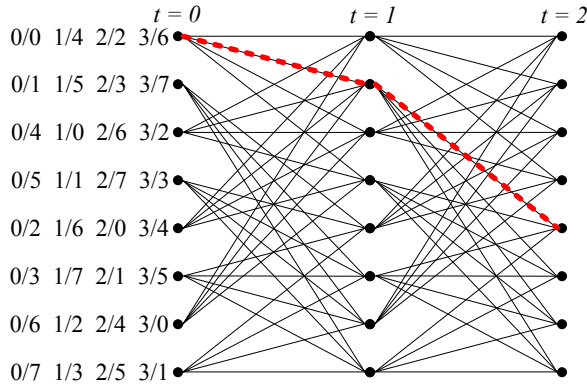


Fig. 6. The Ungerboeck trellis with $S = 8$ states corresponding to the convolutional encoder in Fig. 4. The input/output relations using decimal numbers correspond to state transitions from the top to bottom. The example path $\mathbf{p}_2 = [1, 2, 5]$ that corresponds to binary input sequence $[01, 00]^T$ (or decimal input $[1, 0]^T$) and binary output sequence $[100, 001]^T$ (or decimal output $[4, 1]^T$) is highlighted.

s_{t+1} , corresponds to a point in the source constellation. For example, a transition from state 4 to state 8 corresponds to the binary output sequence 011 which corresponds to the constellation point $\frac{1}{\sqrt{2M_t}}(-1 + j)$ in Fig. 5. Note that, in this setup, a single entry is chosen at each stage where it is possible to choose more; this is done by using intermediate codebooks for each stage of the trellis. For more details on this method and the design of the codebooks, the reader is referred to [31].

To optimize over the trellis, the first task is to define a path metric. Let \mathbf{p}_t be a partial path, or a sequence of states, up to the stage t . For example, the path $\mathbf{p}_2 = [1, 2, 5]$ using state indices is highlighted in Fig. 6. Also, define the two functions $\text{in}(\cdot)$ and $\text{out}(\cdot)$ such that $\text{in}(\mathbf{p}_t)$ outputs the binary input sequence corresponding to path \mathbf{p}_t , and $\text{out}(\mathbf{p}_t)$ gives the sequence of output constellation points corresponding to the path \mathbf{p}_t . Again, using the sample path \mathbf{p}_2 in Fig. 6, we can see that

$$\text{in}(\mathbf{p}_2) = [01, 00]^T, \quad \text{out}(\mathbf{p}_2) = \frac{1}{\sqrt{M_t}} \left[-1, \frac{1}{\sqrt{2}}(1 + j) \right]^T.$$

With these definitions, we can define the path metric, $m(\cdot)$, as

$$m(\mathbf{p}_t, \theta) = \|\bar{\mathbf{h}}_t - e^{j\theta} \text{out}(\mathbf{p}_t)\|_2^2,$$

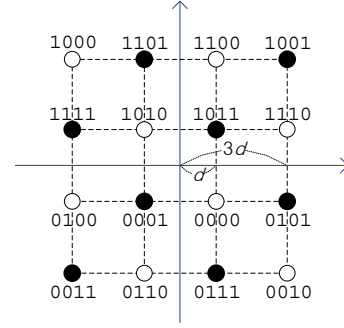


Fig. 7. 16QAM constellation points used in NTCQ are labeled with binary sequences.

where $\theta \in [0, 2\pi)$ and $\bar{\mathbf{h}}_t$ is the vector created by truncating of normalized MISO channel vector $\bar{\mathbf{h}}$ to the first t entries. Note that $\alpha = 1$ because all constellation points have the same magnitude in the 8PSK case. It is easy to check that minimizing over the path metric will minimize the Euclidean distance. It is also important to notice that the path metric can be written recursively as

$$m(\mathbf{p}_t, \theta) = m(\mathbf{p}_{t-1}, \theta) + |\bar{h}_t - e^{j\theta} \text{out}([p_{t-1} \ p_t]^T)|^2,$$

where \bar{h}_t and p_t are the t^{th} entry of $\bar{\mathbf{h}}$ and \mathbf{p}_t , respectively. The above path metric can be efficiently computed via the Viterbi algorithm. The path metric is computed in parallel for each quantized value of $\theta \in \Theta = \{\theta_1, \theta_2, \dots, \theta_{K_\theta}\}$. Then the best path \mathbf{p}_{best} and the phase θ_{best} that minimize the path metric can be found as

$$\min_{\theta \in \Theta} \min_{\mathbf{p}_{M_t} \in \mathbb{P}_{M_t}} m(\mathbf{p}_{M_t}, \theta)$$

where \mathbb{P}_{M_t} denotes all possible paths up to stage M_t . Finally, the beamforming vector \mathbf{f} is calculated as

$$\mathbf{c}_{\text{opt}} = \text{out}(\mathbf{p}_{\text{best}}), \quad \mathbf{f} = \frac{\mathbf{c}_{\text{opt}}}{\|\mathbf{c}_{\text{opt}}\|_2}.$$

Note that $\|\mathbf{c}_{\text{opt}}\|_2 = 1$ for 8PSK; therefore $\mathbf{f} = \mathbf{c}_{\text{opt}}$.

It is important to point out that minimizing over θ only increases the complexity of quantization, not the feedback overhead because the transmitter does not have to know the value of θ_{best} that minimizes the path metric during the beamforming vector reconstruction process. However, there is additional feedback overhead with NTCQ. Since we test all paths in the trellis, the transmitter has to know the starting state of \mathbf{p}_{best} , which causes additional $\log_2 S$ bits of feedback overhead where S is the number of states in the trellis. Therefore, the total feedback overhead is

$$B_{\text{tot}} = BM_t + \log_2 S.$$

The additional feedback overhead $\log_2 S$ bits can vary depending on the trellis used in NTCQ.

C. NTCQ with 16QAM (3 bits/entry)

For the 16QAM constellation, the rate 3/4 convolution encoder is shown in Fig. 4. The source constellation is shown in Fig. 7 where $d = \frac{\Delta}{2\sqrt{M_t}}$ with $\Delta = \sqrt{\frac{6}{M-1}}$ with $M = 16$ to have $E[\|\mathbf{c}_i\|_2^2] = 1$ where expectation is taken over

\mathbf{c}_i assuming all constellation points are selected with equal probability.

The procedure of NTCQ using 16QAM is basically the same as the 8PSK case. The difference arising for 16QAM is that we have to take α into account during the path metric computation as

$$m(\mathbf{p}_t, \alpha, \theta) = \|\bar{\mathbf{h}}_t - \alpha e^{j\theta} \text{out}(\mathbf{p}_t)\|_2^2 \quad (12)$$

where $\theta \in \Theta = \{\theta_1, \theta_2, \dots, \theta_{K_\theta}\}$ and $\alpha \in \mathbb{A} = \{\alpha_1, \alpha_2, \dots, \alpha_{K_\alpha}\}$. Similar to the 8PSK case, additional $\log_2 S$ feedback bits are needed to indicate the starting state of \mathbf{p}_{best} to the transmitter in the 16QAM case.

D. Complexity

NTCQ relies on a trellis search to quantize the beamforming vector, and the trellis search is performed by the Viterbi algorithm. In each state transition of the trellis, one channel entry is quantized with one of 2^B constellation points. This computation is performed for S states in each state transition (stage) and there are M_t state transitions in total. Thus, the complexity of the Viterbi algorithm becomes $O(2^B S M_t)$.

The Viterbi algorithm has to be executed $K_\theta \cdot K_\alpha$ times in NCTQ, which gives the overall complexity of $O(K_\theta K_\alpha 2^B S M_t)$. In the limit of large M_t , Theorem 1 tells us that we can get away with $K_\theta \rightarrow 1$ and $K_\alpha \rightarrow 1$ without performance loss. However, even for moderate values of M_t , our results in Section V-A show that small values of K_θ and K_α can be employed with minimal performance degradation. The key aspect to note is the linear scaling of complexity with the number of transmit antennas M_t , which makes NTCQ particularly attractive for massive MIMO systems for which conventional look-up based approaches are computationally infeasible.

E. Variations of NTCQ

We can also construct several variations of NTCQ with minor tradeoffs between the total number of feedback bits, B_{tot} , and performance. We explain one of the variations briefly below.

- *Variation: Fixing the starting state for the trellis search.*

Because NTCQ searches paths which start from every possible state in the first stage in the trellis, we need an additional $\log_2 S$ bits of feedback overhead to indicate the starting state of \mathbf{p}_{best} . One variation is to fix the first state to eliminate these additional bits, so that the total feedback overhead incurred is exactly $B M_t$ bits. We do incur a small performance loss by doing this, since allowing starting from different states effectively leads to considering more possible values of the scaling parameters α and θ . However, this loss becomes negligible as M_t gets large (consistent with Theorem 1).

For other variations, we can fix the first entry of \mathbf{c}_{opt} to a constant in the trellis search or adopt a tail-biting convolutional code.

IV. ADVANCED NTCQ EXPLOITING CHANNEL CORRELATIONS

In practice, channels are temporally and/or spatially correlated. In this section, we propose advanced NTCQ schemes

that exploit these correlations to improve the performance or reduce the feedback overhead.

A. Differential Scheme for Temporally Correlated Channels

A useful model of this correlation is the first-order Gauss-Markov process [41]

$$\mathbf{h}[k] = \eta \mathbf{h}[k-1] + \sqrt{1-\eta^2} \mathbf{g}[k]$$

where $\mathbf{g}[k] \in \mathbb{C}^{M_t}$ denotes the process noise, which is modeled as having i.i.d. entries distributed with $\mathcal{CN}(0, 1)$. We assume that the initial state $\mathbf{h}[0]$ is independent of $\mathbf{g}[k]$ for all k . The temporal correlation coefficient η ($0 \leq \eta \leq 1$) represents the correlation between elements $h_t[k-1]$ and $h_t[k]$ where $h_t[k]$ is the t^{th} entry of $\mathbf{h}[k]$.

If η is close to one, two consecutive channels are highly correlated and the difference between the previous channel $\mathbf{h}[k-1]$ and the current channel $\mathbf{h}[k]$ might be small. Differential codebooks in [19]–[26] utilize this property to reduce the channel quantization error with an assumption that both the transmitter and the receiver know η perfectly. Most of the previous literature, however, focused on the case with a fixed and small number of transmit antennas and moderate feedback overhead, e.g., $M_t = 4$ and $B_{\text{tot}} = 4$. Therefore, we have to come up with a new differential feedback scheme to accommodate massive MIMO with large feedback overhead.

We denote $\mathbf{f}[k-1]$ as the quantized beamforming vector at block $k-1$ and

$$\mathbf{f}_{\text{opt}}[k] = \frac{\mathbf{h}[k]}{\|\mathbf{h}[k]\|_2}$$

as the unquantized optimal beamforming vector at time k . In our differential NTCQ scheme, instead of quantizing $\mathbf{h}[k]$ directly at time k , the receiver quantizes $\mathbf{f}_{\text{diff}}[k]$ which is given as

$$\mathbf{f}_{\text{diff}}[k] = (\mathbf{I}_{M_t} - \mathbf{f}[k-1]\mathbf{f}^H[k-1]) \mathbf{f}_{\text{opt}}[k].$$

Note that $\mathbf{f}_{\text{diff}}[k]$ is a projection of $\mathbf{f}_{\text{opt}}[k]$ to the null space of $\mathbf{f}[k-1]$. We let $\hat{\mathbf{f}}_{\text{diff}}[k]$ denote the quantized version of $\mathbf{f}_{\text{diff}}[k]$ by NTCQ with $\|\hat{\mathbf{f}}_{\text{diff}}[k]\|_2^2 = 1$. The receiver then constructs candidate beamforming vectors $\mathbf{f}_{\bar{\alpha}, \bar{\theta}}$ with weights $\bar{\alpha} \in \bar{\mathbb{A}} = \{\bar{\alpha}_1, \dots, \bar{\alpha}_{K_\alpha}\}$ and $\bar{\theta} \in \bar{\Theta} = \{\bar{\theta}_1, \dots, \bar{\theta}_{K_\theta}\}$ as

$$\mathbf{f}_{\bar{\alpha}, \bar{\theta}} = \frac{\eta \mathbf{f}[k-1] + \bar{\alpha} e^{j\bar{\theta}} \sqrt{1-\eta^2} \hat{\mathbf{f}}_{\text{diff}}[k]}{\left\| \eta \mathbf{f}[k-1] + \bar{\alpha} e^{j\bar{\theta}} \sqrt{1-\eta^2} \hat{\mathbf{f}}_{\text{diff}}[k] \right\|_2}. \quad (13)$$

The receiver selects the optimal weights $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ by optimizing

$$\max_{\bar{\alpha} \in \bar{\mathbb{A}}} \max_{\bar{\theta} \in \bar{\Theta}} \left| \bar{\mathbf{h}}^H[k] \mathbf{f}_{\bar{\alpha}, \bar{\theta}} \right|^2, \quad (14)$$

and the final beamforming vector is given as

$$\mathbf{f}[k] = \mathbf{f}_{\bar{\alpha}_{\text{opt}}, \bar{\theta}_{\text{opt}}}.$$

To construct candidate beamforming vectors as in (13), we have to define sets of weights $\bar{\mathbb{A}}$ and $\bar{\Theta}$. It is easy to conclude that $\bar{\Theta} = [0, 2\pi)$ because the quantization process uses beamformer phase invariance. To derive the range of the set $\bar{\mathbb{A}}$, we make the following proposition.

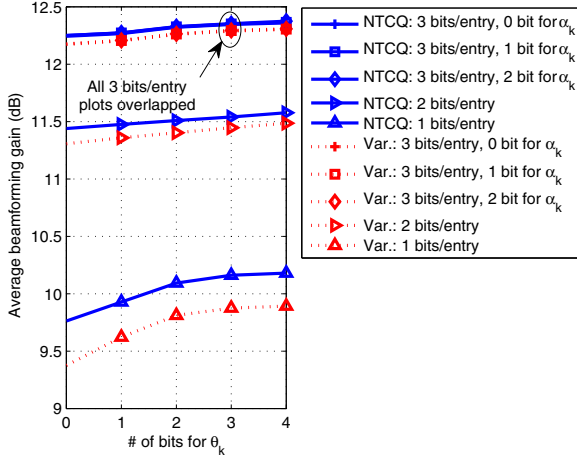


Fig. 8. $J_{\text{avg}}^{\text{dB}}$ vs. different quantization levels of θ_k and α_k with $M_t = 20$ in i.i.d. Rayleigh fading channels.

Proposition 1. When $\eta \rightarrow 1$, the range of $\bar{\alpha}$ can be set as

$$\frac{1 - \eta}{\sqrt{1 - \eta^2}} \leq \bar{\alpha} \leq \frac{1 + \eta}{\sqrt{1 - \eta^2}}. \quad (15)$$

Proof: First, we define $\mathbf{f}_{\bar{\alpha}, \bar{\theta}}^{\text{nom}}$ as the numerator of (13) as

$$\mathbf{f}_{\bar{\alpha}, \bar{\theta}}^{\text{nom}} = \eta \mathbf{f}[k-1] + \bar{\alpha} e^{j\bar{\theta}} \sqrt{1 - \eta^2} \hat{\mathbf{f}}_{\text{diff}}[k].$$

Then, the norm square of $\mathbf{f}_{\bar{\alpha}, \bar{\theta}}^{\text{nom}}$ becomes

$$\begin{aligned} \|\mathbf{f}_{\bar{\alpha}, \bar{\theta}}^{\text{nom}}\|_2^2 &= \eta^2 + \bar{\alpha}^2(1 - \eta^2) \\ &\quad + 2\bar{\alpha}\sqrt{1 - \eta^2} \text{Re} \left\{ e^{j\bar{\theta}} \mathbf{f}^H[k-1] \hat{\mathbf{f}}_{\text{diff}}[k] \right\}. \end{aligned}$$

Because $-1 \leq \text{Re} \left\{ e^{j\bar{\theta}} \mathbf{f}^H[k-1] \hat{\mathbf{f}}_{\text{diff}}[k] \right\} \leq 1$, we have

$$\left(\eta - \bar{\alpha}\sqrt{1 - \eta^2} \right)^2 \leq \|\mathbf{f}_{\bar{\alpha}, \bar{\theta}}^{\text{nom}}\|_2^2 \leq \left(\eta + \bar{\alpha}\sqrt{1 - \eta^2} \right)^2. \quad (16)$$

Note that $\mathbf{f}^H[k-1] \hat{\mathbf{f}}_{\text{diff}}[k] \approx 0$ with a good quantizer. Moreover, with the assumption of a slowly varying channel which is typically assumed in the differential codebook literature, we approximate $\eta \approx 1$. Then we have $\|\mathbf{f}_{\bar{\alpha}, \bar{\theta}}^{\text{nom}}\|_2^2 = 1$, and plugging this into (16) gives the range of $\bar{\alpha}$ in (15). ■

Note that the range in (15) can be further optimized numerically. In Section V-B, we set $\frac{1 - \eta}{\sqrt{1 - \eta^2}} \leq \bar{\alpha} \leq \frac{1 + \eta}{3\sqrt{1 - \eta^2}}$ for simulation. Once the receiver selects the optimal weights $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ by (14), it feeds back $\hat{\mathbf{f}}_{\text{diff}}[k]$, $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ to the transmitter over the feedback link and the transmitter reconstructs $\mathbf{f}[k]$ as in (13). Additional feedback overhead caused by $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ can be very small compared to the feedback overhead for $\hat{\mathbf{f}}_{\text{diff}}[k]$. Simulation indicates that 1 bit for $\bar{\alpha}_{\text{opt}}$ and 3 bits for $\bar{\theta}_{\text{opt}}$ is sufficient to have near-optimal performance in a low mobility scenario.

B. Adaptive Scheme for Spatially Correlated Channels

If the transmit antennas are closely spaced, which is likely for a massive MIMO scenario, channels tend to be spatially correlated and can be modeled as

$$\mathbf{h}[k] = \mathbf{R}^{\frac{1}{2}} \mathbf{h}_w[k]$$

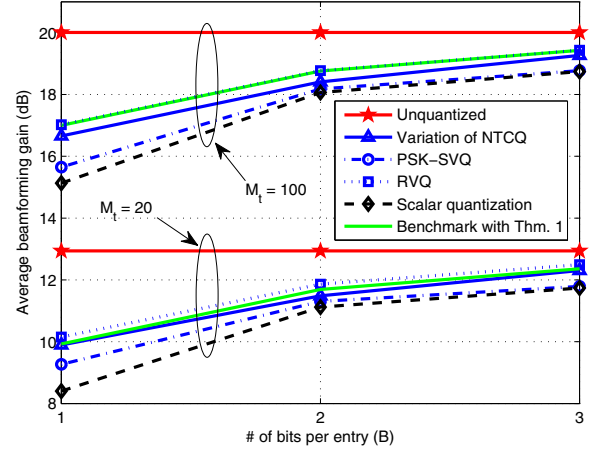


Fig. 9. $J_{\text{avg}}^{\text{dB}}$ vs. B with $M_t = 20$ and 100 in i.i.d. Rayleigh fading channels. PSK-SVQ is from [35]. All limited feedback schemes have the same B_{tot} .

where $\mathbf{h}_w[k]$ is an uncorrelated MISO channel vector with i.i.d. complex Gaussian entries and $\mathbf{R} = E[\mathbf{h}[k]\mathbf{h}^H[k]]$ is a correlation matrix of the channel where expectation is taken over k . We assume that \mathbf{R} is a full-rank matrix. For spatially correlated MISO channels, codebook skewing methods were proposed in [16]–[18] such that codewords in a VQ codebook are rotated and normalized with respect to \mathbf{R} to quantize only the local space of the dominant eigenvector of \mathbf{R} . It was shown in [16]–[18] that this skewing method can significantly reduce the quantization error with the same feedback overhead. With NTCQ, however, there are no fixed VQ codewords for channel quantization which precludes the normal approach for skewing. Therefore, we propose the following method to mimic skewing with NTCQ for spatially correlated MISO channels.

We assume that both the transmitter and the receiver know \mathbf{R} in advance³. At the receiver side, $\mathbf{h}_w[k]$ is obtained by decorrelating $\mathbf{h}[k]$ with $\mathbf{R}^{-\frac{1}{2}}$, i.e.,

$$\mathbf{h}_w[k] = \mathbf{R}^{-\frac{1}{2}} \mathbf{h}[k].$$

Then the receiver quantizes $\mathbf{h}_w[k]$ with NTCQ and get $\hat{\mathbf{h}}_w[k]$. The receiver feeds back $\hat{\mathbf{h}}_w[k]$, and the transmitter reconstructs $\mathbf{f}[k]$ as

$$\mathbf{f}[k] = \frac{\mathbf{R}^{\frac{1}{2}} \hat{\mathbf{h}}_w[k]}{\left\| \mathbf{R}^{\frac{1}{2}} \hat{\mathbf{h}}_w[k] \right\|_2}.$$

This procedure effectively decouples the procedure of exploiting spatial correlation from that of quantization, while providing the same performance gain as standard skewing of fixed codewords.

V. PERFORMANCE EVALUATION

In this section, we present Monte-Carlo simulation results to evaluate the performance of NTCQ in i.i.d. channels, temporally correlated channels, and spatially correlated channels. In each scenario, we simulate the original NTCQ and its variation, differential NTCQ, and spatially adaptive NTCQ

³In practice, the transmitter can acquire an approximate knowledge of \mathbf{R} by averaging $\mathbf{f}[k]$, i.e., $\mathbf{R} \approx E[\mathbf{f}[k]\mathbf{f}^H[k]]$ where expectation is taken over k .

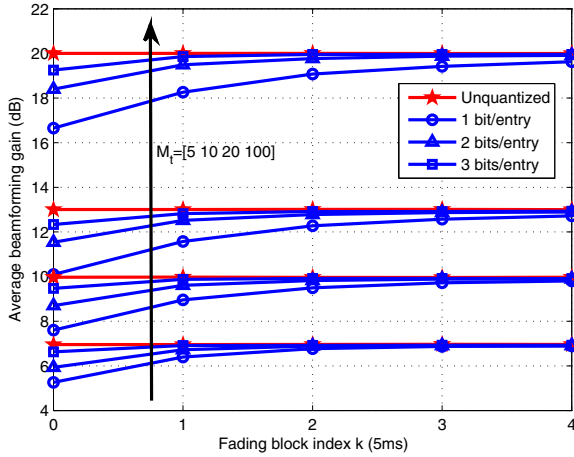


Fig. 10. $J_{\text{avg}}^{\text{dB}}$ vs. fading block index k with $v = 3\text{km}/h$ in temporally correlated channels. Without feedback delay.

explained in Sections III, IV-A, and IV-B, respectively. We use the average beamforming gain in dB scale

$$J_{\text{avg}}^{\text{dB}} = 10 \log_{10} (E[|\mathbf{h}^H \mathbf{f}|^2])$$

as a performance metric where the expectation is over \mathbf{h} .

A. i.i.d. Rayleigh fading Channels

For i.i.d. Rayleigh fading channels, $\mathbf{h}[k]$ is drawn from i.i.d. complex Gaussian entries (i.e., $\mathbf{h}[k] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$). In Fig. 8, we first plot $J_{\text{avg}}^{\text{dB}}$ of NTCQ and its variation in i.i.d. channels with $M_t = 20$ transmit antennas depending on different quantization levels for θ_k and α_k . Clearly, the variation of NTCQ gives strictly lower $J_{\text{avg}}^{\text{dB}}$ than the original NTCQ. Note that it is enough to have $K_\theta = 4$ (2 bits for θ_k) for 1 bit/entry (QPSK) to achieve near-maximal performance of NTCQ and its variation. Interestingly, we can fix $\alpha_k = 1$ with 3 bits/entry (16QAM) for NTCQ and its variation without having any performance loss. This is because when optimizing (12), it is likely to have $E[\|\mathbf{c}_{\text{opt}}\|_2^2] = 1$ since the objective variable is the normalized channel vector $\bar{\mathbf{h}}$ which has a unit norm, i.e., $\|\bar{\mathbf{h}}\|_2^2 = 1$. We fix $K_\theta = 16$ (4 bits for θ_k) for simulations afterward regardless of the number of bits per entry to have a fair comparison. We also fix $\alpha_k = 1$ for 3 bits/entry quantization.

In Fig. 9, we plot $J_{\text{avg}}^{\text{dB}}$ for variation of NTCQ (to have the same feedback overhead $B_{\text{tot}} = BM_t$ with the other limited feedback schemes) as a function of the number of quantization bits per entry, B , in i.i.d. Rayleigh channel realizations. We also plot $J_{\text{avg}}^{\text{dB}}$ for unquantized beamforming, RVQ, PSK-SVQ in [35], scalar quantization, and the benchmark from Theorem 1 which is given as $M_t(1 - 2^{-B})$ (in linear scale). The performance of RVQ is plotted using the analytical approximation in (3) as $M_t(1 - 2^{-\frac{B_{\text{tot}}}{M_t-1}})$ (in linear scale), because it is computationally infeasible to simulate when the number of feedback bits grows large. In scalar quantization, B bits are used to quantize only the phase, not the amplitude, of each channel entry because the phase is generally more important than the amplitude in beamforming [42].

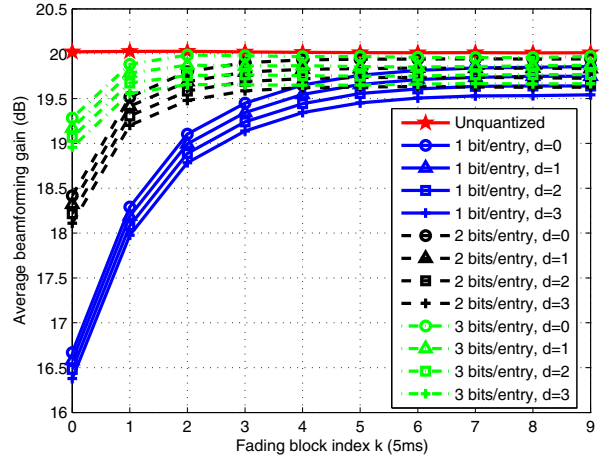


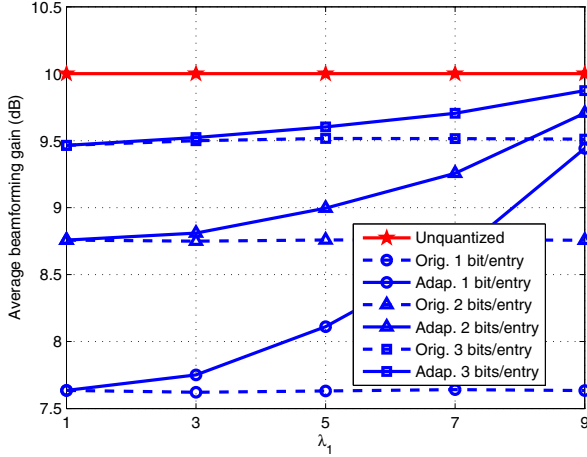
Fig. 11. $J_{\text{avg-delay}}^{\text{dB}}[d]$ vs. fading block index k with $M_t = 100$, d blocks of feedback delay, and $v = 3\text{km}/h$ in temporally correlated channels.

As the number of feedback bits increases, the gap between the unquantized case and all limited feedback schemes decreases as expected. RVQ gives the best performance among limited feedback schemes with the same number of feedback bits. However, the difference between $J_{\text{avg}}^{\text{dB}}$ for RVQ and variation of NTCQ is small for all B . The plots of the benchmark using Theorem 1 well approximate $J_{\text{avg}}^{\text{dB}}$ of NTCQ for all B and M_t , which shows the near-optimality of NTCQ. Note that variation of NTCQ achieves better $J_{\text{avg}}^{\text{dB}}$ than PSK-SVQ regardless of B and M_t , and the gap becomes larger as M_t increases. This gap comes from the coding gain of NTCQ. As shown in Table I, NTCQ can exploit 2^{B+1} constellation points while PSK-SVQ only utilizes 2^B constellation points with B bits quantization per entry. The coding gain of variation of NTCQ is around 0.25 to 1dB depending on M_t and B . Although we do not plot the performance of QAM-SVQ which relies on QAM constellations, it has the same structure as PSK-SVQ meaning that QAM-SVQ roughly experiences the same performance degradation compared to NTCQ.

B. Temporally Correlated Channels

To simulate the differential feedback schemes with the original NTCQ algorithm in temporally correlated channels, we adopt Jakes' model [43] to generate the temporal correlation coefficient $\eta = J_0(2\pi f_D \tau)$, where $J_0(\cdot)$ is the 0th order Bessel function of the first kind, f_D denotes the maximum Doppler frequency, and τ denotes the channel instantiation interval. We assume a carrier frequency of 2.5GHz and $\tau = 5\text{ms}$. We set the quantization level for the combiners $\bar{\theta}$ and $\bar{\alpha}$ in (13) as 3 bits and 1 bit, respectively, which causes 4 bits of additional feedback overhead.

In Fig. 10, we plot the performance of the proposed differential NTCQ feedback schemes with the velocity $v = 3\text{km}/h$ ($\eta = 0.9881$) assuming no feedback delay. The differential NTCQ schemes, even with 1 bit/entry quantization, achieve almost the same performance as unquantized beamforming regardless of M_t . Thus, if we can adjust the feedback overhead as a function of time, we can switch from NTCQ with 2 or 3 bits/entry quantization to 1bit/entry quantization in differential NTCQ to reduce the overall feedback overhead.


 Fig. 12. $J_{\text{avg}}^{\text{dB}}$ vs. λ_1 with $M_t = 10$ in spatially correlated channels.

To see the effect of feedback delay in temporally correlated channels, we simulate the $M_t = 100$ case with different numbers of delay d measured in fading blocks (one fading block corresponds to 5ms) in Fig. 11 such that

$$J_{\text{avg-delay}}^{\text{dB}}[d] = 10 \log_{10} (E[|\mathbf{h}^H[k+d]\mathbf{f}[k]|^2]) .$$

It is shown that the effect of feedback delay is negligible, i.e., around 0.1dB loss with one additional block delay for all cases, which confirms the practicality of the differential NTCQ scheme. Moreover, we can reduce the frequency of the feedback updates to reduce the total amount of feedback overhead without significant performance degradation when the velocity of the receiver is low.

C. Spatially Correlated Channels

To generate spatially correlated channels, we adopt the Kronecker model for the spatial correlation matrix \mathbf{R} which is given as $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ where \mathbf{U} and $\mathbf{\Sigma}$ are $M_t \times M_t$ eigenvector and diagonal eigenvalue matrices, respectively. The performance of the adaptive scheme will highly depend on the amount of spatial correlation. To see the effect of spatial correlation, we assume the eigenvalue matrix $\mathbf{\Sigma}$ has a structure given by

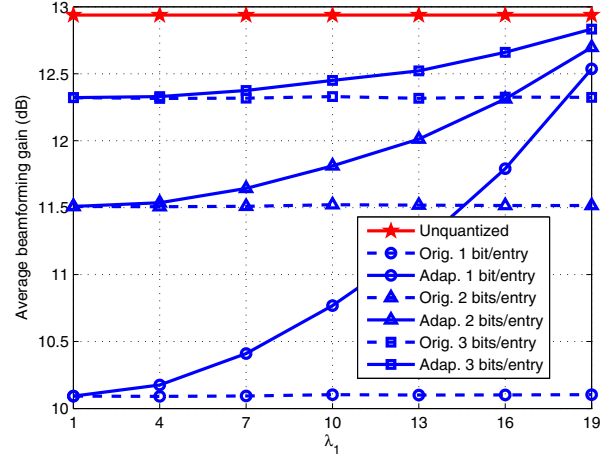
$$\mathbf{\Sigma} = \text{diag} \left\{ \lambda_1, \frac{M_t - \lambda_1}{M_t - 1}, \dots, \frac{M_t - \lambda_1}{M_t - 1} \right\}$$

where $1 \leq \lambda_1 < M_t$ is the dominant eigenvalue of \mathbf{R} . If λ_1 is small (large), the channels are loosely (highly) correlated in spatial domain. Note that channels are i.i.d. when $\lambda_1 = 1$.

In Fig. 12, and 13, we plot $J_{\text{avg}}^{\text{dB}}$ as a function of λ_1 for $M_t = 10$ and 20 cases. The performance of spatially adaptive NTCQ become closer to that of unquantized beamforming as λ_1 increases with the same feedback overhead as original NTCQ. This shows the effectiveness of the proposed adaptive NTCQ scheme for spatially correlated channels.

VI. CONCLUSIONS

In this paper, we have proposed an efficient channel quantization method for massive MIMO systems employing limited


 Fig. 13. $J_{\text{avg}}^{\text{dB}}$ vs. λ_1 with $M_t = 20$ in spatially correlated channels.

feedback beamforming. While the quantization criterion (maximization of beamforming gain or minimization of chordal distance) is associated with the Grassmann manifold, the key to the proposed NTCQ approach is to leverage efficient encoding (via the Viterbi algorithm) and codebook design (via TCQ) in Euclidean space. Efficient encoding relies on the mapping of quantization on the Grassmann manifold to noncoherent sequence detection and the near-optimal implementation of noncoherent detection using a bank of coherent detectors (i.e., Euclidean space quantizers). Standard rate-distortion theory and asymptotic results for RVQ tell us that good Euclidean codebooks should work well in Grassmannian space. Our numerical results show that the NTCQ provides better performance than uncoded schemes such as those considered in [35].

The advantages of NTCQ include flexibility and scalability in the number of channel coefficients: additional coefficients can be accommodated simply by increasing the blocklength, and the encoding complexity is linear in the number of transmit antennas. It can also be easily modified to take advantage of channel conditions such as temporal and spatial correlations. Our numerical results show that these advanced schemes can improve the performance significantly or reduce feedback overhead considerably depending on the system requirement.

While we have developed an efficient channel quantization method for massive MIMO systems, we note that limitations on feedback overhead would typically prevent scaling to an indefinitely large number of antennas. However, the feedback overhead may be reasonable for the moderately large number of antennas (32 to 64) expected in initial deployments [30], and NTCQ represents a computationally efficient approach to generating such feedback.

Finally, in order to make FDD massive MIMO practical, it is also crucial to develop scalable sounding schemes for channel estimation. Current sounding methods that transmit pilot signals from all transmit antennas using different time and/or frequency resources are not appropriate for massive MIMO systems because the pilot signals will dominate the downlink resources. Initial work on this topic was conducted in [44] and extended in [45].

REFERENCES

- [1] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis-coded quantization for massive MIMO limited feedback beamforming," *2013 UCSD Information Theory and Applications Workshop*.
- [2] J. Choi, D. J. Love, and U. Madhow, "Limited feedback in massive MIMO systems: exploiting channel correlations via noncoherent trellis-coded quantization," *Proc. 2013 Conference on Information Sciences and Systems*.
- [3] T. L. Marzetta, "How much training is required for multiuser MIMO?" *Proc. 2006 IEEE Asilomar Conference on Signals, Systems, and Computers*.
- [4] —, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
- [5] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: how many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
- [6] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
- [7] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, E. O. and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
- [8] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.
- [9] P. Ding, D. J. Love, and M. D. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3417–3428, Jul. 2007.
- [10] D. J. Love, R. W. Heath Jr., V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.
- [11] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.
- [12] 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, Jul. 2005.
- [13] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.
- [14] W. Santipach and M. L. Honig, "Capacity of multiple-antenna fading channel with quantized precoding matrix," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1218–1234, Mar. 2009.
- [15] C. K. 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.
- [16] D. J. Love and R. W. Heath Jr., "Grassmannian beamforming on correlated MIMO channels," *Proc. 2004 IEEE Global Telecommunications Conference*.
- [17] —, "Limited feedback diversity techniques for correlated channels," *IEEE Trans. Veh. Technol.*, vol. 55, no. 2, pp. 718–722, Mar. 2006.
- [18] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1853–1863, Mar. 2006.
- [19] B. Banister and J. Zeidler, "Feedback assisted transmission subspace tracking for MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 21, pp. 452–463, May 2003.
- [20] J. Yang and D. Williams, "Transmission subspace tracking for MIMO systems with low-rate feedback," *IEEE Trans. Commun.*, vol. 55, no. 8, pp. 1629–1639, Aug. 2007.
- [21] R. W. Heath Jr., T. Wu, and A. C. K. Soong, "Progressive refinement of beamforming vectors for high-resolution limited feedback," *EURASIP J. Advances Signal Process.*, vol. 2009, no. 6, Feb. 2009.
- [22] K. Huang, R. W. Heath Jr., and J. G. Andrews, "Limited feedback beamforming over temporally-correlated channels," *IEEE Trans. Signal Process.*, vol. 57, no. 5, pp. 1959–1975, May 2009.
- [23] D. Sacristan and A. Pascual-Iserte, "Differential feedback of MIMO channel gram matrices based on geodesic curves," *IEEE Trans. Wireless Commun.*, vol. 9, no. 12, pp. 3714–3727, Dec. 2010.
- [24] T. Kim, D. J. Love, and B. Clerckx, "MIMO system with limited rate differential feedback in slow varying channel," *IEEE Trans. Commun.*, vol. 59, no. 4, pp. 1175–1180, Apr. 2010.
- [25] J. Choi, B. Clerckx, N. Lee, and G. Kim, "A new design of polar-cap differential codebook for temporally/spatially correlated MISO channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 703–711, Feb. 2012.
- [26] J. Choi, B. Clerckx, and D. J. Love, "Differential codebook for general rotated dual-polarized MISO channels," *Proc. 2012 IEEE Global Telecommunications Conference*.
- [27] W. Santipach and M. L. Honig, "Asymptotic performance of MIMO wireless channels with limited feedback," in *Proc. 2003 IEEE Military Communications Conference*.
- [28] S. Lin and D. J. Costello, Jr., *Error Control Coding*. Prentice Hall, 2004.
- [29] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 82–93, Jan. 1990.
- [30] Y. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013.
- [31] C. K. Au-Yeung, D. J. Love, and S. Sanayei, "Trellis coded line packing: large dimensional beamforming vector quantization and feedback transmission," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1844–1853, Jun. 2011.
- [32] C. K. Au-Yeung, A. Jalali, and D. J. Love, "Insights into feedback and feedback signaling for beamformer design," *2009 UCSD Information Theory and Applications Workshop*.
- [33] C. K. Au-Yeung and S. Sanayei, "Enhanced trellis based vector quantization for coordinated beamforming," *Proc. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.
- [34] M. Xu, D. Guo, and M. K. Honig, "MIMO precoding with limited rate feedback: simple quantizers work well," *Proc. 2009 IEEE Global Telecommunications Conference*.
- [35] D. J. Ryan, I. V. L. Clarkson, I. B. Collings, D. Guo, and M. L. Honig, "QAM and PSK codebooks for limited feedback MIMO beamforming," *IEEE Trans. Commun.*, vol. 57, no. 4, pp. 1184–1196, Apr. 2009.
- [36] D. J. Ryan, I. B. Collings, and I. V. L. Clarkson, "GLRT-optimal noncoherent lattice decoding," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3773–3786, Jul. 2007.
- [37] W. Sweldens, "Fast block noncoherent decoding," *IEEE Commun. Lett.*, vol. 5, no. 4, pp. 132–134, Apr. 2001.
- [38] D. Warrier and U. Madhow, "Spectrally efficient noncoherent communication," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 652–668, Mar. 2002.
- [39] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.
- [40] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inf. Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.
- [41] R. H. Etkin and D. N. C. Tse, "Degree of freedom in some underspread MIMO fading channel," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1576–1608, Apr. 2006.
- [42] D. J. Love and R. W. Heath Jr., "Equal gain transmission in multiple-input multiple-output wireless systems," *IEEE Trans. Commun.*, vol. 51, no. 7, pp. 1102–1110, Jul. 2003.
- [43] J. G. Proakis, *Digital Communication*, 4th ed. McGraw-Hill, 2000.
- [44] D. J. Love, J. Choi, and P. Bidigare, "A closed-loop training approach for massive MIMO beamforming systems," *Proc. 2013 Conference on Information Sciences and Systems*.
- [45] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Topics Signal Process.*, submitted for publication. Available: <http://arxiv.org/abs/1309.7712>



Junil Choi (S'12) received his B.S. (with honors) and M.S. degrees from Seoul National University in Seoul, Korea in 2005 and 2007, respectively. He is currently working toward the Ph.D. degree in the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. From 2007 to 2011, he was a member of the technical staff at Samsung Advanced Institute of Technology (SAIT) and Samsung Electronics in Korea, where he contributed advanced codebook and feedback framework designs to 3GPP LTE-Advanced and

IEEE 802.16m standards. His research interests are in the design and analysis of adaptive communication and massive MIMO systems. Mr. Choi was a co-recipient of the 2008 Global Samsung Technical Conference Best Paper Award. Recently, he was awarded the Michael and Katherine Birk Fellowship from Purdue University in 2011; and the Korean Government Scholarship Program for Study Overseas in 2011–2012.



Zachary Chance (S'08-M'12) received the B.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 2007 and 2012, respectively. In the summer and fall of 2011, he was with the Naval Research Laboratory, Washington, D.C., and MIT Lincoln Laboratory, Lexington, MA. Since August 2012, he has been with MIT Lincoln Laboratory. His research interests include wireless communications, feedback systems, and tracking.

Dr. Chance is an active reviewer for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the *EURASIP Journal on Wireless Communications*. He received the Ross Fellowship from Purdue University along with the Frederic R. Miller Scholarship, Mary Bryan Scholarship, Schlumberger Scholarship, and the Bostater-Tellkamp-Power Scholarship.



David J. Love (S'98 - M'05 - SM'09) received the B.S. (with highest honors), M.S.E., and Ph.D. degrees in electrical engineering from the University of Texas at Austin in 2000, 2002, and 2004, respectively. During the summers of 2000 and 2002, he was with Texas Instruments, Dallas, TX. Since August 2004, he has been with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, where he is now a Professor and recognized as a University Faculty Scholar. He has served as an Associate Editor for both the IEEE

TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and he has also served as a guest editor for special issues of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and the *EURASIP Journal on Wireless Communications and Networking*. His research interests are in the design and analysis of communication systems and MIMO array processing. He has published over 120 technical papers in these areas and filed more than 20 U.S. patents, 18 of which have issued.

Dr. Love has been inducted into Tau Beta Pi and Eta Kappa Nu. Along with co-authors, he was awarded the 2009 IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY Jack Neubauer Memorial Award for the best systems paper published in the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in that year. He was the recipient of the Fall 2010 Purdue HKN Outstanding Teacher Award and was an invited participant to the 2011 NAE Frontiers of Engineering Education Symposium. In 2003, Dr. Love was awarded the IEEE Vehicular Technology Society Daniel Noble Fellowship.



Upamanyu Madhow is Professor of Electrical and Computer Engineering at the University of California, Santa Barbara. His research interests broadly span communications, signal processing and networking, with current emphasis on millimeter wave communication and bio-inspired approaches to networking and inference. He received his bachelor's degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1985, and his Ph.D. degree in electrical engineering from the University of Illinois, Urbana-Champaign in 1990.

He has worked as a research scientist at Bell Communications Research, Morristown, NJ, and as a faculty at the University of Illinois, Urbana-Champaign. Dr. Madhow is a recipient of the 1996 NSF CAREER award, and co-recipient of the 2012 IEEE Marconi prize paper award in wireless communications. He has served as Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON INFORMATION THEORY, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He is the author of the textbook *Fundamentals of Digital Communication*, published by Cambridge University Press in 2008.