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.

I. INTRODUCTION

The concept of adding 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. Indeed, informed transmitter strategies are critical for realizing the potentially transformative gains provided by order of magnitude increases in the number of transmit antennas. 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]. MIMO systems can also adapt using statistical side information [27]–[30].

It is difficult to scale the preceding codebook-based techniques to massive MIMO. In order to maintain the same level of channel quantization error, the feedback overhead must increase proportional to the number of transmit antennas [15]. Furthermore, it has been shown in [14] that an RVQ codebook is asymptotically optimal when the number of transmit antennas gets large assuming a fixed number of feedback bits per antenna. It is interesting to point out that the problem of designing practical and high performance limited feedback codebooks is very similar to the problem of designing good channel codes. It is well known that Gaussian codes or spherical codes are theoretically optimal; however, they have exponential encoding/decoding complexity [31], which prevents the use of such codes in real systems. Thus, there have been thousands of papers dedicated to practical channel code designs, and many practical codes such as convolutional codes, Reed-Solomon codes, turbo codes, and LDPC codes are implemented in practice [32]. Unlike the tremendous volume of research on channel codes, however, there is little work on structured, high performance limited feedback codebooks.

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 direct look-up for feedback generation infeasible. Thus, we propose noncoherent trellis-coded quantization (NTCQ) which tackles this computational bottleneck: the encoding complexity scales linearly with the number of antennas, while 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) [33], 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 [34]. 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 [35]–[38], but the path metrics used for the trellis search are ad hoc. On the other hand, the mapping to noncoherent sequence detection has been pointed out in [39], but does not consider coding. The use of nontrivial trellis codes as proposed here significantly enhances performance compared to PSK/QAM singular vector quantization (SVQ) proposed in [39]. Furthermore, [39] employs optimal noncoherent block demodulation, derived in [40], [41], for quantization, incurring complexity $O(M_t^3)$ for QAM constellations (QAM-SVQ) and $O(M_t \log M_t)$ for PSK constellations (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 [42],
Fig. 1: Multiple-input, single-output communications system with feedback.

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 variations 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 $\ell$ in the $k$th fading block can be written as:

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

where $h[k] \in \mathbb{C}^{M_t}$ is the MISO channel vector, $f[k] \in \mathbb{C}^{M_t}$ is the beamforming vector with $\|f[k]\|_2^2 = 1$, $s_\ell[k] \in \mathbb{C}$ 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 $h[k]$ 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 $h[k]$ into a binary vector $b[k]$, which is sent over a limited rate feedback channel. The transmitter

---

1Lower- and upper-case bold symbols denote vectors and matrices, respectively. The two-norm of a vector $x$ is denoted as $\|x\|_2$. The transpose and Hermitian transpose of a vector $x$ are denoted by $x^T$, $x^H$ 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 $\sigma^2$. 

May 23, 2013 DRAFT
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 \textit{normalized beamforming gain}, defined as the square of the normalized inner product between the transmit beamformer and the channel vector as

\[
J(\mathbf{f}, \mathbf{h}) = \frac{\| \mathbf{h} \|^2}{\| \mathbf{f} \|^2} |\mathbf{h}^H \mathbf{f}|^2. \tag{1}
\]

Although \( \| \mathbf{f} \|_2 = 1 \), we still normalize with \( \| \mathbf{f} \|_2 \) in (1) for better explanation afterward. We know from the Cauchy-Schwarz inequality that the largest possible value for this cost function is one and is attained when \( \mathbf{f} \) is a scalar multiple of \( \mathbf{h} \). An equivalent approach is to minimize the \textit{chordal distance} between \( \mathbf{f} \) and \( \mathbf{h} \), defined as

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

These performance measures require searching for codewords on the Grassmann manifold, a projective space in which vectors are mapped to one-dimensional complex subspaces. The chordal distance between two vectors is related to the minimum angle between the corresponding subspaces.

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

\[
\mathbf{c}_{\text{opt}} = \arg \max_{\mathbf{c} \in \mathcal{C}} J(\mathbf{c}, \mathbf{h}) = \arg \min_{\mathbf{c} \in \mathcal{C}} d_c(\mathbf{c}, \mathbf{h}), \tag{2}
\]

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 \( \mathbf{f} = \frac{\mathbf{c}_{\text{opt}(\mathbf{b})}}{\| \mathbf{c}_{\text{opt}(\mathbf{b})} \|^2} \) 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 \( O(M_2 2^{B_{\text{tot}}}) \), 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 \( \|c\|_2 \equiv 1 \)). 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 \( B_{\text{tot}} \) (or codebook size \( 2^{B_{\text{tot}}} \)) and the performance of MIMO systems has been thoroughly investigated for i.i.d. Rayleigh fading channels. In single user (SU) MISO channels with the \( B_{\text{tot}} \) bits RVQ codebook, the loss in normalized beamforming gain is given as [15]

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

where \( \mathcal{F}_{\text{RVQ}} \) 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 \( h \) and \( \mathcal{F}_{\text{RVQ}} \). 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 (i.e., the chordal distance) at a certain level.

In the MU-MIMO zero-forcing beamforming (ZFBF) case, 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, \( B_{\text{user}} \), must scale linearly with SNR (in dB) and \( M_t \) as

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

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

\[
B_{\text{tot}} \triangleq BM_t + q
\]

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 number of transmit elements is typically acceptable in terms of overhead, 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 $O(M_t2^B M_t)$. Thus, we need to develop new techniques to quantize the 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., [35], [39]). 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 [39], but we provide a self-contained derivation in Section II-C. 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

$$y = \beta x + n,$$

where $\beta \in \mathbb{C}$ is an unknown complex channel gain, $x \in \mathbb{C}^N$ is a vector of $N$ transmitted symbols, $n \in \mathbb{C}^N$ is complex Gaussian noise, and $y \in \mathbb{C}^N$ is the received signal. Using the generalized likelihood ratio test (GLRT) as in [39], [42], the estimate of the transmitted vector, $\hat{x}$, is given by

$$\hat{x} = \arg\min_{x \in \mathbb{C}^N} \min_{\beta \in \mathbb{C}} \|y - \beta x\|_2^2$$

$$= \arg\min_{x \in \mathbb{C}^N} \min_{\beta \in \mathbb{C}} \|y\|_2^2 + \|\beta x\|_2^2 - 2 \text{Re}(\beta y^H x)$$

$$= \arg\min_{x \in \mathbb{C}^N} \min_{\alpha \in \mathbb{R}_+} \min_{\theta \in [0, 2\pi]} \|y\|_2^2 + \alpha^2 \|e^{j\theta} x\|_2^2 - 2\alpha \text{Re}(e^{j\theta} y^H x)$$

$$= \arg\min_{x \in \mathbb{C}^N} \min_{\alpha \in \mathbb{R}_+} \|y\|_2^2 + \alpha^2 \|x\|_2^2 - 2\alpha |y^H x|$$

$$= \arg\max_{x \in \mathbb{C}^N} \frac{|y^H x|^2}{\|x\|_2^2}.$$

May 23, 2013 DRAFT
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

$$\min_{\theta \in [0, 2\pi)} \left\{ -\text{Re}(e^{j\theta}y^Hx) \right\} = -|y^Hx|.$$ 

To derive (7), we differentiate (6) with respect to $\alpha$ and set to 0 as

$$\frac{d}{d\alpha} \left\{ \|y\|^2 + \alpha^2 \|x\|^2 - 2\alpha |y^Hx| \right\} = 2\alpha \|x\|^2 - 2|y^Hx| = 0,$$

which gives $\alpha^* = \frac{|y^Hx|}{\|x\|^2}$. Note that $\alpha^*$ is the global minimizer of (6) because (6) is a quadratic function of $\alpha$. We can derive (7) after plugging in $\alpha^*$ 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 [39], we proved the duality of (4) and (2) in a different way than [39]). Therefore, we can find $c_{\text{opt}}$ for a MISO beamforming system with a Euclidean distance quantizer (or noncoherent block demodulator)

$$\min_{\alpha \in \mathbb{R}^+} \min_{\theta \in [0, 2\pi]} \min_{c_i \in \mathbb{C}} \| \tilde{h} - \alpha e^{j\theta}c_i \|^2, \quad (8)$$

where $\tilde{h} = \frac{h}{\|h\|^2}$ 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 [42] that the noncoherent block demodulator in (8) can be implemented near-optimally using a bank of coherent demodulators over the optimized discrete sets of $\alpha \in A = \{\alpha_1, \alpha_2, \ldots, \alpha_{K_{\alpha}}\}$ and $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_{K_{\theta}}\}$. While optimal noncoherent detection can be accomplished with quadratic complexity in $M_t$ [39], 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 $h \sim \mathcal{CN}(0, \sigma_h^2 I)$ 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

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

where $\mathcal{G} = \{g_1, \ldots, g_{2^{B_{\text{tot}}}}\}$, $B_{\text{tot}} = BM_t$, $g_i \sim \mathcal{CN}(0, (\sigma_h^2 - 2D)I)$ 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

$$d_c(h, g_{ED}) \xrightarrow{M_t \to \infty} 2^{-B} \quad (10)$$

**Proof:** By expanding $\|h - g_{ED}\|_2^2$, we have

$$\|h - g_{ED}\|_2^2 = \sum_{t=1}^{M_t} \left[ \{\text{Re}(h_t) - \text{Re}(g_{ED,t})\}^2 + \{\text{Im}(h_t) - \text{Im}(g_{ED,t})\}^2 \right]$$

where $h_t$ and $g_{ED,t}$ are the $t^{th}$ entry of $h$ and $g_{ED}$, 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 $\text{Re}(g_{ED,t})$ and $\text{Im}(g_{ED,t})$ 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 [43], we can achieve the rate-distortion bound

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

as $M_t \to \infty$. Thus, by the weak law of large numbers, the following convergences hold\(^2\)

$$\|h - g_{ED}\|_2^2 \xrightarrow{P} 2M_t E \left[ \{\text{Re}(h_t) - \text{Re}(g_{ED,t})\}^2 \right] = 2M_t D,$n$$

$$\|h\|_2^2 \xrightarrow{P} 2M_t E[\{\text{Re}(h_t)\}^2] = M_t \sigma_h^2,$n$$

$$\|g_{ED}\|_2^2 \xrightarrow{P} 2M_t E[\{\text{Re}(g_{ED,t})\}^2] = M_t (\sigma_h^2 - 2D)$$

\(^2\)Let $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ and $\mu = E[X_i]$ for all $i$. We say $\bar{X}_n$ converges to $\mu$ in probability as $\bar{X}_n \xrightarrow{P} \mu$ for $n \to \infty$ when $\lim_{n \to \infty} \text{Pr}(|\bar{X}_n - \mu| > \epsilon) = 0$ for any $\epsilon > 0$.\n
May 23, 2013 DRAFT
as $M_t \to \infty$. Moreover, $|h^H g_{ED}|^2$ can be lower bounded as

$$|h^H g_{ED}|^2 \geq \left( \text{Re}(h^H g_{ED}) \right)^2 \left( \frac{\|h\|^2 + \|g_{ED}\|^2}{2} - \|h - g_{ED}\|^2 \right)^2 P_t \left( M_t (\sigma_h^2 - 2D) \right)^2.$$  

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

$$d_c(h, g_{ED}) = 1 - \frac{|h^H g_{ED}|^2}{\|h\|^2 \|g_{ED}\|^2} \leq \frac{2D}{\sigma_h^2} = 2^{-B}.$$  

$$d_c(h, g_{ED}) \geq 2^{-\frac{BM_t}{M_t-1}}$$

where $(a)$ follows from the optimality of the RVQ codebook in large asymptotic regime [14]. As $M_t \to \infty$, the lower bound of $d_c(h, g_{ED})$ 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 $M_t \to \infty$ (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 $G$, is also an asymptotically optimal way to quantize the channel vector $h$. Of course, in practice, for finite constellations and number of antennas, we must “align” the codewords $g_i$ with the channel $h$, using parallel branches with different amplitude scaling $\alpha$ and phase rotations $\theta$ 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 [39] will not achieve optimal quantization performance. On the other hand, the constellation expansion employed in the NTCQ schemes considered here is sufficient to achieve near-optimal performance.

We now provide a non-asymptotic result regarding the chordal distances associated with Grassmannian line packing (GLP) attained by codebooks optimized using Euclidean metrics.

Let $N = 2^{B_{\text{tot}}}$ and $U_{M_t}^N \in \mathbb{C}^{M_t \times N}$ 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 $C$, we have to maximize the minimum Euclidean distance between all possible
codeword pairs
\[ d_{E,\text{min}}(\mathcal{C}) \triangleq \min_{1 \leq k < l \leq N} d_E(\mathbf{c}_k, \mathbf{c}_l) \]
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 \( \mathcal{C}_{\text{ED}} \) denote an optimized Euclidean distance codebook that maximizes the minimum Euclidean distance as
\[ \mathcal{C}_{\text{ED}} = \arg\max_{\mathcal{C} \in \mathcal{U}^N_{M_t}} d_{E,\text{min}}(\mathcal{C}). \]
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
\[ d_{c,\text{min}}(\mathcal{C}) \triangleq \min_{1 \leq k < l \leq N} d_c(\mathbf{c}_k, \mathbf{c}_l) \]
A GLP codebook is given as \[ \mathcal{C}_{\text{GLP}} = \arg\max_{\mathcal{C} \in \mathcal{U}^N_{M_t}} d_{c,\text{min}}(\mathcal{C}). \]
Note that the optimization metrics of \( \mathcal{C}_{\text{GLP}} \) and \( \mathcal{C}_{\text{ED}} \) 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
\[ d_c^2(\mathbf{x}, \mathbf{y}) \leq 1 - \left( 1 - \frac{1}{2} d_E^2(\mathbf{x}, \mathbf{y}) \right)^2 = d_E^2(\mathbf{x}, \mathbf{y}) - \frac{1}{4} d_E^4(\mathbf{x}, \mathbf{y}). \]

**Proof:** Let us define \( d_\theta^2(\mathbf{x}, \mathbf{y}) \) as
\[ d_\theta^2(\mathbf{x}, \mathbf{y}) \triangleq \min_{\theta \in [0, 2\pi]} d_E^2(\mathbf{x}, e^{i\theta} \mathbf{y}) = \| \mathbf{x} \|_2^2 + \| \mathbf{y} \|_2^2 - 2 \max_{\theta \in [0, 2\pi]} \Re \{ e^{i\theta} \mathbf{x}^H \mathbf{y} \} = 2 - 2|\mathbf{x}^H \mathbf{y}| \leq d_E^2(\mathbf{x}, \mathbf{y}). \]
Then, the squared chordal distance of \( \mathbf{x} \) and \( \mathbf{y} \) is upper bounded as
\[ d_c^2(\mathbf{x}, \mathbf{y}) = 1 - |\mathbf{x}^H \mathbf{y}|^2 = 1 - \left( 1 - \frac{1}{2} d_\theta^2(\mathbf{x}, \mathbf{y}) \right)^2 \leq 1 - \left( 1 - \frac{1}{2} d_E^2(\mathbf{x}, \mathbf{y}) \right)^2, \]
which finishes the proof.

Moreover, Lemma 1 can be directly extended to the following corollary.
Fig. 2: The minimum chordal distances of different codebooks with $M_t = 8$. GLP and ED codebook are numerically optimized according to their metrics, while the minimum distance of the RVQ codebook is averaged over 1000 different RVQ codebooks.

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

$$d_{c,\text{min}}(C_{ED}) \leq d_{E,\text{min}}(C_{ED}).$$

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

### III. NONCOHERENT TREVILLIS-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 [33],
Fig. 3: Quantization and reconstruction processes for a Euclidean distance quantizer using trellis-coded quantization (TCQ).

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 [44]. 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 $C_{TCM}$ denote a fixed codebook with $N$ codewords generated by a TCM channel code. Then $C_{TCM}$ can be mathematically expressed as

$$C_{TCM} = \arg\max_{C \in \mathcal{V}_N^{M_t}} d_{E_{\min}}(C)$$

where $\mathcal{V}_M^{N} \subset \mathcal{U}_M^{N}$ 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 $C_{TCM}$ is a Euclidean distance codebook within a given set $\mathcal{V}_M^{N}$. Thus, $C_{TCM}$ 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 $x \in \mathbb{C}^N$ 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
TABLE I: Mapping of quantizing bits/entry ($B$) and constellations.

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

the Euclidean distance from a given source vector as

$$c_{opt} = \arg\min_{c_i \in C_{TCM}} \|x - c_i\|_2^2,$$

which is then mapped to a binary sequence $b = \text{bin}(\text{opt})$. The quantized source vector $\hat{x}$ is reconstructed by passing the binary sequence $b$ 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 $b$ represents a unique quantized vector $\hat{x}$.

NTCQ adopts TCQ to quantize the CSI. Note that (11) is the same optimization problem as (8) with a given $\alpha \in \Lambda = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ and $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$. 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 $\alpha$ and $\theta$ 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 $c_i$’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 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 [44], as shown in Fig. 4. The source constellation is assumed to be 8PSK as in Fig. 5a. Note that all constellation points are normalized with the
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., $b_{in,1}$ is the least significant input bit and $b_{in,3}$ is the most significant input bit.

Fig. 5: Constellation points used in NTCQ are labeled with binary sequences.

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 $\vec{h}$ is quantized with one of the QPSK subconstellations marked by black or white circles in Fig.
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 $p_2 = [1, 2, 5]$ 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.

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 $c_i$ in (8) for a given $\alpha$ and $\theta$. 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, $s_t \in \{1, 2, \ldots, S\}$, in the trellis to a state at the $(t+1)^{th}$ stage, $s_{t+1}$, 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. 5a. 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 [35].

To optimize over the trellis, the first task is to define a path metric. Let $p_t$ be a partial path, or a sequence of states, up to the stage $t$. For example, the path $p_2 = [1, 2, 5]$ using state indices is highlighted in Fig. 6. Also, define the two functions $\text{in(} \cdot \text{)}$ and $\text{out(} \cdot \text{)}$ such that $\text{in}(p_t)$ outputs the binary input sequence corresponding to path $p_t$, and $\text{out}(p_t)$ gives the sequence of output constellation points corresponding to the path $p_t$. Again, using the sample path $p_2$ in Fig. 6, we
can see that
\[ \text{in}(p_2) = [01, 00]^T, \quad \text{out}(p_2) = \frac{1}{\sqrt{M_t}} \left[ -1, \frac{1}{\sqrt{2}}(1 + j) \right]^T. \]

With these definitions, we can define the path metric, \( m(\cdot) \), as
\[ m(p_t, \theta) = \| \bar{h}_t - e^{j\theta} \text{out}(p_t) \|_2^2, \]
where \( \theta \in [0, 2\pi) \) and \( \bar{h}_t \) is the vector created by truncating of normalized MISO channel vector \( \hat{h} \) 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
\[ m(p_t, \theta) = m(p_{t-1}, \theta) + \| \bar{h}_t - e^{j\theta} \text{out} \left( [p_{t-1}, p_t]^T \right) \|_2^2, \]
where \( \bar{h}_t \) and \( p_t \) are the \( t^{th} \) entry of \( \bar{h} \) and \( p_t \), respectively. The above path metric can be efficiently computed via the Viterbi algorithm. The path metric is computed parallel for each quantized value of \( \theta \in \Theta = \{ \theta_1, \theta_2, \ldots, \theta_{K_\theta} \} \). Then the best path \( p_{\text{best}} \) and the phase \( \theta_{\text{best}} \) that minimize the path metric can be found as
\[ \min_{\theta \in \Theta} \min_{p_{M_t} \in \mathcal{P}_{M_t}} m(p_{M_t}, \theta) \]
where \( \mathcal{P}_{M_t} \) denotes all possible paths up to stage \( M_t \). Finally, the beamforming vector \( f \) is calculated as
\[ c_{\text{opt}} = \text{out}(p_{\text{best}}), \quad f = \frac{c_{\text{opt}}}{\| c_{\text{opt}} \|_2}. \]
Note that \( \| c_{\text{opt}} \|_2 = 1 \) for 8PSK; therefore \( f = c_{\text{opt}} \).

It is important to point out that minimizing over \( \theta \) only increases the complexity of quantization, not the feedback overhead because the transmitter does not have to know the value of \( \theta_{\text{best}} \) that minimizes the path metric during the beamforming vector reconstruction process. However, there is additional feedback overhead with NTCQ. Since we test all the paths in the trellis, the transmitter has to know the starting state of \( p_{\text{best}} \), 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 $B_{\text{tot}}$ becomes

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

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. 5b where $d = \frac{\Delta}{2\sqrt{M_t}}$ with $\Delta = \sqrt{\frac{6}{M-1}}$ with $M = 16$ to have

$$E[\|c_i\|^2_2] = \frac{M_t}{16} \left[ 2d^2 \times 4 + \left\{ d^2 + (3d)^2 \right\} \times 8 + 2 (3d)^2 \times 4 \right] = 1$$

where expectation is taken over $c_i$ 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 $\alpha$ into account during the path metric computation as

$$m(p_t, \alpha, \theta) = \|\tilde{h}_t - \alpha e^{j\theta} \text{out}(p_t)\|^2_2$$  \hspace{1cm} (12)$$

where $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_{K_\theta}\}$ and $\alpha \in \mathbb{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{K_\alpha}\}$. Similar to the 8PSK case, additional $\log_2 S$ feedback bits are needed to indicate the starting state of $p_{\text{best}}$ 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 $O(2^B S M_t)$.

In NTCQ, the Viterbi algorithm has to be executed $K_\theta \cdot K_\alpha$ times during channel quantization, which gives the overall complexity of $O(K_\theta K_\alpha 2^B S M_t)$. However, $K_\theta \rightarrow 1$ and $K_\alpha \rightarrow 1$ as $M_t \rightarrow \infty$ without any performance loss by Theorem I. Moreover, as shown in Section V-A, the number of $K_\theta$ and $K_\alpha$ can be small even with moderate numbers of $M_t$, which gives a
negligible impact on the complexity in practice. Note that the complexity of NTCQ is linear to the number of transmit antennas $M_t$, which can significantly reduce the complexity of channel quantization in massive MIMO systems.

**E. Variations of NTCQ**

We can also construct several variations of NTCQ that tradeoff the total feedback bits $B_{\text{tot}}$ with the performance. We explain two variations briefly below.

- **Variation 1: Fixing the starting state for the trellis search.**

Because NTCQ searches paths that start from every states of the first stage in the trellis, we need additional $\log S$ bits of feedback overhead to indicate the starting state of $p_{\text{best}}$. However, we can only search paths that start from the first state of the first stage, which eliminates additional $\log_2 S$ bits of feedback overhead. The total feedback overhead becomes

$$B_{\text{tot}}^{(1)} = BM_t.$$

This variation, however, comes with a performance degradation because NTCQ loses the diversity gain of having multiple starting states for the path metric optimization process.

- **Variation 2: Fixing the first entry of $c_{\text{opt}}$.**

It is well known that we can define an equivalence relation between the two beamforming vectors $f_1 \in \mathbb{C}^{M_t}$ and $f_2 \in \mathbb{C}^{M_t}$ as $f_1 \equiv f_2$ in terms of subspace when $f_1 = e^{j\theta} f_2$ for some $\theta \in [0, 2\pi)$ [13] because $f_1$ and $f_2$ have the same beamforming gain $|h^H f_1|^2 = |h^H f_2|^2$. Therefore, we can fix the first entry of all $c_i$’s to $\frac{1}{\sqrt{M_t}}$ for the trellis optimization process. The first entry of the normalized channel vector $\bar{h}_1$ can be artificially quantized by $\alpha e^{j\theta}$ in (8). Then the total feedback overhead of this variation becomes

$$B_{\text{tot}}^{(2)} = B(M_t - 1) + \log_2 S.$$

We can also combine Variation 1 and 2 or adopt a tail-biting convolutional code for other variations.

**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 \[45\]

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

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

If \( \eta \) is close to one, two consecutive channels are highly correlated and the difference between the previous channel \( h[k-1] \) and the current channel \( h[k] \) 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 \( \eta \) 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 \( B_{tot} = 4 \). Therefore, we have to come up with a new differential feedback scheme to accommodate massive MIMO with large feedback overhead.

We denote \( f[k-1] \) as the quantized beamforming vector at block \( k-1 \) and \( f_{opt}[k] = \frac{h[k]}{||h[k]||_2} \) as the unquantized optimal beamforming vector at time \( k \). In our differential NTCQ scheme, instead of quantizing \( h[k] \) directly at time \( k \), the receiver quantizes \( f_{diff}[k] \) which is given as

\[
f_{diff}[k] = (I_{M_t} - f[k-1]f^H[k-1]) f_{opt}[k].
\]

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

\[
f_{\alpha,\theta} = \frac{\eta f[k-1] + \alpha e^{j\theta} \sqrt{1-\eta^2} \hat{f}_{diff}[k]}{||\eta f[k-1] + \alpha e^{j\theta} \sqrt{1-\eta^2} \hat{f}_{diff}[k]||_2}.
\]

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

\[
\max_{\alpha} \max_{\theta} ||f^H[k]f_{\alpha,\theta}||^2,
\]

May 23, 2013 DRAFT
and the final beamforming vector is given as $f[k] = f_{\bar{\alpha}_{\text{opt}}, \bar{\theta}_{\text{opt}}}$. To construct candidate beamforming vectors as in (13), we have to define sets of weights $\bar{A}$ 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 $\bar{A}$, we make the following proposition.

**Proposition 1.** When $\eta \to 1$, the range of $\bar{A}$ can be set as

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

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

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

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

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

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

$$
\left( \eta - \bar{\alpha} \sqrt{1 - \eta^2} \right)^2 \leq \|f_{\alpha, \bar{\theta}}^{\text{nom}}\|^2 \leq \left( \eta + \bar{\alpha} \sqrt{1 - \eta^2} \right)^2.
$$

Note that $f^H[k - 1] \hat{f}_{\text{diff}}[k] \approx 0$ 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 $\|f_{\alpha, \bar{\theta}}^{\text{nom}}\|^2 = 1$, and plugging in this to (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 $\hat{f}_{\text{diff}}[k]$, $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ to the transmitter over the feedback link and the transmitter reconstructs $f[k]$ 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 $\hat{f}_{\text{diff}}[k]$.

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

\[ h[k] = R^{\frac{1}{2}}h_w[k] \]

where \( h_w[k] \) is an uncorrelated MISO channel vector with i.i.d. complex Gaussian entries and \( R \) is a correlation matrix of the channel. We assume that \( R \) 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 \( R \) to quantize only the local space of the dominant eigenvector of \( R \). 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 \( R \) in advance. At the receiver side, \( h_w[k] \) is obtained by decorrelating \( h[k] \) with \( R^{-\frac{1}{2}} \), i.e., \( h_w[k] = R^{-\frac{1}{2}}h[k] \). Then the receiver quantizes \( h_w[k] \) with NTCQ and get \( \hat{h}_w[k] \). The receiver feeds back \( \hat{h}_w[k] \), and the transmitter reconstructs \( f[k] \) as

\[ f[k] = \frac{R^{\frac{1}{2}}\hat{h}_w[k]}{||R^{\frac{1}{2}}\hat{h}_w[k]||_2} \]

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 variations, differential NTCQ, and spatially adaptive NTCQ explained in Sections III, IV-A and IV-B, respectively. We use the averaged normalized beamforming gain in dB scale \( J_{\text{avg}}^{\text{dB}} = 10 \log_{10} (E[J(f, h)]) \) as a performance metric where the expectation is over \( h \).
A. i.i.d. Rayleigh fading Channels

For i.i.d. Rayleigh fading channels, $h[k]$ is drawn from i.i.d. complex Gaussian entries (i.e., $h[k] \sim \mathcal{CN}(0, I)$). In Fig. 7a, we first plot $J_{\text{avg}}^{dB}$ of NTCQ and its variations in i.i.d. channels with $M_t = 20$ transmit antennas depending on different quantization levels for $\theta_k$ and $\alpha_k$. Clearly, the two variations of NTCQ have strictly lower $J_{\text{avg}}^{dB}$ than the original NTCQ. Note that it is enough to have $K_\theta = 4$ for 1 bit/entry (QPSK) to achieve near-maximal performance of NTCQ and its variations. For 2 bits/entry (8PSK), Variation 2, which fixes the first entry of each $c_i$ to $e^{j\theta_k \sqrt{M_t}}$, has a notable gain with additional bits for $\theta_k$ because it has degrees of freedom for the first entry of the $c_i$’s only through $e^{j\theta_k}$. Interestingly, we can fix $\alpha_k = 1$ with 3 bits/entry (16QAM) for NTCQ and its variations without having any performance loss. This is because when optimizing (12), it is likely to have $E[\|c^{\text{opt}}\|^2] = 1$ since the objective variable is the normalized channel vector $\bar{h}$ which has a unit norm, i.e., $\|\bar{h}\|^2 = 1$. We fix $K_\theta = 16$ for simulations afterward regardless of the number of bits per entry to have a fair comparison. We also fix $\alpha_k = 1$ for 16QAM.

In Fig. 7b, we plot $J_{\text{avg}}^{dB}$ for Variation 1 of NTCQ (to have the same feedback overhead $B_{\text{tot}} = BM_t$ as 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 $J_{\text{avg}}^{dB}$ for unquantized beamforming, scalar quantization (SQ) with 6 bits/entry (4 bits for phase, 2 bits for amplitude), RVQ, PSK-SVQ in [39], and the benchmark from Theorem 1 which is given as $M_t \left(1 - 2^{-B}\right)$ (in linear scale). The performance of RVQ is plotted using the analytical approximation in (3) as $M_t \left(1 - 2^{-B_{\text{tot}}/M_t}\right)$ (in linear scale), because it is computational infeasible to simulate when...
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 $J_d^\text{dB}_{\text{avg}}$ for RVQ and Variation 1 of NTCQ is small for all $B$. Moreover, RVQ and Variation 1 of NTCQ both give better $J_d^\text{dB}_{\text{avg}}$ than SQ with 6 bits/entry with much less feedback in the $M_t = 100$ case. The plots of the benchmark using Theorem 1 well approximates $J_d^\text{dB}_{\text{avg}}$ of NTCQ for all $B$ and $M_t$, which shows the near-optimality of NTCQ.

Note that Variation 1 of NTCQ achieves better $J_d^\text{dB}_{\text{avg}}$ 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 NTCQ is 1 dB when $M_t = 100$ and $B = 1$. 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 [46] 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 $\tau$ denotes the channel instantiation interval. We assume a
carrier frequency of 2.5 GHz and $\tau = 5ms$. 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. 8a, we plot the performance of the proposed differential NTCQ feedback schemes with the velocity $v = 3km/h$ ($\eta = 0.9881$) assuming no feedback delay. The differential NTCQ schemes, even with 1 bit/entry quantization, achieve almost the same performance as unquantized beamforming regardless of $M_t$. To see the effect of feedback delay in temporally correlated channels, we simulate the $M_t = 100$ case with 5ms feedback delay (one fading block corresponds to 5ms) in Fig. 8b such that $J_{\text{avg-delay}}^{\text{dB}} = 10 \log_{10} \left( E[J(h^H[k], f[k-1]^2)] \right)$. It is shown that the effect of feedback delay is negligible, which confirms the practicality of the differential NTCQ scheme.

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 $\mathbf{R} = \mathbf{U}\Psi\mathbf{U}^H$ where $\mathbf{U}$ and $\Psi$ 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 $\Psi$ has a structure given by

$$\Psi = \text{diag} \left\{ \lambda_1, \frac{M_t - \lambda_1}{M_t - 1}, \ldots, \frac{M_t - \lambda_1}{M_t - 1} \right\}$$
where $1 \leq \lambda_1 < M_t$ is the dominant eigenvalue of $R$. If $\lambda_1$ 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. 9a and 9b, we plot $J_{dB}^{avg}$ as a function of $\lambda_1$ for $M_t = 10$ and 20 cases. The performance of spatially adaptive NTCQ become closer to that of unquantized beamforming as $\lambda_1$ 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 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 [39].

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, it is also crucial to develop scalable sounding schemes for channel estimation to make massive MIMO systems practical in FDD systems. 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 [47].
REFERENCES

[1] J. Choi, Z. Chance, D. J. Love, and U. Madhow, “Noncoherent trellis-coded quantization for massive MIMO limited feedback beamforming,” UCSD Information Theory and Applications Workshop, Feb. 2013.

[2] J. Choi, D. J. Love, and U. Madhow, “Limited feedback in massive MIMO systems—exploiting channel correlations via noncoherent trellis-coded quantization,” Proceedings of Conference on Information Sciences and Systems, Mar. 2013.

[3] T. L. Marzetta, “How much training is required for multiuser MIMO?” Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers, Oct. 2006.

[4] ——, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” IEEE Transactions on Wireless Communications, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[5] J. Hoydis, S. ten Brink, and M. Debbah, “Massive MIMO: how many antennas do we need?” Proceedings of Allerton Conference on Communication, Control, and Computing, Sep. 2011.

[6] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Uplink power efficiency of multiuser MIMO with very large antenna arrays,” Proceedings of Allerton Conference on Communication, Control, and Computing, Sep. 2011.

[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 Processing Magazine, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[8] N. Jindal, “MIMO broadcast channels with finite rate feedback,” IEEE Transactions on Information 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 Transaction on Signal Processing, 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 Journal on Selected Areas in Communications, 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 Transactions on Information 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 Transactions on Wireless Communications, 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 Transactions on Information 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 Transactions on Information 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 Transactions on Wireless Communications, vol. 6, no. 2, pp. 458–462, Feb. 2007.

[16] D. J. Love and R. W. Heath Jr., “Grassmannian beamforming on correlated MIMO channels,” Proceedings of IEEE Global Telecommunications Conference, Dec. 2004.

[17] ——, “Limited feedback diversity techniques for correlated channels,” IEEE Transactions on Vehicular Technology, 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 Transactions on Signal Processing, vol. 54, no. 5, pp. 1853–1863, Mar. 2006.

[19] B. Banister and J. Zeidler, “Feedback assisted transmission subspace tracking for MIMO systems,” IEEE Journal on Selected Areas in Communications, vol. 21, pp. 452–463, May 2003.

[20] J. Yang and D. Williams, “Transmission subspace tracking for MIMO systems with low-rate feedback,” IEEE Transactions on Communications, 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 Transaction on Signal Processing, 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 Transactions on Wireless Communications, 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 Transactions on Communications, 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 Transactions on Wireless Communications, 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,” Proceedings of IEEE Global Telecommunications Conference, Dec. 2012.

[27] E. Visotsky and U. Madhow, “Space-time transmit precoding with imperfect feedback,” IEEE Transactions on Information Theory, vol. 47, no. 6, pp. 2632–2639, Jun. 2001.

[28] H. Sampth and A. Paulraj, “Linear precoding for space-time coded systems with known fading correlations,” Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers, Nov. 2001.

[29] S. Zhou and G. B. Giannakis, “Optimal transmitter eigen-beamforming and space-time block coding based on channel correlations,” IEEE Transactions on Information Theory, vol. 49, no. 7, pp. 1673–1690, Jul. 2003.

[30] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” IEEE Journal on Selected Areas in Communications, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[31] D. Tse and P. Viswanath, Fundamentals of wireless communication. New York: Cambridge University Press, 2005.

[32] C. K. Au-Yeung, D. J. Love, and S. Sanayei, “Trellis coded line packing: large dimensional beamforming vector quantization and feedback transmission,” IEEE Transactions on Wireless Communications, vol. 10, no. 6, pp. 1844–1853, Jun. 2011.

[33] M. W. Marcellin, “Trellis coded quantization of memoryless and Gaussian-Markov sources,” IEEE Transactions on Communications, vol. 38, no. 1, pp. 82–93, Jan. 1990.

[34] “Technologies for Rel-12 and onwards,” Samsung Electronics. [Online]. Available: "http://www.3gpp.org/Future-Radio-in-3GPP-300-attend"

[35] C. K. Au-Yeung, D. J. Love, and S. Sanayei, “Trellis coded line packing: large dimensional beamforming vector quantization and feedback transmission,” IEEE Transactions on Wireless Communications, vol. 10, no. 6, pp. 1844–1853, Jun. 2011.

[36] C. K. Au-Yeung, A. Jalali, and D. J. Love, “Insights into feedback and feedback signaling for beamformer design,” UCSD Information Theory and Applications Workshop, Feb. 2009.

[37] C. K. Au-Yeung and S. Sanayei, “Enhanced trellis based vector quantization for coordinated beamforming,” Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 2010.

[38] M. Xu, D. Guo, and M. K. Honig, “MIMO precoding with limited rate feedback: simple quantizers work well,” Proceedings of IEEE Global Telecommunications Conference, Dec. 2009.

[39] 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 Transactions on Communications, vol. 57, no. 4, pp. 1184–1196, Apr. 2009.

[40] D. J. Ryan, I. B. Collings, and I. V. L. Clarkson, “Girt-optimal noncoherent lattice decoding,” IEEE Transaction on Signal Processing, vol. 55, no. 7, pp. 3773–3786, Jul. 2007.

[41] W. Sweldens, “Fast block noncoherent decoding,” IEEE Communications Letters, vol. 5, no. 4, pp. 132–134, Apr. 2001.

[42] D. Warrier and U. Madhow, “Spectrally efficient noncoherent communication,” IEEE Transactions on Information Theory, vol. 48, no. 3, pp. 652–668, Mar. 2002.

[43] R. G. Gallager, Information Theory and Reliable Communication. New York: Wiley, 1968.

[44] G. Ungerboeck, “Channel coding with multilevel/phase signals,” IEEE Transactions on Information Theory, vol. 28, no. 1, pp. 55–67, Jan. 1982.

[45] R. H. Etkin and D. N. C. Tse, “Degree of freedom in some underspread MIMO fading channel,” IEEE Transactions on Information Theory, vol. 52, pp. 1576–1608, Apr. 2006.

[46] J. G. Proakis, Digital communication, 4th ed. New York: McGraw-Hill, 2000.

[47] D. J. Love, J. Choi, and P. Bidigare, “A closed-loop training approach for massive MIMO beamforming systems,” Proceedings of Conference on Information Sciences and Systems, Mar. 2013.