Task-Based Quantization for Massive MIMO Channel Estimation

Nir Shlezinger, Yonina C. Eldar, and Miguel R. D. Rodrigues

Abstract—**Massive multiple-input multiple-output (MIMO) systems are the focus of increasing research attention. In such setups, there is an urgent need to utilize simple low-resolution quantizers, due to power and memory constraints. In this work we study massive MIMO channel estimation with quantized measurements, when the quantization system is designed to minimize the channel estimation error, as opposed to the quantization distortion. We first consider vector quantization, and characterize the minimal error achievable. Next, we focus on practical systems utilizing scalar uniform quantizers, and design the analog and digital processing as well as the quantization dynamic range to optimize the channel estimation accuracy. Our results demonstrate that the resulting massive MIMO system which utilizes low-resolution scalar quantizers can approach the minimal estimation error dictated by rate-distortion theory, achievable using vector quantizers.**

Index terms— **Massive MIMO, quantization.**

I. Introduction

Modern digital signal processing and communications systems use quantized digital representations of continuous-amplitude physical signals [1]. The quantized representations are typically designed to accurately match the original analog signal, by minimizing some distortion measure between the analog signal and the digital signal [2], regardless of the task of the system. Nonetheless, in many cases, the task of the system is not to recover the analog signal, but to extract some other information from its quantized representation [3]. It is therefore possible that in such systems – which we refer to as using *task-based quantization* – one can obtain further performance improvements in terms of the quantization rate necessary to achieve a certain performance.

As practical quantization systems typically utilize scalar uniform analog-to-digital convertors (ADCs) [1], recent years have witnessed a growing interest in systems implementing task-based quantization using low-resolution scalar quantizers. One of the main applications is massive multiple-input multiple-output (MIMO) [4]–[12]. In such systems, each base station (BS) in a wireless network is equipped with a large number of antennas [13], [14]. The BS uses a quantized representation of the received signal to estimate the underlying channel [4]–[7] and/or recover the transmitted messages [5]–[12]. For large number of BS antennas, accurate quantizers become costly in terms of power and memory usage, and low resolution quantizers are desirable. As the task in massive MIMO is not to recover the input signal, but to estimate the channel or decode the transmitted message, reasonable performance with low-resolution scalar quantizers has been observed [4]–[12]. However, prior works typically assume uniform quantization with a fixed dynamic range, thus, they do not characterize the performance that can be achieved when the quantizers and the dynamic range also account for the system task.

Joint (vector) quantization is known to outperform scalar quantization in the sense of achieving minimal distortion [15, Ch. 10]. Task-based vector quantization can be considered as a special case of the indirect lossy-source coding setup [2]. In such scenarios, one wishes to recover a desired signal based on a discrete representation of an observed signal, which is statistically related to the desired signal [16]. For the mean-squared error (MSE) distortion, it was shown in [17] that the optimal system which achieves the rate-distortion curve applies vector quantization to the minimum mean-squared error (MMSE) estimate of the desired signal. Nonetheless, in massive MIMO systems, vector quantization becomes infeasible, and practical quantization approaches are required.

Task-based quantization with scalar uniform ADCs, referred to as *hardware-limited task-based quantization*, can be realized in practice by allowing analog linear processing prior to quantization [18]. In the context of MIMO systems with quantized inputs, [19] compared the achievable-rate versus power efficiency tradeoff for various analog combining systems, [9] and [20] jointly designed precoding with analog combining to maximize the achievable rate and the MSE in signal recovery, respectively, with full channel state information (CSI), while [10] proposed a bit allocation algorithm for minimizing the quantization error when the analog combining is set to the largest channel eigenmodes, using high rate quantization analysis. Additionally, [11] studied the achievable rate in the presence of imperfect CSI when distinct sets of inputs are each combined in analog to maximize the receive power, while [21] characterized bounds on the capacity of MIMO communications with analog combining and one-bit quantization. It is emphasized that previous works which studied analog combining for MIMO receivers, e.g., [9], [10], [20], required knowledge of the underlying channel in their design, and thus cannot be utilized for channel estimation. In particular, the joint design of the analog combining, quantization rule, and digital processing, to optimize the accuracy of massive MIMO channel estimation with scalar ADCs has not yet been studied to the best of our knowledge.

In this work we study task-based quantization for massive MIMO channel estimation. In our previous work [18], we showed that for the generic parameter acquisition task with fixed input size and number of quantization bits, hardware-limited task-based quantization can approach the optimal performance dictated by indirect rate-distortion theory. Here, we extend our analysis of [18] to massive MIMO systems, where the dimensions of the input signal are asymptotically large, while the number of bits per input sample is fixed.

We first characterize the minimal achievable average MSE for any quantization system operating with a fixed quantization rate, namely, a fixed number of bits per input sample, revealing the fundamental limits of massive MIMO channel estimation from quantized measurements. We then proceed to study practical task-based quantization with scalar uniform ADCs, by allowing analog combining prior to quantization. We consider two types of analog combining: The first allows the inputs to be gathered over different antennas as well as over different time instances, while the second allows only inputs taken at the same time instance to be combined. For each setup we derive the optimal system, characterize its analog combining matrix, its digital processing, and the achievable average MSE. Our simulations demonstrate that the proposed quantizers, which utilize practical low-resolution scalar uniform quantizers, are capable of approaching the fundamental performance limits dictated by indirect rate-distortion theory, achievable with vector quantizers.

The rest of this paper is organized as follows: Section II reviews some preliminaries in quantization theory, and presents the system model. Section III studies the task-based quantization systems. Sec-

This project has received funding from the European Unions Horizon 2020 research and innovation program under grant No. 646804-ERC-COG-BNYQ, from the Israel Science Foundation under grant No. 0100101, and from the Royal Society International Exchange scheme IE 160348.

N. Shlezinger and Y. C. Eldar are with the faculty of Math and CS, Weizmann Institute of Science, Rehovot, Israel (e-mail: nirshlezinger1@gmail.com; yonina@weizmann.ac.il). M. R. D. Rodrigues is with the department of EE, University College, London, UK (e-mail: m.rodrigues@ucl.ac.uk).

tion IV presents a numerical study.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$, and its ith element is denoted $(\boldsymbol{x})_i$. Boldface upper-case letters denote matrices, e.g., $\boldsymbol{M}$, and $(\boldsymbol{M})_{i,j}$ is its (i,j)th element. Hermitian transpose, transpose, stochastic expectation, and mutual information are written as $(\cdot)^H$, $(\cdot)^T$, $\mathbb{E}\{\cdot\}$, and $I(\cdot\,;\cdot)$, respectively. We use $a^+ \triangleq \max(a,0)$, $\mathrm{Tr}(\cdot)$ is the trace operator, $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, $\delta_{(\cdot)}$ is the indicator function, $\otimes$ is the Kronecker product, $\mathcal{R}$ and $\mathcal{C}$ are the sets of real and complex numbers, respectively.

II. Preliminaries and System Model

A. Preliminaries in Quantization Theory

To formulate the problem, we first briefly review the standard quantization setup. While parts of this review also appear in our previous work [18], it is included for completeness. We begin with the definition of a quantizer:

Definition 1 (Quantizer)**.** *A quantizer $Q_M^{n,k}(\cdot)$ with $\log M$ bits, input size n, input alphabet $\mathcal{X}$, output size k, and output alphabet $\hat{\mathcal{X}}$, consists of:* 1) *An encoding function $g_n^{\mathrm{e}} : \mathcal{X}^n \mapsto \{1,2,\ldots,M\} \triangleq \mathcal{M}$ which maps the input from $\mathcal{X}^n$ into a discrete index $i \in \mathcal{M}$.* 2) *A decoding function $g_k^{\mathrm{d}} : \mathcal{M} \mapsto \hat{\mathcal{X}}^k$ which maps each index $i \in \mathcal{M}$ into a codeword $\boldsymbol{q}_i \in \hat{\mathcal{X}}^k$.*

We write the quantizer output given input $\boldsymbol{x}$ as $\hat{\boldsymbol{x}} = Q_M^{n,k}(\boldsymbol{x})$. *Scalar quantizers* operate on scalar inputs, i.e., $n = 1$ and $\mathcal{X}$ is a scalar space, while *vector quantizers* have multivariate inputs. For equally sized input and output, we write $Q_M^n(\cdot) \triangleq Q_M^{n,n}(\cdot)$.

In the standard quantization problem, a $Q_M^n(\cdot)$ quantizer is designed to minimize some distortion measure between its input and its output. Characterizing the optimal quantizer and the optimal tradeoff between distortion and quantization rate is in general a very difficult task. In the special case when the input consists of i.i.d. random variables (RVs), and the distortion measure is separable, the optimal distortion in the limit $n \to \infty$ for a rate R is given by the distortion-rate function:

Definition 2 (Distortion-rate function)**.** *The distortion-rate function for input $\boldsymbol{x} \in \mathcal{X}$ with distortion measure $d : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{R}^+$ is defined as*

$$D_{\boldsymbol{x}}(R) = \min_{f_{\hat{\boldsymbol{x}}|\boldsymbol{x}} : I(\hat{\boldsymbol{x}};\boldsymbol{x}) \leq R} \mathbb{E}\left\{ d\left(\hat{\boldsymbol{x}}, \boldsymbol{x}\right)\right\}. \tag{1}$$

B. System Model

We study pilot-aided channel estimation from quantized measurements in a noncooperative multi-cell multi-user MIMO system with n_c cells. In each cell, a BS equipped with n_t antennas serves n_u single-antenna user terminals (UTs). We focus on the *massive MIMO regime*, namely, the number of antennas n_t is sufficiently large to carry out large-scale (asymptotic) analysis.

We consider the block-fading massive MIMO channel model, as in [13]. To formulate the model, let $\boldsymbol{D}_{k,l}$ be an $n_u \times n_u$ diagonal matrix with positive diagonal entries $\{d_{k,l,m}\}_{m=1}^{n_u}$ representing the attenuation between the mth UT of the lth cell and the kth BS, $k,l \in \{1,2,\ldots,n_c\} \triangleq \mathcal{N}_c$. Without loss of generality, we assume that for each $k \in \mathcal{N}_c$, the coefficients $\{d_{k,k,m}\}_{m=1}^{n_u}$ are arranged in descending order. Furthermore, let $\boldsymbol{H}_{k,l} \in \mathcal{C}^{n_t \times n_u}$ be a random proper-complex zero-mean Gaussian matrix with i.i.d. entires of unit variance, representing the instantaneous channel response between the UTs of the lth cell and the kth BS, $k,l \in \mathcal{N}_c$. Let $\boldsymbol{G}_{k,l} = \boldsymbol{H}_{k,l}\boldsymbol{D}_{k,l}$ be the random channel matrix from the UTs in the kth cell to the lth BS. We assume a block-fading model for $\{\boldsymbol{H}_{k,l}\}_{k,l \in \mathcal{N}_c}$, in which the channel coefficients $\{\boldsymbol{H}_{k,l}\}_{k,l \in \mathcal{N}_c}$ are unknown. Let $\boldsymbol{w}_k[i] \in \mathcal{C}^{n_t}$, $k \in \mathcal{N}_c$, be an i.i.d. zero-mean proper-complex Gaussian signal with covariance matrix $\sigma_W^2 \boldsymbol{I}_{n_t}$, $\sigma_W^2 > 0$, representing the additive channel noise at the kth BS.

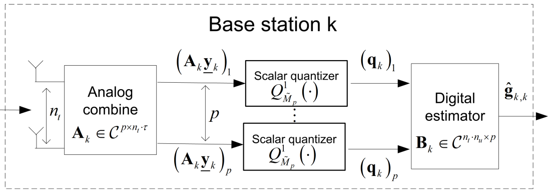
Base station k
$\underline{\mathbf{y}}_k$
n_t
Analog combine
$\mathbf{A}_k \in \mathcal{C}^{p \times n_t \cdot \tau}$
$\left(\mathbf{A}_k \underline{\mathbf{y}}_k\right)_1$
p
$\left(\mathbf{A}_k \underline{\mathbf{y}}_k\right)_p$
Scalar quantizer $Q^1_{\tilde{M}_p}(\cdot)$
Scalar quantizer $Q^1_{\tilde{M}_p}(\cdot)$
$(\mathbf{q}_k)_1$
$(\mathbf{q}_k)_p$
Digital estimator
$\mathbf{B}_k \in \mathcal{C}^{n_t \cdot n_u \times p}$
$\hat{\mathbf{g}}_{k,k}$

Fig. 1. Massive MIMO channel estimation with scalar ADCs.

Channel estimation is carried out in a time-division duplex fashion. Each UT sends an orthogonal pilot sequence (PS) of τ pilot symbols, where the PSs are the same in all cells. The BSs use the knowledge of the PSs to estimate the channel. Let $s_m[i]$ denote the ith pilot symbol of the mth user in each cell, $m \in \{1,\ldots,n_u\} \triangleq \mathcal{N}_u$, and define $\boldsymbol{s}[i] \triangleq [s_1[i],\ldots,s_{n_u}[i]]^T$. The channel output at the kth BS, $k \in \mathcal{N}_c$, is given by

$$\boldsymbol{y}_k[i] = \sum_{l=1}^{n_c} \boldsymbol{G}_{k,l}\boldsymbol{s}[i] + \boldsymbol{w}_k[i], \qquad i \in \{1,\ldots,\tau\}. \tag{2}$$

Alternatively, by defining $\underline{\boldsymbol{y}}_k \triangleq \mathrm{vec}\big(\boldsymbol{y}_k[1],\ldots,\boldsymbol{y}_k[\tau]\big)$, $\underline{\boldsymbol{w}}_k \triangleq \mathrm{vec}\big(\boldsymbol{w}_k[1],\ldots,\boldsymbol{w}_k[\tau]\big)$, $\underline{\boldsymbol{g}}_{k,l} \triangleq \mathrm{vec}\big(\boldsymbol{G}_{k,l}\big)$, and the $n_u \times \tau$ deterministic matrix $\boldsymbol{S} \triangleq \big(\boldsymbol{s}[1],\ldots,\boldsymbol{s}[\tau]\big)$, the observed signal used for channel estimation can be written as

$$\underline{\boldsymbol{y}}_k = \sum_{l=1}^{n_c} \left(\boldsymbol{S}^T \otimes \boldsymbol{I}_{n_t}\right) \underline{\boldsymbol{g}}_{k,l} + \underline{\boldsymbol{w}}_k. \tag{3}$$

Since the PSs are orthogonal and $\tau \geq n_u$, it holds that $\boldsymbol{S}\boldsymbol{S}^H = \tau \cdot \boldsymbol{I}_{n_u}$, and the covariance matrix of $\underline{\boldsymbol{y}}_k$ is given by $\boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}_k} = \boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}'_k} \otimes \boldsymbol{I}_{n_t}$, where $\boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}'_k} \triangleq \sum_{l=1}^{n_c} \boldsymbol{S}^T \boldsymbol{D}_{k,l}^2 \boldsymbol{S}^* + \sigma_W^2 \boldsymbol{I}_\tau$. Furthermore, the PS length, τ, must not be smaller than n_u [13, Sec. III-A]. Each BS uses up to $\log M$ bits to represent $\underline{\boldsymbol{y}}_k$, from which an estimate of its channel $\underline{\boldsymbol{g}}_{k,k}$, denoted $\hat{\underline{\boldsymbol{g}}}_{k,k}$, is produced.

Our goal is to derive the achievable average MSE in estimating the channel matrix at a given cell with index $k \in \mathcal{N}_c$, and to characterize the corresponding quantization scheme. In our analysis, we fix the quantization rate, defined as $R \triangleq \frac{1}{n_t \cdot \tau} \log M$, and derive the achievable MSE in massive MIMO regime, i.e.,

$$\mu_k \triangleq \lim_{n_t \to \infty} \frac{1}{n_t \cdot n_u} \mathbb{E}\Big\{ \big\| \underline{\boldsymbol{g}}_{k,k} - \hat{\underline{\boldsymbol{g}}}_{k,k} \big\|^2 \Big\}. \tag{4}$$

We consider the following quantization systems architectures:

Vector Quantizers: In the optimal task-based quantization system, the quantizer $Q_M^{\tau \cdot n_t, n_u \cdot n_t}(\cdot)$ minimizes the distortion between the quantized representation $\hat{\boldsymbol{g}}$ and $\boldsymbol{g}$. The performance of this system represents the fundamental limit of massive MIMO channel estimation with quantized measurements.

Hardware-Limited Quantizers: Vector quantization may be difficult to implement, especially for large input sizes. As discussed in the introduction, practical systems typically implement quantization using scalar ADCs. Here, each continuous-amplitude sample is converted into a discrete representation using a single uniform quantization rule. In particular, we consider the system depicted in Fig. 1. The observed vector $\underline{\boldsymbol{y}}_k$, is projected into $\mathcal{C}^p$ using some pre-quantization processing carried out in the analog domain. We assume that p, which represents the number of scalar quantizers, is not larger than the size of the observed vector $\tau \cdot n_t$. In the following, we write the number of scalar quantizers in terms of its quotient and remainder with respect to n_t, denoted n_p and n_q, respectively, i.e.,

$$p = n_p \cdot n_t + n_q, \qquad 0 < n_q < n_t. \tag{5}$$

The motivation for (5) stems from the fact that in massive MIMO, the number of antennas n_t tends to infinity, and thus n_p and n_q represent how p scales accordingly. Since arbitrary processing may be difficult to implement in analog, we restrict our attention to linear pre-quantization processing only. Thus, we allow the quantization system to utilize *analog combining*, modeled via the matrix $\boldsymbol{A}_k \in \mathcal{C}^{p \times \tau \cdot n_t}$.

The real and imaginary parts of each entry of $\boldsymbol{A}_k \boldsymbol{y}$ are quantized using the same scalar quantizer with resolution $\tilde{M}_p \triangleq \lfloor M^{1/2p} \rfloor$, denoted $Q^1_{\tilde{M}_p}(\cdot)$. By defining the combining ratio

$$r \triangleq \frac{p}{\tau \cdot n_t} = \frac{n_p}{\tau} + \frac{n_q}{\tau \cdot n_t}, \tag{6}$$

it holds that $\tilde{M}_p = \lfloor 2^{\frac{R}{2 \cdot r}} \rfloor$, and the overall quantization rate is $\frac{2 \cdot p}{\tau \cdot n_t} \log\left(\tilde{M}_p\right) \le \frac{1}{\tau \cdot n_t} \log M = R$. The scalar quantizers $Q^1_{\tilde{M}_p}$ implement non-subtractive uniform dithered quantization [24], see [18, Sec. II] for a detailed discussion on the benefits of carrying out our analysis assuming dithered quantizers.

To formulate the input-output relationship of the scalar ADC, let γ denote the dynamic range of the quantizer, and define $\Delta_p \triangleq \frac{2\gamma}{\tilde{M}_p}$ as the quantization spacing. The uniform quantizer is designed to operate within the dynamic range. To guarantee this, we fix γ to be some multiple η of the maximal standard deviation of the input. We assume that $\eta < \sqrt{3}\tilde{M}_p$, such that the variable $\kappa_p \triangleq \eta^2 \big(1 - \frac{2\eta^2}{3\tilde{M}_p^2}\big)^{-1}$ is strictly positive. Note that $\eta = 2$ satisfies this requirement for any $\tilde{M}_p \ge 2$. The output of the serial scalar ADC with input sequence $y_1, \ldots, y_p$ can be written as $Q^1_{\tilde{M}_p}(y_i) = q_p\left(\mathrm{Re}\{y_i + z_i\}\right) + j \cdot q_p\left(\mathrm{Im}\{y_i + z_i\}\right)$, where $z_1, \ldots, z_p$ are i.i.d. RVs with i.i.d. real and imaginary parts uniformly distributed over $\left[-\frac{\Delta_p}{2}, \frac{\Delta_p}{2}\right]$, independent of the input. The uniform quantization function $q_p(\cdot)$ is given by

$$q_p(y) = \begin{cases} -\gamma + \Delta_p\left(l + \frac{1}{2}\right) & \begin{array}{c} y - l \cdot \Delta_p + \gamma \in [0, \Delta_p] \\ l \in \{0, 1, \ldots, \tilde{M}_p - 1\} \end{array} \\ \mathrm{sign}\,(y)\left(\gamma - \frac{\Delta_p}{2}\right) & |y| > \gamma. \end{cases}$$

Note that when $\tilde{M}_p = 2$, the resulting quantizer is a standard one-bit sign quantizer of the form $q_p(y) = c \cdot \mathrm{sign}(y)$, where the constant $c > 0$ is determined by the dynamic range γ.

Finally, in the digital domain, the system computes the linear MMSE channel estimate based on the ADC output, denoted $\boldsymbol{q}_k \in \mathcal{C}^p$, where $(\boldsymbol{q})_i = Q^1_{\tilde{M}_p}\left((\boldsymbol{A}_k \underline{\boldsymbol{y}}_k)_i\right)$. The estimated vector can be written as $\hat{\boldsymbol{g}} = \boldsymbol{B}_k \boldsymbol{q}_k$ for some $n_u \cdot n_t \times p$ matrix $\boldsymbol{B}_k$.

III. Task-Based Quantization Systems

In the following we study task-based quantization for massive MIMO channel estimation. As a preliminary step, we characterize the average MSE without quantization, i.e., the average MMSE. We then characterize the achievable average MSE when the BS uses vector quantizers, which is the optimal quantization strategy [22, Ch. 23]. Next, we derive the achievable average MSE when the BS uses hardware-limited quantizers. Finally, we note that it may be preferable in massive MIMO to combine only channel outputs received at the same time instance. By accounting for this constraint, we derive the achievable average MSE and the resulting quantization system for this form of restricted hardware-limited quantization.

To derive the MMSE, define $\phi_{k,m} \triangleq \sqrt{f_{k,m}} d_{k,k,m}$ where

$$f_{k,m} \triangleq \frac{\tau d^2_{k,k,m}}{\sigma^2_W + \tau \sum\limits_{l'=1}^{n_c} d^2_{k,l',m}}, \quad k, l \in \mathcal{N}_c, m \in \mathcal{N}_u, \tag{7}$$

as well as the $n_u \times n_u$ diagonal matrices $\{\boldsymbol{\Phi}_k\}_{k \in \mathcal{N}_c}$ and $\{\boldsymbol{F}_k\}_{k \in \mathcal{N}_c}$ with diagonal entries $\{\phi_{k,m}\}_{m=1}^{n_u}$ and $\{f_{k,m}\}_{m=1}^{n_u}$, respectively. The MMSE channel estimation error is stated in the following lemma:

Lemma 1. *The average MSE of the MMSE estimate, denoted $\underline{\tilde{\boldsymbol{g}}}_{k,k}$, is given by*

$$\mu_k^{\mathrm{MMSE}} = \frac{1}{n_u} \sum_{m=1}^{n_u} \left(d^2_{k,k,m} - \phi^2_{k,m}\right). \tag{8}$$

Furthermore, $\underline{\tilde{\boldsymbol{g}}}_{k,k}$ is a zero-mean $n_t \cdot n_u \times 1$ Gaussian random vector with covariance matrix $\mathbb{E}\{\underline{\tilde{\boldsymbol{g}}}_{k,k} \underline{\tilde{\boldsymbol{g}}}_{k,k}^H\} = \left(\boldsymbol{\Phi}_k^2 \otimes \boldsymbol{I}_{n_t}\right)$.

Proof: The lemma follows from [14, Appendix B]. ■

Having characterized the MMSE channel estimate for the massive MIMO setup without quantization, we are now ready to introduce quantization constraints in the following subsections.

A. Vector Quantization

We now obtain the minimal achievable average MSE of any quantization system with quantization rate R using indirect rate-distortion theory. To that aim, let $D_{\mathrm{G}}(R, \boldsymbol{\Sigma})$ be the distortion-rate function of a zero-mean proper-complex Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}$, obtained from Def. 2 via the reverse waterfilling algorithm [15, Ch. 10.3]. The minimal average MSE is stated in the following proposition:

Proposition 1. *The average MSE of the optimal vector quantizer is*

$$\mu_k^{\mathrm{Opt}} = \mu_k^{\mathrm{MMSE}} + \frac{1}{n_u} D_{\mathrm{G}}\left(\frac{\tau}{n_u} \cdot R, \boldsymbol{\Phi}_k^2\right). \tag{9}$$

Proof: The proposition follows from indirect rate-distortion theory, as the distortion is minimized by applying optimal vector quantization to the MMSE estimate [17], see also [3]. ■

Note that (9) represents the fundamental performance limits of massive MIMO channel estimation with quantized observations, and is achievable using vector quantizers.

B. Hardware-Limited Quantization: Spatial-Temporal Combining

Utilizing vector quantization in massive MIMO systems is likely to be infeasible due to its extremely high complexity for large-scale inputs. Thus, practical massive MIMO systems utilize serial scalar uniform ADCs, corresponding to the hardware-limited quantization setup described in Subsection II-B.

By setting the analog combining matrix $\boldsymbol{A}_k$ to be $\boldsymbol{I}_{n_t}$, the system specializes to the standard model for MIMO channel estimation from quantized inputs, as in [4]–[7]. Thus, the analog processing of $\underline{\boldsymbol{y}}_k$ combined with the adjustment of the dynamic range γ are the main difference between our model and existing models in the literature. In Section IV we numerically illustrate that jointly optimizing the analog combining and the quantization dynamic range significantly improves the estimation accuracy, and can approach the fundamental limits achievable with vector quantizers.

We next characterize the minimal achievable average MSE in estimating massive MIMO channels using hardware-limited quantizers. Our analysis follows similar guidelines to [18, Thm. 1], with the exception that [18] considered real-valued signals with fixed size and fixed number of bits. Here, the signals are complex with asymptotically large dimensions while the number of bits grows proportionally. We define the non-negative function $\beta_{k,m}(x) \triangleq (x \cdot \phi_{k,m} - 1)^+$. Our derivation reveals the optimal analog combining matrix and the linear MMSE matrix, denoted $\boldsymbol{A}_k^{\circ}$ and $\boldsymbol{B}_k^{\circ}$, respectively, and the corresponding dynamic range γ, for fixed quantization rate R and analog combining ratio r, as stated in the following proposition:

Proposition 2. *The minimal achievable average MSE of the hardware-limited task-based quantization system is given by*

$$\mu^{\mathrm{HL}} = \mu^{\mathrm{MMSE}} + \frac{1}{n_u} \sum_{m=1}^{\min(n_u, n_p)} \frac{\phi^2_{k,m}}{\beta_{k,m}(\zeta) + 1} + \delta_{(n_p < n_u)} \frac{1}{n_u}$$
$$\times \left(\sum_{m=n_p+1}^{n_u} \phi^2_{k,m} - (r \cdot \tau - n_p) \frac{\beta_{k,n_p+1}(\zeta) \cdot \phi^2_{k,n_p+1}}{\beta_{k,n_p+1}(\zeta) + 1} \right), \tag{10a}$$

where ζ is set such that

$$\frac{1}{\tau} \sum_{m=1}^{n_p} \beta_{k,m}(\zeta) + (r \cdot \tau - n_p)\, \beta_{k,n_p+1}(\zeta) = \frac{3\tilde{M}_p^2 \cdot r}{4\kappa_p}. \tag{10b}$$

In the optimal system, the analog combining matrix is given by $\boldsymbol{A}_k^{\circ} = \boldsymbol{U}_{\boldsymbol{A}} \boldsymbol{\Lambda}_{\boldsymbol{A}} \left(\boldsymbol{V}_{\boldsymbol{A}}^H \boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}_k}^{-1/2} \otimes \boldsymbol{I}_{n_t}\right)$, where $\boldsymbol{V}_{\boldsymbol{A}} \in \mathcal{C}^{\tau \times \tau}$ is the right *singular vectors matrix of $\tau^{-1}\boldsymbol{F}_k\boldsymbol{S}^*\boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}'_k}^{1/2}$, $\boldsymbol{\Lambda}_{\boldsymbol{A}} \in \mathcal{C}^{p\times\tau\cdot n_t}$ is a diagonal matrix with diagonal entries $(\boldsymbol{\Lambda}_{\boldsymbol{A}})^2_{l,l} = \frac{4\kappa_p}{3\tilde{M}_p^2\cdot r}\beta_{k,\lfloor\frac{l-1}{n_t}\rfloor+1}(\zeta)\cdot\delta_{(l<n_u\cdot n_t)}$, and $\boldsymbol{U}_{\boldsymbol{A}} \in \mathcal{C}^{p\times p}$ is a unitary matrix such that $\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^H\boldsymbol{U}_{\boldsymbol{A}}^H$ has identical diagonal entries [25, Ch. 2].*

The dynamic range of the ADC is given by $\gamma^2 = \frac{\kappa_p}{r}$, and the digital processing matrix is given by

$$\boldsymbol{B}_k^{\circ} = \left(\boldsymbol{D}_{k,k}^2\boldsymbol{S}^*\otimes\boldsymbol{I}_{n_t}\right)\left(\boldsymbol{A}_k^{\circ}\right)^H\left(\boldsymbol{A}_k^{\circ}\boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}_k}(\boldsymbol{A}_k^{\circ})^H + \frac{4\gamma^2}{3\tilde{M}_p^2}\boldsymbol{I}_p\right)^{-1}.$$

Proof: The proof is obtained by repeating the arguments in [18, Appendix C] for complex inputs, while letting their size tend to infinity, and is omitted due to space limitations. ■

We note that $\boldsymbol{A}_k^{\circ}$ linearly combines samples taken at different time instances, i.e., temporal combining. Unlike spatial combining, which can be implemented using simple hardware, see, e.g., [20], temporal combining requires storing samples for different durations in analog, which may be difficult when the PS length τ is large. Consequently, we next characterize the optimal system when $\boldsymbol{A}_k$ is restricted to implement only spatial combining.

C. Spatial Analog Combining

In Proposition 2 we characterized the minimal achievable MSE when the input to the scalar ADCs can be written as any linear transformation of all the channel outputs, $\underline{\boldsymbol{y}}_k$. Since it may be preferable not to combine samples received at different time instances in analog, we now restrict the analog processing to combine only samples received at the same time instance. It should be noted that this model for analog combining in MIMO systems was also considered in the works [9], [10], [21].

To formulate the setup, let $\tilde{\boldsymbol{A}}_k \in \mathcal{C}^{n_t\times\tilde{p}}$ represent the analog combining, applied to each received channel output. Here, at each time index i, the vector $\tilde{\boldsymbol{A}}_k\boldsymbol{y}_k[i]$ is quantized using $\tilde{p}$ identical scalar quantizers with resolution $\tilde{M}_{\tau\cdot\tilde{p}}$, where $\tilde{M}_{\tau\cdot\tilde{p}} = \lfloor M^{1/(2\tau\cdot\tilde{p})}\rfloor$. This setup is a special case of the model illustrated in Fig. 1, where the analog combining matrix can be written as $\boldsymbol{A}_k = \boldsymbol{I}_\tau\otimes\tilde{\boldsymbol{A}}_k$ and $p = \tilde{p}\cdot\tau$. Consequently, $r = \frac{p}{\tau\cdot n_t} = \frac{\tilde{p}}{n_t}$, thus, letting n_t grow arbitrarily large implies that $\tilde{p}$ grows proportionally. Let σ_k be the maximal diagonal entry of $\boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}'_k}$. Under this setting, the optimal system and its average MSE are stated in the following proposition, given without proof due to space limitations.

Proposition 3. *The minimal achievable average MSE when only spatial analog combining is carried out is given by*

$$\mu_k^{\rm sHL} = \mu_k^{\rm MMSE} + \frac{1}{n_u}\sum_{m=1}^{n_u}\phi_{k,m}^2 - \frac{r}{n_u}\sum_{m=1}^{n_u}\frac{\phi_{k,m}^4}{\phi_{k,m}^2 + \frac{4\kappa_{\tilde{p}\cdot\tau}\cdot\sigma_k}{3\tilde{M}_{\tilde{p}\cdot\tau}^2\cdot\tau}\cdot f_{k,m}^2}.$$

The optimal analog combining matrix $\tilde{\boldsymbol{A}}_k^{\circ}$ is diagonal with identical diagonal entries $\left(\tilde{\boldsymbol{A}}_k^{\circ}\right)^2_{i,i} = \frac{3\tilde{M}_{\tilde{p}\cdot\tau}^2}{4\kappa_{\tilde{p}\cdot\tau}\cdot\sigma_k}$. The dynamic range is $\gamma^2 = \frac{3\tilde{M}_{\tilde{p}\cdot\tau}^2}{4}$, and the digital processing matrix is

$$\tilde{\boldsymbol{B}}_k^{\circ} = \left(\boldsymbol{D}_{k,k}^2\boldsymbol{S}^*\otimes\left(\tilde{\boldsymbol{A}}_k^{\circ}\right)^H\right)\left(\left(\boldsymbol{\Sigma}_{\underline{\boldsymbol{y}}'_k}\otimes\tilde{\boldsymbol{A}}_k^{\circ}\left(\tilde{\boldsymbol{A}}_k^{\circ}\right)^H\right) + \boldsymbol{I}_{\tau\cdot\tilde{p}}\right)^{-1}.$$

Note that $\boldsymbol{y}_k[i]$ has i.i.d. entries. Therefore, intuitively, combining the entries of $\boldsymbol{y}_k[i]$ into a lower dimension may result in an inaccurate estimation. Furthermore, it follows from Proposition 3 that $\tilde{\boldsymbol{A}}_k^{\circ}$ merely multiplies each input by a constant, whose purpose is to guarantee that the quantized entries are within the dynamic range of the uniform scalar quantizers. Consequently, when the quantizer cannot combine samples taken at different time instances in the analog domain, most of the performance gain is a result of the digital processing and the scaling of the input to be in the dynamic range.

IV. Numerical Results

In this section we numerically evaluate the performance channel estimation systems studied in Section III. We consider a massive MIMO system with $n_c = 7$ hexagonal cell of radius 400 m, with $n_u = 10$ UTs uniformly distributed in each cell. We focus on the BS of the first cell. Following [13], the attenuation coefficients are set to $d_{k,l,m} = \frac{z_{k,l,m}}{\rho_{k,l,m}^2}$, where $\{z_{k,l,m}\}$ are the shadow fading coefficients, randomized from a log-normal distribution with standard deviation of 8 dB, and $\{\rho_{k,l,m}\}$ represent the range between the mth UT of the lth cell and the kth BS. The PSs are obtained from the columns of the $\tau\times\tau$ discrete Fourier transform matrix [5, Sec. II-A].

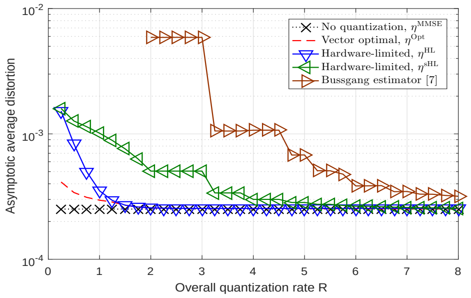

Fig. 2. Asymptotic average MSE vs. R.

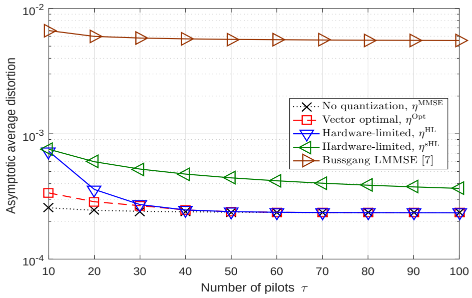

Fig. 3. Asymptotic average MSE vs. τ for $R = 2$.

We compare the performance of our proposed hardware-limited quantizers to the fundamental limit in (9), as well as to the channel estimator of [7], which extends the 1-bit Bussgang-LMMSE estimator of [5] to multiple bits. We set $\eta = 2$ and the analog combining ratio to $r = \min\left(\frac{n_u}{\tau}, \frac{R}{2}\right)$ and $r = \min\left(1, \frac{R}{2}\right)$ for the systems of Proposition 2-3, respectively. In Fig. 2 we fix the PS length to $\tau = 40$, and evaluate the achievable average MSE versus $R \in [0.5, 8]$; in Fig. 3 we fix $R = 2$ and compute the performance versus $\tau \in [10, 100]$. Observing Fig. 2, we note that the performance of the hardware-limited quantizer of Proposition 2 approaches the fundamental limits, achievable with vector quantizers, for quantization rate as small as $R = 1.5$. The hardware-limited quantizer with spatial combining approaches the fundamental limits for $R \geq 5$, and outperforms the estimator of [7] for all considered quantization rates. From Fig. 3 we note that as τ increases, $\mu^{\rm HL}$ approaches the optimal performance $\mu^{\rm Opt}$ for a fixed quantization rate R, as its analog combining ratio $\frac{n_u}{\tau}$ decreases. When this happens, uniform quantization can be carried out at more accurately for the same R, resulting in a negligible quantization error. Both hardware-limited systems significantly outperform the estimator of [7], which implements no analog combining.

The results above demonstrate the fundamental performance limits of channel estimation in massive MIMO systems, and illustrate that properly designed hardware-limited quantizers can approach these limits at relatively low quantization rates.

V. Conclusions

In this work we studied task-based quantization for massive MIMO channel estimation. We derived the average achievable MSE with vector quantization system and with scalar ADCs. Our numerical study demonstrates that our proposed quantization systems utilizing scalar ADCs can approach the fundamental limits of massive MIMO channel estimation, achievable with optimal vector quantizers.

References

[1] Y. C. Eldar. *Sampling Theory: Beyond Bandlimited Systems*. Cambridge Press, 2015.

[2] R. M. Gray and D. L. Neuhoff. “Quantization”. *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998, pp. 2325-2383.

[3] M. R. D. Rodrigues, N. Deligiannis, L. Lai, and Y. C. Eldar. “Rate-distortion trade-offs in acquisition of signal parameters”. *Proc. IEEE ICASSP*, New-Orleans, LA, Mar. 2017, pp. 6105-6109.

[4] J. Mo, P. Schniter, and R. W. Heath. “Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs”. *IEEE Trans. Signal Process.*, vol. 66, no. 5, Mar. 2018, pp. 1141-1154.

[5] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu. “Channel estimation and performance analysis of one-bit massive MIMO systems”. *IEEE Trans. Signal Process.*, vol. 65, no. 15, Aug. 2017, pp. 4075-4089.

[6] J. Choi, J. Mo, and R. W. Heath. “Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs”. *IEEE Trans. Commun.*, vol. 64, no. 5, May 2016, pp. 2005-2018.

[7] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer. “Throughput analysis of massive MIMO uplink with low-resolution ADCs”. *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, Jun.. 2017, pp. 4038 - 4051.

[8] J. Choi, J. Sung, B. L. Evans, and A. Gatherer. “Antenna selection for large-scale MIMO systems with low-resolution ADCs”. *Proc. IEEE ICASSP*, Calgary, Canada, Apr. 2018.

[9] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath. “Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate tradeoffs”. *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, Apr. 2017, pp. 2274-2287.

[10] J. Choi, B. L. Evans, and A. Gatherer. “Resolution-adaptive hybrid MIMO architectures for millimeter wave communications”. *IEEE Trans. Signal Process.*, vol. 65, no. 23, Dec. 2017, pp. 6201-6216.

[11] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek. “A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect CSI”. *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 3, Jun. 2018, pp. 484-498.

[12] T. C. Zhang, C. K. Wen, S. Jin, and T. Jiang. “Mixed-ADC massive MIMO detectors: Performance analysis and design optimization”. *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, Nov. 2016, pp. 7738-7752.

[13] T. L. Marzetta. “Noncooperative cellular wireless with unlimited numbers of base station antenna”. *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, Nov. 2010, pp. 3950–3600.

[14] N. Shlezinger and Y. C. Eldar. “On the spectral efficiency of noncooperative uplink massive MIMO systems”. *arXiv preprint*, arXiv:1712.03485, 2017.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[16] H. Witsenhausen. “Indirect rate distortion problems”. *IEEE Trans. Inform. Theory*, vol. 26, no. 5, Sep. 1980, pp. 518-521.

[17] J. K. Wolf and J. Ziv. “Transmission of noisy information to a noisy receiver with minimum distortion”. *IEEE Trans. Inform. Theory*, vol. 16, no. 4, Jul. 1970, pp. 406-411.

[18] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues. “Hardware-limited task-based quantization”. *arXiv preprint*, arXiv:1807.08305, 2018.

[19] W. B. Abbas, F. Gomez-Cuba, and M. Zorzi. “Millimeter wave receiver efficiency: A comprehensive comparison of beamforming schemes with low resolution ADCs”. *IEEE Trans. Wireless Commun.* vol. 16, no. 12, Dec. 2017, pp. 8131-8146.

[20] S. Stein and Y. C. Eldar. “Hybrid analog-digital beamforming for massive MIMO systems”. *arXiv preprint*, arXiv:1712.03485, 2017.

[21] S. Rini, L. Barlett , E. Erkip, and Y. C. Eldar. “A general framework for MIMO receivers with low-resolution quantization”. *Proc. IEEE ITW*, Kaohsiung, Taiwan, Nov. 2017.

[22] Y. Polyanskiy and Y. Wu. *Lecture Notes on Information Theory*. 2015.

[23] J. Li, N. Chaddha, and R. M. Gray. “Asymptotic performance of vector quantizers with a perceptual distortion measure”. *IEEE Trans. Inform. Theory*, vol. 45, no. 4, May 1999, pp. 1082-1091.

[24] R. M. Gray and T. G. Stockholm. “Dithered quantization”. *IEEE Trans. Inform. Theory*, vol. 39, no. 3, Mar. 1993, pp. 805-812.

[25] D. P. Palomar and Y. Jiang. *MIMO Transceiver Design via Majorization Theory*. Now Publishers, 2007.