Hardware-Limited Task-Based Quantization

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

***Abstract*—Quantization plays a critical role in digital signal processing systems. Quantizers are typically designed to obtain an accurate digital representation of the input signal, operating independently of the system task, and are commonly implemented using scalar analog-to-digital converters (ADCs). In this work, we study hardware-limited task-based quantization, where a system utilizing a serial scalar ADC is designed to provide a suitable representation in order recover a parameter vector underlying the input signal. We propose hardware-limited task-based quantization systems for a fixed and finite quantization resolution, and characterize their achievable distortion. Our results illustrate that properly designed hardware-limited systems can approach the optimal performance achievable with vector quantizers, and that by taking the underlying task into account, the quantization error can be made negligible with a relatively small number of bits.**

***Index terms*— Quantization, Analog-to-digital conversion.**

I. Introduction

Quantization refers to the representation of a continuous-amplitude signal using a finite dictionary, or equivalently, a finite number of bits [1]. Quantizers are implemented in digital signal processing systems using analog-to-digital convertors (ADCs), which typically operate in a serial scalar manner due to hardware-limitations. In such systems, each incoming continuous-amplitude sample is represented in digital form using the same mechanism [2]. The quantized representation is commonly selected to accurately match the original signal, such that the signal can be recovered with minimal error from the quantized measurements [3, Ch. 10], [4].

Quantization design is typically performed regardless of the system task. However, in many signal processing applications, the goal is not to recover the actual signal, but to capture certain underlying parameters from the quantized signal [5]. We refer to systems where one wishes to extract some information from the quantized signal, rather than recovering the signal, as *task-based quantization*, and to such systems operating with serial scalar ADCs as *hardware-limited task-based quantization* systems.

Hardware-limited quantization with low resolution is the focus of growing interest over recent years. Common applications considered with low resolution hardware-limited quantization include multiple-input multiple-output (MIMO) communications [6]–[11], channel estimation [10]–[15], subspace estimation [16], time difference of arrival estimation [17], and direction of arrival estimation [18], [19]. These works assumed that quantization is carried out *separately from the system task*, typically using *fixed uniform* low-precision quantizers. Thus, they do not provide guidelines to designing quantization systems with a small and finite number of bits by acknowledging the task of the system.

When hardware-limitations are not present, task-based quantization systems can take advantage of joint vector quantization, which is known to be superior to serial scalar quantization [20, Ch. 22.2]. When the signal parameter is random, task-based quantization can be viewed as an indirect lossy source coding problem [1, Sec. V-G]. For this setup with a stationary source that is related to the observation vector via a stationary memoryless channel, [21] showed that the rate-distortion function, namely, the minimal number of bits required to obtain a given representation accuracy, is asymptotically equivalent to the rate-distortion function for representing the observed signal with a surrogate distortion measure. Under mean-squared error (MSE) distortion, [22] proved that this equivalence also holds for finite signal size. Recently, [23], [24] characterized nonasymptotic bounds on the rate-distortion functions with arbitrary distortion measures, by considering single-shot quantization, and specialized the bounds for i.i.d. signals with separable distortion. The focus in [21]–[24] is on the *optimal* tradeoff between quantization rate and achievable distortion. Consequently, their results do not quantify the achievable performance of practical hardware-limited systems utilizing serial scalar ADCs.

In this work we study quantization for the task of acquiring a random parameter vector taking values on a continuous set, from a statistically dependent observations vector, using practical serial scalar ADCs operating with a fixed number of bits. We focus on the case where the relationship between the desired signal and the observed signal is such that the minimum MSE (MMSE) estimate is a linear function of the observations. Such relationships are commonly encountered in channel estimation and signal recovery problems, e.g., [5], [8]–[15]. We consider practical systems implementing uniform quantization with linear processing, allowing analog combining prior to digital processing. This approach was previously studied in the context of MIMO communications as a method for reducing the number of RF chains [6], [25]–[27]. For this setup, we derive the optimal hardware-limited task-based quantization system, and characterize the achievable distortion. The optimal system accounts for the task by reducing the number of quantized samples via an appropriate linear transformation to be not larger than the size of the desired signal. It then rotates the quantized samples to have identical variance. Quantization is performed based on a waterfilling-type expression, accounting for the serial operation and the limited dynamic range of practical ADCs.

We apply our results to the practical setup of channel estimation from quantized measurements [10]–[15]. We demonstrate that, by properly accounting for the serial scalar ADC, practical hardware-limited systems operating with a relatively small number of bits can approach the optimal performance,

This project has received funding from the European Union's 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 Mathematics and Computer Science, 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).

achievable with vector quantizers, in practical and relevant scenarios.

The rest of this paper is organized as follows: Section II briefly reviews some preliminaries in quantization theory, and formulates the hardware-limited task-based quantization setup. Section III derives the hardware-limited task-based quantizer, and Section IV presents a numerical study.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$; the ith element of $\boldsymbol{x}$ is written as $(\boldsymbol{x})_i$. Matrices are denoted with boldface upper-case letters, e.g., $\boldsymbol{M}$, and $(\boldsymbol{M})_{i,j}$ is its (i,j)th element. Sets are denoted with calligraphic letters, e.g., $\mathcal{X}$. Transpose, Euclidean norm, trace, expectation, and sign are written as $(\cdot)^T$, $\|\cdot\|$, $\mathrm{Tr}(\cdot)$, $\mathbb{E}\{\cdot\}$, and $\mathrm{sign}(\cdot)$, respectively, and $\mathcal{R}$ is the set of real numbers. We use a^+ to denote $\max(a,0)$, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. All logarithms are taken to basis 2.

II. Preliminaries and System Model

A. Preliminaries in Quantization Theory

To formulate the hardware-limited task-based quantization problem, we first review standard quantization notations. To that aim, we recall the definition of a quantizer:

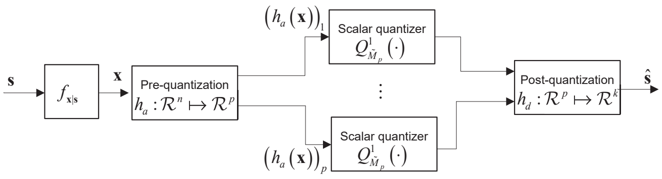
Definition 1 (Quantizer)**.** *A quantizer $Q_M^{n,k}(\cdot)$ with $\log M$ bits, input size n, input alphabet $\mathcal{X}$, output size k, and output alphabet $\hat{\mathcal{X}}$, consists of:* 1) *An encoding function $g_n^{\rm e} : \mathcal{X}^n \mapsto \{1,2,\ldots,M\} \triangleq \mathcal{M}$ which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

We write the output of the quantizer with input $\boldsymbol{x} \in \mathcal{X}^n$ as $\hat{\boldsymbol{x}} = g_k^{\rm d}\left(g_n^{\rm e}(\boldsymbol{x})\right) \triangleq Q_M^{n,k}(\boldsymbol{x})$. *Scalar quantizers* operate on a scalar input, i.e., $n = 1$ and $\mathcal{X}$ is a scalar space, while *vector quantizers* have a multivariate input. When the input and output are equally sized, i.e., $n = k$, we write $Q_M^n(\cdot) \triangleq Q_M^{n,n}(\cdot)$.

1) Standard Quantization: In the standard quantization problem, a $Q_M^n(\cdot)$ quantizer is designed to minimize some distortion measure $d_n : \mathcal{X}^n \times \hat{\mathcal{X}}^n \mapsto \mathcal{R}^+$ between its input and its output. The performance of a quantizer is therefore characterized using two measures: The quantization rate, defined as $R \triangleq \frac{1}{n}\log M$, and the expected distortion $\mathbb{E}\{d_n(\boldsymbol{x},\hat{\boldsymbol{x}})\}$. For a fixed input size n and codebook size M, the optimal quantizer is thus given by

$$Q_M^{n,\rm opt}(\cdot) = \mathop{\arg\min}_{Q_M^n(\cdot)} \mathbb{E}\left\{d_n\left(\boldsymbol{x}, Q_M^n(\boldsymbol{x})\right)\right\}. \quad (1)$$

Characterizing the optimal quantizer via (1) and the optimal tradeoff between distortion and quantization rate is in general a very difficult task. Consequently, optimal quantizers are typically studied assuming either high quantization rate, i.e., $R \to \infty$, see, e.g., [28], or asymptotically large input size, namely, $n \to \infty$, typically with i.i.d. inputs, via rate-distortion theory [3, Ch. 10]. Comparing high quantization rate analysis for scalar quantizers and rate-distortion theory for vector quantizers demonstrates the sub-optimality of serial scalar quantization. For example, for i.i.d. Gaussian inputs with the MSE distortion and large R, vector quantization outperforms serial scalar quantization by 4.35 dB [20, Ch. 23.2].



$\mathbf{s}$
$f_{\mathbf{x}|\mathbf{s}}$
$\mathbf{x}$
Pre-quantization
$h_a : \mathcal{R}^n \mapsto \mathcal{R}^p$
$\left(h_a(\mathbf{x})\right)_1$
Scalar quantizer
$Q^1_{\tilde{M}_p}(\cdot)$
$\left(h_a(\mathbf{x})\right)_p$
Scalar quantizer
$Q^1_{\tilde{M}_p}(\cdot)$
Post-quantization
$h_d : \mathcal{R}^p \mapsto \mathcal{R}^k$
$\hat{\mathbf{s}}$

Fig. 1. Hardware-limited task-based quantizer.

2) Task-Based Quantization: In task-based quantization the design objective of the quantizer is some task other than minimizing the distortion between its input and output. In the following, we focus on the generic task of acquiring a zero-mean random vector $\boldsymbol{s} \in \mathcal{R}^k$ from a statistically dependent measured zero-mean random vector $\boldsymbol{x} \in \mathcal{R}^n$, and $n \geq k > 0$. This formulation accommodates a broad range of tasks. A natural distortion measure for such setups is the MSE, which we consider henceforth.

B. System Model

In this work we study task-based quantization with serial scalar ADCs. We focus on scenarios in which the MMSE estimate of $\boldsymbol{s}$ from $\boldsymbol{x}$, $\tilde{\boldsymbol{s}} = \mathbb{E}\{\boldsymbol{s}|\boldsymbol{x}\}$, is a linear function of $\boldsymbol{x}$. Such relationships arise in various channel estimation and signal recovery setups, e.g., [5], [8]–[15]. By focusing on these setups, we are able to explicitly derive the achievable distortion and to characterize the system which achieves minimal distortion.

In the considered setup, each continuous-amplitude sample is converted into a discrete representation using a single quantization rule, this operation can be modeled using *identical scalar quantizers*. Consequently, the system we consider is modeled using the setup depicted in Fig. 1, and consists of three steps:

1) Analog Processing: The observed signal $\boldsymbol{x} \in \mathcal{R}^n$ is projected into $\mathcal{R}^p$, $p \leq n$, using some mapping $h_{\rm a}(\cdot)$, which represents the pre-quantization processing carried out in the analog domain. Since general mappings may be difficult to implement in analog, we henceforth restrict $h_{\rm a}(\cdot)$ to be a linear function, namely, we only allow *analog combining*, as in, e.g., [6], [25]. In this case, $h_{\rm a}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{A} \in \mathcal{R}^{p \times n}$.

2) Scalar Quantization: Each entry of $h_{\rm a}\big(\boldsymbol{x}\big)$ is quantized using the same scalar quantizer with resolution $\tilde{M}_p \triangleq \lfloor M^{1/p} \rfloor$, denoted $Q^1_{\tilde{M}_p}(\cdot)$. The overall number of quantization levels is thus $\big(\tilde{M}_p\big)^p \leq M$. In particular, the identical scalar quantizers $Q^1_{\tilde{M}_p}(\cdot)$ implement non-subtractive uniform dithered quantization [29]. Unlike subtractive dithered quantization, non-subtractive quantizers do not require the realization of the dithered signal to be subtracted from the quantizer output in the digital domain, resulting in a practical structure [29]. To formulate the input-output relationship of the serial 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, namely, the amplitude of the input is not larger than γ with sufficiently large probability. 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 $\kappa_p \triangleq \eta^2\big(1 - \frac{\eta^2}{3\tilde{M}_p^2}\big)^{-1}$ is strictly positive. Note that $\eta = 3$ satisfies this requirement for any $\tilde{M}_p \geq 2$, i.e., the ADC is implemented using scalar quantizers with at least one bit. The output of the serial scalar ADC with input sequence $y_1, y_2, \ldots, y_p$ can be written as $Q^1_{\tilde{M}_p}(y_i) = q_p(y_i + z_i)$, where $z_1, z_2, \ldots, z_p$ are i.i.d. random variables (RVs) uniformly distributed over $\left[-\frac{\Delta_p}{2}, \frac{\Delta_p}{2}\right]$, mutually independent of the input, representing the dither signal. The function $q_p(\cdot)$, which implements the uniform quantization, is given by

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

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

Dithered quantizers significantly facilitate the analysis, due to the following favorable properties: The output can be written as the sum of the input and an additive zero-mean white quantization noise signal, and the quantization noise is uncorrelated with the input. The drawback of adding dither is that it increases the energy of the quantization noise, namely, it results in increased distortion [29]. Nonetheless, the favorable properties of dithered quantization are also satisfied in uniform quantization *without dithering* for inputs with bandlimited characteristic function, and are approximately satisfied for various families of input distributions [30]. Consequently, while in the following analysis we assume dithered quantization, exploiting the fact that the resulting quantization noise is white and uncorrelated with the input, the proposed system can also be applied without dithering. Furthermore, as demonstrated in Section IV, applying the proposed system without dithering yields improved performance, due to the reduced energy of the quantization noise.

3) Digital Processing: The representation of $\boldsymbol{s}$, denoted $\hat{\boldsymbol{s}}$, is obtained as the output of the mapping $h_{\rm d} : \mathcal{R}^p \mapsto \mathcal{R}^k$, applied to the output of the identical scalar quantizers. The mapping $h_{\rm d}(\cdot)$ represents the joint-processing carried out in the digital domain. We restrict the digital mapping $h_{\rm d}(\cdot)$ to be linear, namely, $h_{\rm d}(\boldsymbol{u}) = \boldsymbol{B}\boldsymbol{u}$, $\boldsymbol{B} \in \mathcal{R}^{k \times p}$. This constraint leads to practical systems, and is not expected to have a notable effect on the overall performance, especially when the error due to quantization is small, since the MMSE estimator here is linear.

The novelty of the model in Fig. 1, compared to previous works on quantization for specific tasks with serial scalar ADCs, e.g., [8]–[19], is in the introduction of the additional linear processing carried out in the analog domain, represented by the mapping $h_{\rm a}(\cdot)$. The concept of using analog combining prior to digital processing was previously studied in the context of MIMO communications in [6], [7], [25]–[27]. The motivation for introducing $h_{\rm a}(\cdot)$ is to reduce the dimensionality of the input to the ADC, thus facilitating a more accurate quantization without increasing the overall number of bits, $\log M$. As shown in the following section, by properly designing $h_{\rm a}(\cdot)$, this approach can substantially improve the performance of task-based quantizers operating with serial scalar ADCs.

III. Hardware-Limited Systems Design

We now characterize the optimal hardware-limited task-based quantizer under the system model detailed in the previous section. Our characterization yields the optimal analog combining matrix and digital processing matrix, denoted $\boldsymbol{A}^{\rm o}$ and $\boldsymbol{B}^{\rm o}$, respectively, and the corresponding dynamic range γ. Since for any quantized representation $\hat{\boldsymbol{s}}$, it follows from the orthogonality principle that the MSE, $\mathbb{E}\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\}$, equals the sum of the estimation error of the MMSE estimate, $\mathbb{E}\{\|\boldsymbol{s} - \tilde{\boldsymbol{s}}\|^2\}$, and the distortion with respect to the MMSE estimate, $\mathbb{E}\{\|\tilde{\boldsymbol{s}} - \hat{\boldsymbol{s}}\|^2\}$, in the following we characterize the performance of the proposed systems via the distortion with respect to $\tilde{\boldsymbol{s}}$. The results presented in this section are given without proofs due to space limitations. Detailed proofs can be found in [31].

Let $\boldsymbol{\Gamma}$ be the MSE optimal transformation of $\boldsymbol{x}$, namely, $\tilde{\boldsymbol{s}} = \boldsymbol{\Gamma}\boldsymbol{x}$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ be the covariance matrix of $\boldsymbol{x}$, assumed to be non-singular. Before we derive the optimal hardware-limited task-based quantization system, we first derive the optimal digital processing matrix for a given analog combining matrix $\boldsymbol{A}$ and the resulting MSE, which is stated in the following lemma:

Lemma 1. *For any analog combining matrix $\boldsymbol{A}$ and dynamic range γ which guarantees that $\Pr\left(\left|\left(\boldsymbol{A}\boldsymbol{x}\right)_l + z_l\right| > \gamma\right) \approx 0$, the optimal digital processing matrix is*

$$\boldsymbol{B}^{\rm o}\left(\boldsymbol{A}\right) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T + \frac{2\gamma^2}{3\tilde{M}_p^2}\boldsymbol{I}_p\right)^{-1},$$

and the minimal achievable MSE is given by

$$\begin{aligned}{\rm MSE}\left(\boldsymbol{A}\right) &= \min_{\boldsymbol{B}} \mathbb{E}\left\{\|\tilde{\boldsymbol{s}} - \hat{\boldsymbol{s}}\|^2\right\} \\ &= {\rm Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Gamma}^T - \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T + \frac{2\gamma^2}{3\tilde{M}_p^2}{\bf I}_p\right)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Gamma}^T\right).\end{aligned}$$

The optimal digital processing matrix in Lemma 1 is the linear MMSE estimator of $\tilde{\boldsymbol{s}} = \boldsymbol{\Gamma}\boldsymbol{x}$ from the vector $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e}$ represents the quantization noise, which is white and uncorrelated with $\boldsymbol{A}\boldsymbol{x}$. This stochastic representation is a result of the usage of dithered quantizers. Additionally, it is assumed in Lemma 1 that the input to the quantizers is in the dynamic range of the quantizers, namely, $\Pr\left(\left|\left(\boldsymbol{A}\boldsymbol{x}\right)_l + z_l\right| > \gamma\right) \approx 0$ for each l, and we set the value of γ accordingly. When this requirement is not satisfied, by the law of total expectation, the resulting MSE includes an additional weighted term which accounts for working outside the dynamic range.

We now use Lemma 1 to obtain the optimal analog combining matrix $\boldsymbol{A}^{\rm o}$. Define the matrix $\tilde{\boldsymbol{\Gamma}} \triangleq \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$, and let $\{\lambda_{\tilde{\boldsymbol{\Gamma}},i}\}$ be its singular values arranged in a descending order. Note that for $i > {\rm rank}\big(\tilde{\boldsymbol{\Gamma}}\big)$, $\lambda_{\tilde{\boldsymbol{\Gamma}},i} = 0$. The optimal hardware-limited task-based quantization system is given in the following theorem:

Theorem 1. *For the optimal quantization system, the analog combining matrix is given by $\boldsymbol{A}^{\rm o} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{V}_{\boldsymbol{A}}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1/2}$, where*

- *$\boldsymbol{V}_{\boldsymbol{A}} \in \mathcal{R}^{n \times n}$ is the right singular vectors matrix of $\tilde{\boldsymbol{\Gamma}}$.*

- $\boldsymbol{\Lambda}_{\boldsymbol{A}} \in \mathcal{R}^{p\times n}$ *is a diagonal matrix with diagonal entries*
$$\left(\boldsymbol{\Lambda}_{\boldsymbol{A}}\right)^2_{i,i} = \frac{2\kappa_p}{3\tilde{M}_p^2\cdot p}\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+, \qquad \text{(2a)}$$
where ζ *is set such that* $\frac{2\kappa_p}{3\tilde{M}_p^2\cdot p}\sum\limits_{i=1}^{p}\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+ = 1.$
- $\boldsymbol{U}_{\boldsymbol{A}} \in \mathcal{R}^{p\times p}$ *is a unitary matrix which guarantees that* $\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T\boldsymbol{U}_{\boldsymbol{A}}^T$ *has identical diagonal entries, namely,* $\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T\boldsymbol{U}_{\boldsymbol{A}}^T$ *is weakly majorized by all possible rotations of* $\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T$ *[32, Cor. 2.1]. The matrix* $\boldsymbol{U}_{\boldsymbol{A}}$ *can be obtained via [32, Alg. 2.2].*

The dynamic range of the ADC is given by
$$\gamma^2 = \frac{\kappa_p}{p} = \frac{\eta^2}{p}\left(1-\frac{\eta^2}{3\tilde{M}_p^2}\right)^{-1}, \qquad \text{(2b)}$$
and the digital processing matrix is equal to
$$\boldsymbol{B}^{\rm o}\left(\boldsymbol{A}^{\rm o}\right) = \tilde{\boldsymbol{\Gamma}}\boldsymbol{V}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T\left(\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T + \frac{2\gamma^2}{3\tilde{M}_p^2}\boldsymbol{I}_p\right)^{-1}\boldsymbol{U}_{\boldsymbol{A}}^T. \qquad \text{(2c)}$$
The resulting minimal achievable distortion is
$$\mathbb{E}\left\{\left\|\tilde{\boldsymbol{s}}-\hat{\boldsymbol{s}}\right\|^2\right\} = \begin{cases}\sum\limits_{i=1}^{k}\frac{\lambda^2_{\tilde{\boldsymbol{\Gamma}},i}}{\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+ +1}, & p\geq k\\ \sum\limits_{i=1}^{p}\frac{\lambda^2_{\tilde{\boldsymbol{\Gamma}},i}}{\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+ +1} + \sum\limits_{i=p+1}^{k}\lambda^2_{\tilde{\boldsymbol{\Gamma}},i}, & p<k.\end{cases}$$

We note that, unlike task-based vector quantizers, the optimal hardware-limited system does not recover the MMSE estimate $\tilde{\boldsymbol{s}}$ in the analog domain. Since the quantization is carried out using a serial scalar ADC, the optimal analog combining rotates the input to the ADC such that each entry has *identical variance*, accounting for the fact that the same quantization rule is applied to each entry. Furthermore, the optimal analog combiner includes a waterfilling-type expression over its singular values, which accounts for the finite dynamic range of the ADC. In particular, the waterfilling allows the optimal system to reduce the quantization error by quantizing a non-bijective linear transformation of $\tilde{\boldsymbol{s}}$ instead of $\tilde{\boldsymbol{s}}$ itself. To see this, we note that the matrix $\boldsymbol{\Lambda}_{\boldsymbol{A}}$ determines the dynamic range γ. Consequently, by potentially nulling the diagonal entries corresponding to the less dominant singular values $\{\lambda_{\tilde{\boldsymbol{\Gamma}},i}\}$, the optimal quantizer can reduce the dynamic range. This yields a more precise quantization and reduces the quantization error, at the cost of a small estimation error.

Theorem 1 provides guidelines to selecting the dimensions of the output of the analog combiner, as stated in the following:

Corollary 1. *In order to minimize the MSE,* p *must not be larger than the rank of the covariance matrix of* $\tilde{\boldsymbol{s}}$.

Corollary 1 indicates that analog combining should project the observed vector such that the signal which undergoes the serial scalar quantization has reduced dimensionality, not larger than the rank of the covariance of $\tilde{\boldsymbol{s}}$. This follows since, by reducing the dimensionality of the input to the ADC while keeping the overall number of quantization levels M fixed, the quantization error induced by the scalar quantization is reduced. The exact optimal value of p is determined by the values of the non-zero singular values $\{\lambda_{\tilde{\boldsymbol{\Gamma}},i}\}$. In particular, the MSE expression in Theorem 1 implies that decreasing p below the number of non-zero singular values results in a tradeoff between improving quantization precision and increasing the estimation error. In the numerical analysis in Section IV we demonstrate that using the proposed hardware-limited task-based system, the quantization error is made negligible for relatively small M, and the performance the MMSE.

IV. Applications and Numerical Study

In this section we study the application of the proposed hardware-limited task-based quantization system. We consider the estimation of a scalar intersymbol interference (ISI) channel from quantized observations, as in [12]–[14]. In this scenario, the parameter vector $\boldsymbol{s}$ represents the coefficients of a multipath channel with k taps. The channel is estimated from a set of $n = 120$ noisy observations $\boldsymbol{x}$, given by [12, Eq. (1)]
$$\left(\boldsymbol{x}\right)_i = \sum_{l=1}^{k}\left(\boldsymbol{s}\right)_l a_{i-l+1} + w_i, \qquad i\in\{1,2,\ldots,n\}, \qquad (3)$$
where a_i is a deterministic known training sequence, and $\{w_i\}_{i=1}^n$ are samples from an i.i.d. zero-mean unit variance Gaussian noise process independent of $\boldsymbol{s}$. In particular, the channel $\boldsymbol{s}$ is modeled as a zero-mean Gaussian vector with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}}$, given by $\left(\boldsymbol{\Sigma}_{\boldsymbol{s}}\right)_{i,j} = e^{-|i-j|}$, $i,j \in \{1,2,\ldots,k\} \triangleq \mathcal{K}$, and the training sequence is given by $a_i = \cos\left(\frac{2\pi i}{n}\right)$ for $i>0$ and $a_i = 0$ otherwise. Note that $\boldsymbol{s}$ and $\boldsymbol{x}$ are jointly Gaussian, and thus the MMSE estimator $\tilde{\boldsymbol{s}}$ is a linear function of $\boldsymbol{x}$.

In the following we evaluate the achievable distortion of the resulting hardware-limited task-based quantization for this setup. To that aim, we consider two channels: one with $k=2$ taps and one with $k=8$ taps, and let the overall number of quantization bits be $\log M \in [2\cdot k, 10\cdot k]$. As $\log M$ is strictly smaller than n, any quantization system which is based on applying serial scalar quantization to the observation $\boldsymbol{x}$ without any processing in the analog domain, such as the quantization systems considered in [12], [13], cannot be implemented here.

In the numerical study, we compute the achievable distortion of the optimal system derived in Theorem 1. Since the covariance matrix of $\tilde{\boldsymbol{s}}$ is non-singular for the considered setup, we set $p=k$ following Corollary 1. Furthermore, since dithering increases the energy of the quantization noise, we also compute the achievable rate of the proposed systems when the ADCs implement uniform quantization without dithering. These distortions are compared to the MMSE $\mathbb{E}\left\{\left\|\boldsymbol{s}-\tilde{\boldsymbol{s}}\right\|^2\right\}$, which is the optimal distortion of a system with no quantization. Additionally, we also evaluate a lower bound on the MSE of the optimal vector quantizer of [22], which is given by the sum of the MMSE and the distortion-rate function of $\tilde{\boldsymbol{s}}$, see [31, Sec. III] for details.

Figs. 2-3 depict the distortions for $k=2$ and for $k=8$, respectively. Observing Figs. 2-3, we note that hardware-limited task-based quantizers approach the optimal performance as M increases. In particular, when each scalar quantizer uses at least five bits, i.e., $\log M \geq 5k$, the quantization error becomes negligible and the overall distortion is effectively the minimum achievable estimation error, i.e., the MMSE.

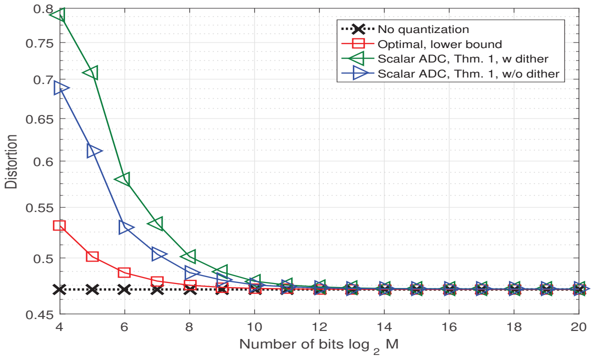


Fig. 2. Distortion comparison, channel estimation, $k = 2$.

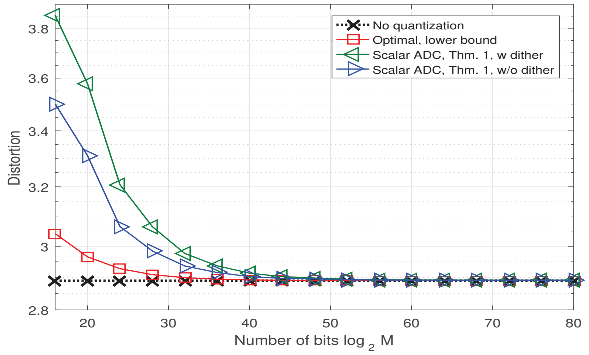


Fig. 3. Distortion comparison, channel estimation, $k = 8$.

Furthermore, we note that the proposed hardware-limited task-based quantizers, designed assuming dithered uniform quantizers, obtain improved performance without dithering. This follows since the favorable properties of dithered quantization, which are accounted for in the design of the systems in Section III, are approximately satisfied also for non-dithered standard quantization, as noted in [30], without the excess distortion induced by dithering. This illustrates that our proposed design can be applied also without dithering, and that the resulting performance is improved compared to systems implementing dithered quantization.

V. Conclusions

In this work we studied hardware-limited task-based quantization systems, operating with practical serial scalar ADCs, for finite-size signals with finite-resolution quantization. We characterized the optimal hardware-limited task-based quantizer when the MMSE estimate of the desired signal is a linear function of the observed signal. We showed that, unlike when vector quantizers are used, quantizing the MMSE estimate is generally not optimal. Finally, we applied our results to channel estimation in ISI channels, and showed that the performance of the optimal task-based vector quantizer can be approached with a practical system utilizing a scalar ADC.

References

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

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

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

[4] T. Berger and J. D. Gibson. "Lossy source coding". *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998, pp. 2693-2723.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[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] 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.

[11] 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.

[12] G. Zeitler, G. Kramer, and A. C. Singer. "Bayesian parameter estimation using single-bit dithered quantization". *IEEE Trans. Signal Process.*, vol. 60, no. 6, Jun. 2012, pp. 2713-2726.

[13] O. Dabeer and U. Madhow. "Channel estimation with low-precision analog-to-digital conversion". *Proc. IEEE ICC*, May 2010.

[14] M. S. Stein, S. Bar, J. A. Nossek, and J. Tabrikian. "Performance analysis for channel estimation with 1-bit ADC and unknown quantization threshold". *IEEE Trans. Signal Process.*, vol. 66, no. 10, May 2018, pp. 2557-2571.

[15] 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.

[16] Y. Chi and H. Fu. "Subspace learning from bits". *IEEE Trans. Signal Process.*, vol. 65, no. 17, Sep. 2017, pp. 4429-4442.

[17] R. M. Corey and A. C. Singer. "Wideband source localization using one-bit quantized arrays". *Proc. IEEE CAMSAP*, Curacao, Dutch Antilles, Dec. 2017.

[18] K. Yu, Y. D. Zhang, M. Bao, Y. Hu, and Z. Wang. "DOA estimation from one-bit compressed array data via joint sparse representation". *IEEE Signal Process. Let.*, vol. 23, no. 8, Sep. 2016, pp. 1279-1283.

[19] C. L Liu and P. P. Vaidyanathan. "One-bit sparse array DOA estimation". *Proc. IEEE ICASSP*, New-Orleans, LA, Mar. 2017.

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

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

[22] 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.

[23] V. Kostina and S. Verdu. "Nonasymptotic noisy lossy source coding". *IEEE Trans. Inform. Theory*, vol. 62, no. 11, Nov. 2016, pp. 6111-6123.

[24] V. Kostina and S. Verdu. "Fixed-length lossy compression in the finite blocklength regime". *IEEE Trans. Inform. Theory*, vol. 58, no. 3, Jun. 2012, pp. 3309-3338.

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

[26] A. AlKhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath. "MIMO precoding and combining solutions for millimeter-wave systems". *IEEE Comm. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 122-131.

[27] 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.

[28] 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.

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

[30] B. Widrow, I. Kollar, and M. C. Liu . "Statistical theory of quantization". *IEEE Trans. Inst. and Measure.*, vol. 45, no. 2, Apr. 1996, pp. 353-361.

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

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