Deep Quantization for MIMO Channel Estimation

Matan Shohat, Georgee Tsintsadze, Nir Shlezinger, and Yonina C. Eldar

***Abstract*—Quantizers play a critical role in digital signal processing systems. In practice, quantizers are typically implemented using scalar analog-to-digital converters (ADCs), commonly utilizing a fixed uniform quantization rule which is ignorant of the task of the system. Recent works have shown that the performance of quantization systems utilizing scalar ADCs can be significantly improved by properly processing the analog signal prior to quantization. However, the implementation of such systems requires complete knowledge of the underlying model, which may not be available in practice. In this work we design task-oriented quantization systems with scalar ADCs using deep learning, focusing on the task of multiple-input multiple-output (MIMO) channel estimation. By utilizing deep learning, we construct a task-based quantization system, overcoming the need to explicitly recover the system model and to find the proper quantization rule for it. Our results indicate that the proposed method results in practical MIMO systems with scalar ADCs which are capable of approaching the optimal performance limits dictated by indirect rate-distortion theory, achievable using vector quantizers and requiring complete knowledge of the underlying statistical model.**

***Index terms*— Quantization, deep learning, channel estimation.**

I. Introduction

Digital signal processing systems typically require a finite-dictionary representation of continuous-amplitude analog signals. The mapping of an analog signal into a digital representation with a finite number of bits is referred to as *quantization* [1]. This representation is commonly selected to accurately match the quantized signal, in the sense of minimizing some distortion measure, such that the signal can be recovered with minimal error from the quantized measurements [2], [3, Ch. 10]. In many relevant scenarios, the task of the system is to recover some underlying parameters, and not to accurately represent the observed signal. In these cases, it was shown that by accounting for the system task in the design of the quantizers, namely, by utilizing *task-based quantization*, the performance can be improved without increasing the number of bits used [4], [5].

In practice, quantizers are typically implemented using analog-to-digital converters (ADCs), which operate on the input signal in a serial scalar manner. In such systems, the quantization rule is based on a uniform partition of a subspace of the real line, determined by the dynamic range of the quantizer. This quantization logic is very limited due to its simplicity, hence, with the exception of the specific case where the input is uniformly distributed over the dynamic range of the quantizer, uniform quantization is far from optimality [6, Sec. 22], namely, a more accurate representation can be obtained with the same number of bits. Furthermore, such quantizers typically do not account for the system task, namely, they are *task-ignorant*.

Quantizers are inherently non-linear systems, thus, the design and implementation of practical quantizers which provide an accurate discrete representation while accounting for the task of the system, is in general difficult for the following reasons: 1) It requires complete knowledge of the stochastic model of the underlying signal [1], [2], which may be unavailable in practice, and 2) Even when the stochastic model is perfectly known, the scalar quantization rule which minimizes the representation error is generally unknown for most distributions under finite resolution quantization [6, Ch. 23.1].

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

M. Shohat and G. Tsintsadze are with the department of EE, Technion, Haifa, Israel ({matan.shohat; tsintsadze}@campus.technion.ac.il).

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).

A promising approach to efficiently implement task-based quantizers without the need to explicitly know the underlying model and to analytically derive the proper quantization rule for it is to use deep learning algorithms. Existing works on deep learning for quantization typically focus on image compression [7]–[10], where the goal is to represent the analog image using a single quantization rule, i.e., non task-based vector-quantization. Alternatively, a large body of deep learning related works consider deep neural network (DNN) model compression [11]–[13], where a DNN operates with quantized instead of continuous weights. To the best of our knowledge, despite the importance of quantization with scalar ADCs in digital signal processing, the application of deep learning in such systems has not yet been studied.

In this paper we consider the application of DNNs for task-based quantization, utilizing practical ADCs. Since continuous-to-discrete mappings applied in the quantization process are inherently non-differentiable, standard deep learning training algorithms, such as stochastic gradient descent (SGD), cannot be applied in a straightforward manner. To overcome this difficulty, we propose two methods for applying deep learning to train task-based quantizers operating with scalar ADCs.

Then, we focus on the problem of channel estimation from quantized observations in multi-user multiple-input multiple-output (MIMO) communications, a setup representing common wireless communications networks. In these scenarios there is an urgent need for efficient low resolution quantization, due to the increasing complexity and bitrate demands of modern communications [14]–[16]. We compare the performance our our proposed DNN-based system to previous channel estimators from task-ignorant quantized measurements, as well as to the optimal task-based estimator proposed in our previous work [5]. We also compare with the fundamental limits on channel estimation performance in MIMO systems with quantized observations, derived using rate-distortion theory which is achievable using optimal vector quantizers [6, Ch. 23]. Our results demonstrate that, even when the DNN-based quantizer is trained with samples taken from setups with different signal-to-noise ratio (SNR), it is still able to approach the performance of the optimal task-based quantizers with ADCs for varying SNRs, which is within a small gap of the fundamental performance limits.

The rest of this paper is organized as follows: Section II introduces the problem formulation; Section III discusses the implementation of scalar quantization systems using DNNs; Section IV presents its application to MIMO channel estimation.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$. Matrices are denoted with boldface upper-case letters, e.g., $\boldsymbol{M}$. Sets are denoted with calligraphic letters, e.g., $\mathcal{X}$. We use $\boldsymbol{I}_n$ to denote the $n \times n$ identity matrix and $\otimes$ as the symbol for the kronecker product. Transpose, Euclidean norm, stochastic expectation, real part, and imaginary part are written as $(\cdot)^T$, $\|\cdot\|$, $\mathbb{E}\{\cdot\}$, $\mathrm{Re}\,(\cdot)$, and $\mathrm{Im}\,(\cdot)$, respectively, $\mathcal{R}$ is the set of real numbers, and $\mathcal{C}$ is the set of complex numbers.

II. Preliminaries and Problem Statement

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 [5], 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 $f_n : \mathcal{X}^n \mapsto \{1, 2, \ldots, M\} \triangleq \mathcal{M}$ which maps the input into a discrete index.* 2) *A decoding function $g_k : \mathcal{M} \mapsto \hat{\mathcal{X}}^k$ 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 $\hat{\boldsymbol{x}} = g_k\left(f_n\left(\boldsymbol{x}\right)\right) \triangleq Q_M^{n,k}\left(\boldsymbol{x}\right)$. *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 size and the output size are equal, $n = k$, we write $Q_M^n\left(\cdot\right) \triangleq Q_M^{n,n}\left(\cdot\right)$.

In the standard quantization problem, a $Q_M^n\left(\cdot\right)$ 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 characterized using two measures: the quantization rate, defined in the standard quantization setup as $R \triangleq \frac{1}{n}\log M$, and the expected distortion $\mathbb{E}\{d_n\left(\boldsymbol{x}, \hat{\boldsymbol{x}}\right)\}$. For a fixed input size n and codebook size M, the optimal quantizer is

$$Q_M^{n,\mathrm{opt}}\left(\cdot\right) = \arg\min_{Q_M^n(\cdot)} \mathbb{E}\left\{d_n\left(\boldsymbol{x}, Q_M^n\left(\boldsymbol{x}\right)\right)\right\}. \tag{1}$$

Characterizing the optimal quantizer via (1) and the optimal trade-off between distortion and quantization rate is in general a very difficult task. Optimal quantizers are thus typically studied assuming either high quantization rate, i.e., $R \to \infty$, see, e.g., [17], or asymptotically large inputs, namely, $n \to \infty$, typically with i.i.d. inputs, via rate-distortion theory [3, Ch. 10].

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 measured zero-mean random vector $\boldsymbol{x} \in \mathcal{R}^n$. This formulation accommodates a broad range of tasks, including channel estimation, covariance estimation, and source localization [4], [5]. For task-based quantization with the the mean-squared error (MSE) distortion, i.e., $d(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2$, it was shown in [18] that the optimal quantizer applies standard quantization to the MMSE estimate of the desired vector $\boldsymbol{s}$ from the observed vector $\boldsymbol{x}$. While the optimal system utilizes vector quantization, its structure indicates that processing the observations in the analog domain is beneficial in task-based quantization.

B. Problem Statement

As discussed in the introduction, practical digital signal processing systems typically obtain a digital representation of physical analog signals using serial scalar ADCs. Since in such systems, each continuous-amplitude sample is converted into a discrete representation using a single quantization rule, this operation can be modeled using *identical scalar quantizers*. In this work we study the implementation of task-based quantization systems with serial scalar ADCs using DNNs.

In particular, the considered task-based deep quantization system with scalar ADCs is modeled using the setup depicted in Fig. 1. We consider the recovery of a vector $\boldsymbol{s} \in \mathcal{R}^k$ based on an observed vector $\boldsymbol{x} \in \mathcal{R}^n$ quantized with up to $\log M$ bits. The observed $\boldsymbol{x}$ is related to $\boldsymbol{s}$ via a conditional probability measure $f_{\boldsymbol{x}|\boldsymbol{s}}$, which is assumed to be unknown. The input to the ADC, denoted $\boldsymbol{z} \in \mathcal{R}^p$, is obtained from $\boldsymbol{x}$ using some pre-quantization mapping carried out in the analog domain. Then, $\boldsymbol{z}$ is quantized using an ADC modeled as p identical scalar quantizers with resolution $\tilde{M} \triangleq \lfloor M^{1/p} \rfloor$. The overall number of bits is $p \cdot \log \tilde{M} \leq \log M$. The ADC output is processed in the digital domain to obtain the quantized representation $\hat{\boldsymbol{s}} \in \mathcal{R}^k$.

In the following, we implement the pre and post quantization processings using dedicated DNNs jointly trained in an end-to-end manner. We elaborate on the design of these networks in the following section. By utilizing DNNs, we expect the resulting system to be able to approach the optimal achievable distortion for fixed quantization rate and input size, without requiring knowledge of the underlying distribution. Such performance is illustrated in the numerical example presented in Section IV.

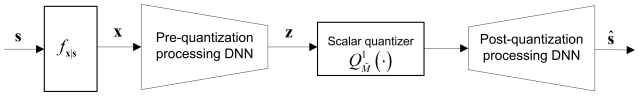


Fig. 1. Deep scalar quantization system model.

III. Deep Scalar Quantization

We now discuss the implementation of the system in Fig. 1 using DNNs. As common in supervised learning, we assume that a labeled data set is given in advance. The training data consists of t independent realizations of $\boldsymbol{s}$ and $\boldsymbol{x}$, denoted $\{\boldsymbol{s}^{(i)}, \boldsymbol{x}^{(i)}\}_{i=1}^t$. It is emphasized that, in general, the training set may taken from a set of joint distributions containing the true (unknown) joint distribution of $\boldsymbol{s}$ and $\boldsymbol{x}$.

Here, the serial scalar ADC which implements the continuous-to-discrete mapping is modeled as an activation function between the two intermediate layers. The system input is the $n \times 1$ observed vector $\boldsymbol{x}$. By letting θ be a vector representing the tunable parameters of the DNN and $q_\theta(\cdot)$ denote the mapping implemented by overall system, the output is given by the $k \times 1$ vector $\hat{\boldsymbol{s}} = q_\theta(\boldsymbol{x})$. The output $\hat{\boldsymbol{s}}$ is used as a representation of the desired vector $\boldsymbol{s}$.

Since the task of the system is to recover $\boldsymbol{s}$, the loss function is the empirical MSE, given by

$$\mathcal{L}(\theta) = \frac{1}{t}\sum_{j=1}^{t}\left\|\boldsymbol{s}^{(j)} - q_\theta\left(\boldsymbol{x}^{(j)}\right)\right\|_2^2. \tag{2}$$

The network is trained to minimize the loss in (2) using the SGD optimization algorithm in an end-to-end manner. Specifically, the pre-qunatization DNN, representing the processing carried out in the analog domain, is jointly trained with the post-quantization DNN to minimize the loss in (2).

Note that we use scalar quantization as an intermediate activation in the system. The non-differentiable nature of such continuous-to-discrete mappings induces a major challenge in applying SGD in the presence of such activations. In particular, quantization activation, which consists of a superposition of step functions, nullifies the gradient of the cost function. Consequently, straight-forward application of SGD fails to properly set the pre-quantization network. To overcome this drawback, we consider two approaches, referred to henceforth as *passing gradient* and *soft-to-hard quantization*.

A. Passing Gradient Quantization

We first present a naive approach which overcomes the fact that, using SGD, the pre-quantization layers cannot be tunned, by essentially ignoring the presence of quantization during training. Here, the training algorithm passes the gradient value through the quantization activation layer. An illustration of this approach is depicted in Fig. 2(a). We expect the resulting system to obtain poor performance when non-negligible distortion is induced by the quantizers. In our numerical study presented in Subsection IV-C, it is illustrated that this method indeed achieves relatively poor performance, as scalar quantization induces an error term which cannot be ignored. It is therefore desirable to formulate a network structure which properly accounts for the presence of the scalar quantizers during training.

B. Soft-to-Hard Quantization

The second approach to deal with the described problem is based on approximating the non-differentiable quantization mapping by a

differentiable one. Here, we replace the continuous-to-discrete mapping with a non-linear activation function which has approximately the same behavior as the quantization function. Specifically, we use a sum of shifted hyperbolic tangents, as such functions are known to closely resemble step functions in the presence of large magnitude inputs. The resulting scalar quantization mapping is given by:

$$\tilde{q}_{\tilde{M}}(x) = \sum_{i=1}^{\tilde{M}-1} a_i \tanh\left(c_i \cdot x - b_i\right), \tag{3}$$

where $\{a_i, b_i, c_i\}$ are a set of real-valued parameters. Note that as the parameters $\{c_i\}$ increase, the corresponding hyperbolic tangents approach step functions. Since we use a differentiable activation to approximate a set of non-differentiable functions, as in [7], we refer to this method *soft-to-hard quantization*.

In addition to learning the DNN weights, here we let the DNN also learn its activation function, and particularly, the best suitable constants $\{a_i\}$ (the amplitudes) and $\{b_i\}$ (the shifts). These tunable parameters are later used to determine the decision regions of the resulting scalar quantizer. The parameters $\{c_i\}$, which essentially control the resemblance of (3) to an actual continuous-to-discrete mapping, can be either fixed, or alternatively, modified using annealing-based optimization [19], where $\{c_i\}$ are manually increased during training. The proposed optimization is achieved by manually defining these parameters as part of the network parameters. Due to the differentiability of (3), one can apply standard SGD to optimize the overall network parameters.

Once training is concluded, we replace the learned $\tilde{q}_{\tilde{M}}(x)$ activation with a scalar quantizer whose decision regions are dictated by the tunable parameters $\{a_i, b_i\}$. In particular, since $\tanh(c \cdot x - b) = 0$ for $x = \frac{b}{c}$, we use the set $\left\{\frac{b_i}{c_i}\right\}$ to determine the decision region of the quantizer, and use the value of $\tilde{q}_{\tilde{M}}(x)$ at each decision region center as its corresponding representation level. Without loss of generality, we assume that $\frac{b_0}{c_0} \leq \frac{b_1}{c_1} \leq \ldots \leq \frac{b_{\tilde{M}-1}}{c_{\tilde{M}-1}}$. The resulting quantizer is given by

$$Q^1_{\tilde{M}}(x) = \begin{cases} -\sum_{i=1}^{\tilde{M}-1} a_i & x \leq \frac{b_0}{c_0} \\ \tilde{q}_{\tilde{M}}\left(\frac{\frac{b_i}{c_i} + \frac{b_{i+1}}{c_{i+1}}}{2}\right) & \frac{b_i}{c_i} < x \leq \frac{b_{i+1}}{c_{i+1}} \\ \sum_{i=1}^{\tilde{M}-1} a_i & \frac{b_{\tilde{M}-1}}{c_{\tilde{M}-1}} < x. \end{cases} \tag{4}$$

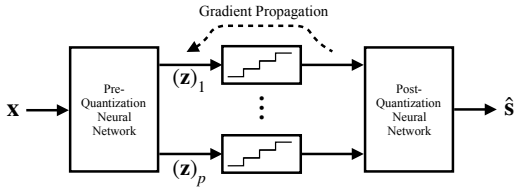
In the simulations presented in Subsection IV-C, it is illustrated that the proposed method, which is capable of accounting for the presence of scalar quantizers during training, can approach the performance of the optimal task-based quantizer with scalar ADCs of [5], which requires complete knowledge of the underlying model, in a MIMO channel estimation scenario.

IV. Application to MIMO Channel Estimation

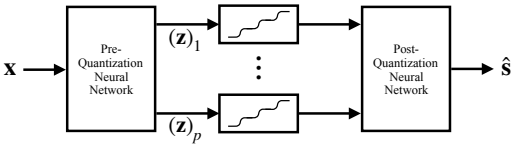
In the following we apply our proposed deep task-based quantizer for channel estimation in multi-user MIMO communications. We first formulate the setup in Subsection IV-A. Then, in Subsection IV-B, we discuss the theoretical performance bounds. Finally, in Subsection IV-C, we numerically compare the achievable distortion of our proposed system to the performance of previously proposed systems as well as to the fundamental performance limits.

A. MIMO Channel Estimation

The problem of MIMO channel estimation with low resolution quantization is the focus of many recent works, including, e.g., [14]–[16]. We consider channel estimation in a single cell baseband multi-user MIMO system, in which n_u single antenna users are served by a base station (BS) with n_t antennas. Channel estimation is carried out in a time diversity duplexing manner using orthogonal pilot sequences



(a) Passing Gradient quantization system



(b) Soft-to-Hard quantization system

Fig. 2. Task-based deep quantization architectures

of length $\tau_p \geq n_u$ with SNR P. Let $\boldsymbol{\Phi} \in \mathcal{C}^{\tau_p \times n_u}$ denote the known pilot sequence matrix, where the orthogonality of the pilots implies that $\boldsymbol{\Phi}^H \boldsymbol{\Phi} = \tau_p \cdot \boldsymbol{I}_{n_u}$. Additionally, let $\boldsymbol{h} \in \mathcal{C}^{n_u \cdot n_t}$ be a random vector whose entires are the i.i.d. zero-mean unit-variance complex normal channel coefficients, and $\boldsymbol{w} \in \mathcal{C}^{\tau_p \cdot n_t}$ be a random vector with i.i.d. zero-mean unit-variance complex normal entries mutually independent of $\boldsymbol{h}$, representing the additive noise at the BS. The observed signal $\boldsymbol{y} \in \mathcal{C}^{\tau_p \cdot n_t}$, used by the BS to estimate $\boldsymbol{h}$, can be written as [15, Eq. (4)]:

$$\boldsymbol{y} = \sqrt{P}\left(\boldsymbol{\Phi} \otimes \boldsymbol{I}_{n_t}\right)\boldsymbol{h} + \boldsymbol{w}. \tag{5}$$

To put the setup in (5) in the framework of our problem formulation, which considers real-valued signals, we write the observations as $\boldsymbol{x} = \left[\mathrm{Re}\left(\boldsymbol{y}\right)^T, \mathrm{Im}\left(\boldsymbol{y}\right)^T\right]^T$ and the unknown channel as $\boldsymbol{s} = \left[\mathrm{Re}\left(\boldsymbol{h}\right)^T, \mathrm{Im}\left(\boldsymbol{h}\right)^T\right]^T$. Consequently, the number of measurements is $n = 2 \cdot \tau_p \cdot n_t$, the number of unknown parameters is $k = 2 \cdot n_u \cdot n_t$, and their ratio is $\rho = \frac{\tau_p}{n_u} \geq 1$. The performance measure for evaluating quantization systems here is the average MSE, namely, $\eta = \frac{1}{k}\mathbb{E}\left\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\right\}$.

B. Theoretical Performance

As a basis for comparison, we review the fundamental performance limits for this setup dictated by indirect rate-distortion theory. To formulate these limits, note that it follows from (5) that: 1) $\boldsymbol{s}$ and $\boldsymbol{x}$ are zero-mean jointly Gaussian random vectors; 2) the covariance matrix of $\boldsymbol{x}$ can be written as $\boldsymbol{\Sigma}_{\underline{\boldsymbol{x}}} \otimes \boldsymbol{I}_{n_t}$, with

$$\boldsymbol{\Sigma}_{\underline{\boldsymbol{x}}} = \frac{1}{2}\begin{bmatrix} \mathrm{Re}\left(P \cdot \boldsymbol{\Phi}\boldsymbol{\Phi}^H + \boldsymbol{I}_{\tau_p}\right), & -\mathrm{Im}\left(P \cdot \boldsymbol{\Phi}\boldsymbol{\Phi}^H + \boldsymbol{I}_{\tau_p}\right) \\ \mathrm{Im}\left(P \cdot \boldsymbol{\Phi}\boldsymbol{\Phi}^H + \boldsymbol{I}_{\tau_p}\right), & \mathrm{Re}\left(P \cdot \boldsymbol{\Phi}\boldsymbol{\Phi}^H + \boldsymbol{I}_{\tau_p}\right) \end{bmatrix};$$

and 3) the minimum mean-squared error (MMSE) estimate of $\boldsymbol{s}$ from $\boldsymbol{x}$ is $\tilde{\boldsymbol{s}} \triangleq \mathbb{E}\{\boldsymbol{s}|\boldsymbol{x}\} = \boldsymbol{\Gamma}\boldsymbol{x} = \left(\underline{\boldsymbol{\Gamma}} \otimes \boldsymbol{I}_{n_t}\right)\boldsymbol{x}$, where

$$\underline{\boldsymbol{\Gamma}} = \frac{\sqrt{P}}{1 + P \cdot \tau_p}\begin{bmatrix} \mathrm{Re}\left(\boldsymbol{\Phi}^H\right), & -\mathrm{Im}\left(\boldsymbol{\Phi}^H\right) \\ \mathrm{Im}\left(\boldsymbol{\Phi}^H\right), & \mathrm{Re}\left(\boldsymbol{\Phi}^H\right) \end{bmatrix}. \tag{6}$$

The covariance of the MMSE estimate $\tilde{\boldsymbol{s}}$ is therefore given by $\left(\underline{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}_{\underline{\boldsymbol{x}}}\underline{\boldsymbol{\Gamma}}^T \otimes \boldsymbol{I}_{n_t}\right) = \frac{P \cdot \tau_p}{2(1 + P \cdot \tau_p)}\left(\boldsymbol{I}_{2n_u} \otimes \boldsymbol{I}_{n_t}\right)$, thus the entries of $\tilde{\boldsymbol{s}}$ are i.i.d. zero-mean Gaussian random variables with variance $\frac{P \cdot \tau_p}{2(1 + P \cdot \tau_p)}$. Based on the above, the average MMSE, which is the optimal performance achievable with no quantization constraints, is given by $\tilde{\eta} = \frac{1}{2(1 + P \cdot \tau_p)}$. In the presence of quantization constraints, the optimal approach is to quantize the MMSE estimate [18], and the resulting average distortion is obtained from rate-distortion theory [3, Ch. 10.3] as

$$\eta_{\mathrm{opt}} = \tilde{\eta} + \frac{P \cdot \tau_p}{2(1 + P \cdot \tau_p)} 2^{-2\rho \cdot R}. \tag{7}$$

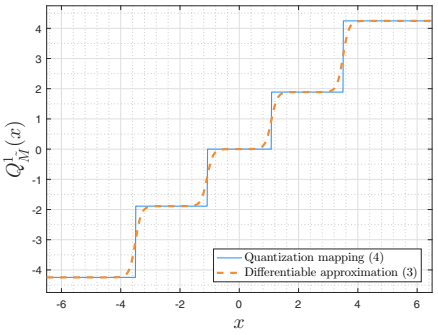


Fig. 3. Soft-to-hard quantization rule illustration.

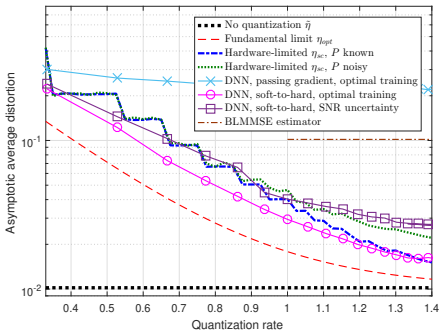


Fig. 4. Numerical results versus theoretical measures.

Note that η_{opt} is achievable using optimal vector quantization in the limit $n_t \to \infty$. For finite n_t and scalar quantizers, (7) serves as a lower bound on the achievable performance. We thus refer to η_{opt} as the *fundamental performance limit*.

C. Numerical Study

We now numerically evaluate our proposed deep task-based quantizers detailed in Section III in a MIMO channel estimation setup. The numerical performance is compared to the fundamental performance limit in (7), as well as to the performance of the task-based quantizer with scalar uniform ADCs designed in [5], denoted η_{sc}, which is optimal in scenarios where the MMSE estimate is linear. It is noted that while our proposed system can modify the quantization regions, the model of [5] assumes fixed uniform quantizers. Consequently, the average MSE of the system of [5] does not necessarily lower bound the performance of our proposed system. We also note that the system of [5] requires full knowledge of the underlying statistical model.

We simulate a multi-user MIMO network in which a BS equipped with $n_t = 10$ antennas serves $n_u = 4$ users. We set the SNR to be $P = 4$ and the number of pilots to $\tau_p = 12$. As in [15], we fix the pilots matrix $\mathbf{\Phi}$ to be the first n_u columns of the $\tau_p \times \tau_p$ discrete Fourier transform matrix. In the implementation the deep quantizers, we set the pre and post quantization DNNs to consist of linear layers. The motivation for using linear layers stems from the fact that for the considered setup, the MMSE estimate is a linear function of the observations. Following [5, Cor. 1], we evaluate the average MSE of our proposed systems with $p = k$ quantizers. We consider two training sets, both of size $t = 2^{15}$: In the first training set, representing *optimal training*, the realizations $\{\boldsymbol{s}^{(i)}, \boldsymbol{x}^{(i)}\}_{i=1}^t$ are sampled from the true joint distribution of $\boldsymbol{s}, \boldsymbol{x}$; In the second training set, representing *SNR uncertainty*, $\{\boldsymbol{s}^{(i)}, \boldsymbol{x}^{(i)}\}_{i=1}^t$ are sampled from the joint distribution of $\boldsymbol{s}, \boldsymbol{x}$ with different values of P, uniformly randomized over the set $[1, 10]$ for each realization. We use an SGD optimizer with momentum 0.5 and learning rate decaying factor of 0.7. In the *passing gradient* method we use uniform codebook under the dynamic range $\boldsymbol{z}_k^i \in [-3, 3]$ for each i, k. In the *soft-to-hard* method we randomize the initial values of $\{a_i\}, \{b_i\}$ from a standard normal distribution. The parameters $\{c_i\}$ are fixed to $c_i = 5$ for each $i \in \{1, \ldots, p\}$. At the end of the training session, we fix the quantizer to implement the continuous-to-discrete rule in (4). An illustration of such mapping is depicted in Fig. 3, where the dashed smooth curve represents the differentiable function after training is concluded, and the straight curve is the resulting scalar quantizer. We numerically evaluate the generalization error of our proposed deep quantizers using 2^{10} realizations.

In Fig. 4 we depict the resulting performance versus the quantization rate $R = \frac{1}{n} \log M$ in the range $R \in [0.33, 1.4]$. The empirical performance is compared to the theoretical measures, representing the MMSE, the fundamental performance limits of channel estimation from quantized measurements (7), and the performance of the optimal task-based quantizer with scalar ADCs [5]. Since [5] requires perfect knowledge of the underlying model, which may not be available in practice, we also consider the case where [5] utilizes an estimation of P corrupted by zero-mean Gaussian noise with variance $\frac{P}{4}$. Furthermore, we also compute the average MSE of the BLMMSE estimator proposed in [15] via [15, Eq. (15)]. Since the BLMMSE estimator quantizes the observed signal without analog pre-processing, it is applicable only for $R \geq 1$.

Observing Fig. 4, we note the gap in performance between the two considered training methods. While the passing gradient fails to approach optimal result, the performance of our soft-to-hard deep quantizer is within a small gap of the fundamental performance limits. Furthermore, the fact that the soft-to-hard method is not restricted to uniform quantizers allows it to outperform η_{sc}, especially in lower quantization rates. Finally, we note that in the presence of SNR uncertainty, the performance of the soft-to-hard method is similar to η_{sc} with noisy SNR estimate, and that both outperform the BLMMSE estimator of [15]. This indicates that our proposed scheme is applicable also when the training data is not generated from the exact same distribution as the test data. Our results demonstrate that feasible and optimal-approaching quantization systems can be implemented using DNNs in practical communications setups.

V. Conclusions

In this work we designed task-based quantization systems, operating with serial scalar ADCs, using DNNs. We studied two methods for handling the non-differentiability of quantization. Our numerical results, which considered MIMO channel estimation, demonstrated that even for a very simple network structure, the performance achievable with our proposed *soft-to-hard* method for training the network is comparable with the fundamental limits for this setup, achievable using optimal vector quantizers.

References

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

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

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

[4] 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.

[5] N. Shlezinger, Y. C. Eldar and M. R. D. Rodrigues. "Hardware-Limited Task-Based Quantization". arXiv preprint, arXiv:1807.08305, 2018.

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

[7] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. van Gool. "Soft-to-hard vector quantization for end-to-end learning compressible representations". *Proc. NIPS*, Long Beach, CA, 2017, pp. 1141–1151.

[8] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell. "Full resolution image compression with recurrent neural networks". *Proc. IEEE CVPR*, Honolulu, Hi, 2017.

[9] J. Balle, V. Laparra, and E. P. Simoncelli. "End-to-end optimized image compression". arXiv preprint, arXiv:1611.01704, 2016.

[10] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. J. Hwang, J. Shor, and G. Toderici. "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks". arXiv preprint, arXiv:1703.10114, 2017.

[11] S. Han, H. Mao, and W. J. Dally. "Compressing deep neural networks with pruning, trained quantization and huffman coding". arXiv preprint, arXiv:1510.00149, 2015.

[12] K. Ullrich, E. Meeds, and M. Welling. "Soft weight-sharing for neural network compression". arXiv preprint, arXiv:1702.04008, 2017.

[13] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. "Quantized neural networks: Training neural networks with low precision weights and activationss". *Journal of Machine Learning Research*, vol. 187, no. 18, Apr. 2018, pp. 1-30.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] K. Rose, E. Gurewitz, and G. C. Fox. "Vector quantization by deterministic annealing". *IEEE Trans. Inform. Theory*, vol. 38, no. 4, Apr. 1992, pp. 1249-1257.