Article

Deep Task-Based Quantization †

Nir Shlezinger [1,*] and **Yonina C. Eldar** [2]

1. School of Electrical and Computer Engineering, Ben-Gurion University of the Negev,
   Beer-Sheva 8410501, Israel
2. Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel;
   yonina.eldar@weizmann.ac.il
* Correspondence: nirshl@bgu.ac.il
† This paper is an extended paper presented in the 2019 IEEE International Conference on Acoustics, Speech
   and Signal Processing (ICASSP) in Brighton, UK, on 12–17 May 2019.

Abstract: Quantizers play a critical role in digital signal processing systems. Recent works have shown that the performance of acquiring multiple analog signals using scalar analog-to-digital converters (ADCs) can be significantly improved by processing the signals prior to quantization. However, the design of such hybrid quantizers is quite complex, and their implementation requires complete knowledge of the statistical model of the analog signal. In this work we design data-driven task-oriented quantization systems with scalar ADCs, which determine their analog-to-digital mapping using deep learning tools. These mappings are designed to facilitate the task of recovering underlying information from the quantized signals. By using deep learning, we circumvent the need to explicitly recover the system model and to find the proper quantization rule for it. Our main target application is multiple-input multiple-output (MIMO) communication receivers, which simultaneously acquire a set of analog signals, and are commonly subject to constraints on the number of bits. Our results indicate that, in a MIMO channel estimation setup, the proposed deep task-bask quantizer is 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. Furthermore, for a symbol detection scenario, it is demonstrated that the proposed approach can realize reliable bit-efficient hybrid MIMO receivers capable of setting their quantization rule in light of the task.

Keywords: analog-to-digtal conversion; task-based quantization; deep learning

1. Introduction

Digital signal processing systems operate on finite-bit representation of continuous-amplitude physical signals. The mapping of an analog signal into a digital representation of a finite dictionary is referred to as *quantization* [1]. This representation is commonly selected to accurately match the quantized signal, in the sense of minimizing a 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 using *task-based quantization*, the accuracy in carrying out the task can be improved without increasing the number of bits used [4–7]. Such task-based quantization was shown to improve performance in channel estimation [8,9] and target identification in multiple-input multiple-output (MIMO) radar [10], when operating under tight bit budgets.

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 support of the quantizer. This quantization logic is very limited due to its simplicity: except for the specific case where the input is uniformly distributed over the support of the quantizer, uniform quantization is far from optimality ([11], 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*. While the distortion induced by such inefficient quantization can be mitigated by assigning more bits for digital representation, i.e., using high-resolution quantizers, it can severely degrade the performance of bit-constrained systems.

Recent years have witnessed a growing interest in systems operating with low-resolution ADCs. In particular, the power consumption of ADCs typically grows with the bandwidth and the quantization resolution [12]. To maintain feasible cost and power usage when acquiring multiple signals at large frequency bands, low-resolution quantizers may be used. An example where such bit-constrained systems are popular is MIMO communication receivers, which simultaneously acquire and process multiple analog signals in order to recover the transmitted symbols and/or estimate the underlying channel, i.e., for a specific task. MIMO receivers operating at large spectral bands, e.g., millimeter wave systems [13], are commonly designed to acquire the channel output with low-resolution quantizers, and a large body of work focuses on schemes for carrying out the aforementioned tasks from coarsely discretized measurements, see, e.g., [14–20].

Quantizers are inherently non-linear systems. Hence, the design and implementation of practical quantizers which provide an accurate discrete representation while accounting for the system task, is difficult in general. Two notable challenges are associated with designing such task-based quantization systems: (1) In order to design the quantization scheme, one must have full knowledge of the stochastic model of the underlying signal [1,2], which may be unavailable in practice; (2) Even when the stochastic model is perfectly known, the scalar continuous-to-discrete rule which minimizes the representation error is generally unknown for most distributions under finite resolution quantization ([11], Ch. 23.1). A possible approach to tackle the second challenge is to use a uniform quantization rule, while applying additional processing in analog prior to quantization, resulting in an analog-digital hybrid system [21,22]. While such hybrid systems were shown to result in substantially improved performance for signal recovery tasks under bit constraints [5–7,9], their design is commonly restricted to a subset of analog mappings, e.g., linear processing [5,7]; and specific stochastic models, such as Gaussian observations [6,9]. Furthermore, these model-based quantization systems assume uniform quantizers, hence, they do not exploit the ability to use arbitrary quantization rules, while requiring accurate knowledge of the underlying statistical model.

An alternative approach to inferring the quantization system from the model, is to learn it from a set of training samples in a data-driven fashion. In particular, by using machine learning methods, one can implement task-based quantizers without the need to explicitly know the underlying model and to analytically derive the proper quantization rule. Existing works on deep learning for quantization typically focus on image compression [23–27], where the goal is to represent the analog image using a single quantization rule from which the image can be recovered and/or processed [28]. Alternatively, a large body of deep learning related works consider deep neural network (DNN) model compression [29–31], where a DNN operates with quantized instead of continuous weights. The work [32] used DNNs to compress and quantize high-dimensional channel state information in a massive MIMO feedback setup. The design of DNNs for processing one-bit quantized measurements in the digital domain, i.e., in the presence of task-ignorant quantizers, was considered for signal recovery in [33]; while DNN-based MIMO receivers with one-bit quantizers were studied in [34,35]. 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 design of data-driven task-based quantizers, using scalar ADCs. Following [5,6], we propose a hybrid quantization system in which the analog mapping, the quantization rule, and the digital processing, are learned from training in

an end-to-end fashion. The operation of the scalar ADCs is modeled as an intermediate activation layer. Unlike previous works which combined fixed uniform quantizers as part of a neural network [25,26,32], our method is specifically designed for learning scalar quantization mappings. We consider two generic tasks: estimating a set of parameters taking values in a continuous set from the quantized observations, and classifying the acquired signals. Our main target application is bit constrained MIMO receivers, in which these tasks may represent, for example, channel estimation and symbol detection, respectively.

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 straight-forward manner. To overcome this difficulty, previous works used a simplified model of the quantizer, in which the quantization error is replaced by additive i.i.d. noise [25,26,32]. As the quantization error is a deterministic function of the analog input [36], the resulting model is relatively inaccurate, inducing a mismatch which, as we numerically demonstrate, degrades the ability to properly optimize the system in light of the task. Furthermore, this model is limited to fixed uniform continuous-to-discrete mappings, namely the quantization mapping cannot be learned during training. Here, we follow the soft-to-hard approach used in [23], approximating the continuous-to-discrete mapping during training with a differentiable one which faithfully represents the operation of the quantizer. This approach facilitates the application of back-propagation, while allowing learning the quantization mapping as part of an end-to-end network.

We numerically evaluate the performance of our proposed DNN-based system in MIMO communication scenarios. We first consider channel estimation, and compare our data-driven task-based quantizer to previous channel estimators from task-ignorant quantized measurements, as well as to the model-based task-based quantization system 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 indirect rate-distortion theory, which are achievable using optimal vector quantizers ([11], 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.

Next, we test the data-driven quantizer for the task of symbol detection in multi-user MIMO communications. Here, we show that our quantizer achieves performance which is comparable to applying the maximum a-posteriori probability (MAP) rule without any quantization constraints, and is notably more robust to inaccurate channel state information (CSI). Furthermore, our deep task-based quantizer significantly outperforms the previously used approach of modeling quantization as additive noise during training, and we illustrate that the gap stems from the usage of a more accurate model for the quantization mapping. We also discuss how the proposed approach can be exploited to construct trainable task-based ADCs, by combining neuromorphic electronic systems [37] with digital neural networks, giving rise to robust, efficient, and accurate, data-driven methods for acquisition of analog signals.

The rest of this paper is organized as follows: Section 2 formulates the problem. Implementation of the data-driven task-based quantizer is presented in Section 3. Section 4 numerically evaluates the proposed quantizer in MIMO communication scenarios. Finally, Section 5 provides some concluding remarks.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$, and boldface upper-case letters for matrices, e.g., $\boldsymbol{M}$. Sets are denoted with calligraphic letters, e.g., $\mathcal{X}$. We use $\boldsymbol{I}_n$ to represent the $n \times n$ identity matrix. 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.

2. Preliminaries and Problem Statement

2.1. 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 $q_i \in \hat{\mathcal{X}}^k$.*

We write the output of the quantizer with input $x \in \mathcal{X}^n$ as $\hat{x} = g_k(f_n(x)) \triangleq Q_M^{n,k}(x)$. *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 size and the output size are equal, $n = k$, 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 $d : \mathcal{X}^n \times \hat{\mathcal{X}}^n \mapsto \mathcal{R}^+$ between its input and its output. The performance of a quantizer is characterized using two measures: the quantization rate, defined as $R \triangleq \frac{1}{n} \log M$, and the expected distortion $\mathbb{E}\{d(x, \hat{x})\}$. For a fixed input size n and codebook size M, the optimal quantizer is

$$Q_M^{n,\text{opt}}(\cdot) = \underset{Q_M^n(\cdot)}{\arg\min} \, \mathbb{E}\{d(x, Q_M^n(x))\}. \tag{1}$$

Characterizing the optimal quantizer via (1) and its 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., [38], or asymptotically large inputs, namely $n \to \infty$, commonly 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 random vector $s \in \mathcal{S}^k \subseteq \mathcal{R}^k$ from a statistically dependent random vector $x \in \mathcal{R}^n$. The set $\mathcal{S}$ represents the possible values of the unknown vector: It can be continuous, representing an estimation task; discrete, for classification tasks; or binary, for detection tasks. This formulation accommodates a broad range of applications, including channel estimation and symbol detection, that are the common tasks considered in bit-constrained hybrid MIMO communications receivers [9], which are the main target systems considered in this work.

When quantizing for the task of estimation, under the objective of minimizing the mean-squared error (MSE) distortion, i.e., $d(s, \hat{s}) = \|s - \hat{s}\|^2$, it was shown in [39] that the optimal quantizer applies vector quantization to the minimum MSE (MMSE) estimate of the desired vector s from the observed vector x. While the optimal system uses vector quantization, the fact that such pre-quantization processing can improve the performance in estimation tasks was also demonstrated in [5], which considered scalar quantizers. However, it was also shown in [5,6] that the pre-quantization processing which is optimal with vector quantizers, i.e., recovery of the MMSE estimate of s from x, is no longer optimal when using scalar quantization, and that characterizing the optimal pre-quantization processing in such cases is very difficult in general. The fact that processing the observations in the analog domain is beneficial in task-based quantization motivates the hybrid system model which is the focus of the current work, and detailed in the following subsection. Due to the difficulty in analytically characterizing the optimal hybrid system, we consider a data-driven design, described in Section 3.

2.2. Problem Statement

As discussed in the introduction, practical digital signal processing systems typically obtain a digital representation of physical analog signals using 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 scalar ADCs in a data-driven fashion.

The considered signal acquisition system with scalar ADCs is modeled using the hybrid setup depicted in Figure 1, where a set of analog signals are converted to digital in order to extract some desired information from them. This model can represent, e.g., sensor arrays or MIMO receivers, and specializes the case of a single analog input signal. While acquiring a set of analog signals in digital hardware includes both sampling, i.e., continuous-to-discrete time conversion, as well as quantization, namely continuous-to-discrete amplitude mapping, we henceforth focus only the quantization aspect assuming a fixed sampling mechanism, and leave the data-driven design of the overall system for future investigation.
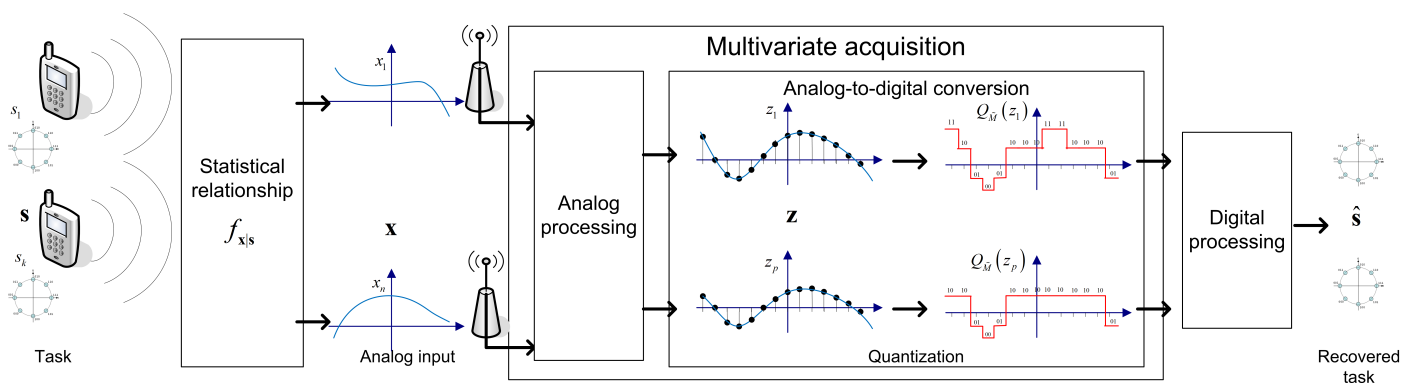
Figure 1. Hybrid task-based quantization system model. For illustration, the task is recovering a set of constellation symbols in uplink MIMO communications.

We consider the recovery of an unknown random vector $s \in \mathcal{S}^k$ based on an observed vector $x \in \mathcal{R}^n$ quantized with up to $\log M$ bits. The observed x is related to s via a conditional probability measure $f_{x|s}$, which is assumed to be unknown. For example, in a communications setup. the conditional probability measure $f_{x|s}$ encapsulates the noisy channel. The input to the ADC, denoted $z \in \mathcal{R}^p$, where p denotes the number of scalar quantizers, is obtained from x using some pre-quantization mapping carried out in the analog domain. Then, 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 $\hat{s} \in \mathcal{S}^k$.

Our goal is to design a generic machine-learning based architecture for task-based quantization with scalar ADCs. The proposed system operates in a data-driven manner, namely it is capable of learning the analog transformation, quantization mapping, and digital processing, from a training data set, consisting of t independent realizations of s and x, denoted $\{s^{(i)}, x^{(i)}\}_{i=1}^{t}$. In general, the training samples may be taken from a set of joint distributions, and not only from the true (unknown) joint distribution of s and x, as we consider in our numerical study in Section 4. We focus on two tasks which are relevant for MIMO receivers: An estimation task, in which $\mathcal{S} = \mathcal{R}$, representing, e.g., channel estimation; and classification, where $\mathcal{S}$ is a finite set, modeling, e.g., symbol detection. Our design is based on machine-learning methods, and specifically, on the application of DNNs.

3. Deep Task-Based Quantization

In the following, we present a deep task-based quantizer, which implements the system depicted in Figure 1 in a data-driven fashion using DNNs. To that aim, we first discuss the proposed network architecture in Section 3.1. Then, in Section 3.2 we elaborate on the discrete-to-continuous mapping and its training method, and provide a discussion on the resulting system in Section 3.3.

3.1. DNN Architecture

We propose to implement a data-driven task-based quantizer using machine-learning methods. In particular, we realize the pre and post quantization mappings using dedicated DNNs, jointly trained in an end-to-end manner, as illustrated in Figure 2.
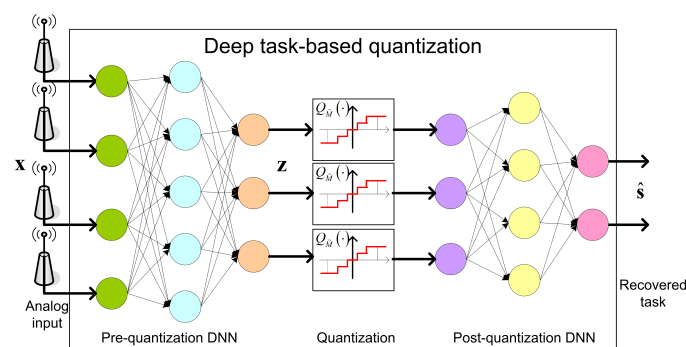

Figure 2. Deep task-based quantization system architecture.

In the proposed architecture, the serial scalar ADC, which implements the continuous-to-discrete mapping, is modeled as an activation function between the two intermediate layers. The trainable parameters of this activation function determine the quantization rule, allowing it to be learned during training. The DNN structure cannot contain any skip connections between the multiple layers prior to quantization (analog domain) and those after quantization (digital domain), representing the fact that all analog values must be first quantized before processed in digital. The pre and post quantization networks are henceforth referred to as the *analog DNN* and the *digital DNN*, respectively. The system input is the $n \times 1$ observed vector $\boldsymbol{x}$, and we use $\boldsymbol{\theta}$ to denote the trainable parameters of the network. As detailed in Section 2.2, we consider two main types of tasks:

- Estimation: Here, the deep task-based quantizer should learn to recover a set of k unknown parameters taking values on a continuous set, i.e., $\mathcal{S} = \mathcal{R}$. By letting $\psi_{\boldsymbol{\theta}}(\cdot)$ denote the mapping implemented by the overall system, the output is given by the $k \times 1$ vector $\hat{\boldsymbol{s}} = \psi_{\boldsymbol{\theta}}(\boldsymbol{x})$, which is used as a representation of the desired vector $\boldsymbol{s}$. The loss function is the empirical MSE, given by

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

- Classification: In such tasks, the deep task-based quantization should decide between a finite number of options based on its analog input. Here, $\mathcal{S}$ is a finite set, and we use $|\mathcal{S}|$ to denote its cardinality. The last layer of the digital DNN is a softmax layer, and thus the network mapping $\psi_{\boldsymbol{\theta}}(\cdot)$ is a $|\mathcal{S}|^k \times 1$ vector, whose entries represent the conditional probability for each different value of $\boldsymbol{s}$ given the input $\boldsymbol{x}$. By letting $\psi_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\alpha})$ be the output value corresponding to $\boldsymbol{\alpha} \in \mathcal{S}^k$, the decision is selected as the most probable one, i.e., $\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{\alpha} \in \mathcal{S}^k} \psi_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\alpha})$. The loss function is the empirical cross-entropy, given by

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{t} \sum_{j=1}^{t} -\log \psi_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(j)}; \boldsymbol{s}^{(j)}\right). \tag{3}$$

By using DNNs, we expect the resulting system to be able to approach the optimal achievable distortion for fixed quantization rate $R = \frac{1}{n} \log M$ and input size n, without requiring explicit knowledge of the underlying distribution $f_{x|s}$. Such performance is illustrated in the numerical example presented in Section 4.1.

The proposed architecture is generic, and its main novelty is in the introduction of the learned quantization layer, detailed in the following subsection. Our structure can thus be combined with existing dedicated networks, which are trainable in an end-to-end manner, as a form of transfer learning. For example, sliding bidirectional recursive neural networks (SBRNNs) were shown to achieve good performance for the task of symbol detection in non-quantized communication systems with long memory [40]. Consequently, one can design a deep symbol detector operating under quantization constraints, as common in, e.g., millimeter wave communications [13], by implementing the digital DNN of Figure 2 as an SBRNN. In this work we focus on fully connected analog and digital DNNs, and leave the analysis of combination with dedicated networks to future investigation.

3.2. Quantization Activation

Our proposed deep task-based quantizer implements scalar quantization as an intermediate activation in a joint analog-digital hybrid DNN. This layer converts its continuous-amplitude input into a discrete digital representation. The non-differentiable nature of such continuous-to-discrete mappings induces a major challenge in applying SGD for optimizing the trainable parameters of the network. In particular, quantization activation, which can be modeled as a superposition of step functions determining the continuous regions jointly mapped into a single value, 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 first review the common approach, referred to henceforth as *passing gradient*. Then discuss how one can backpropagate via quantization mappings while faithfully accounting for its operation as well as enabling its optimization in the training procedure via *soft-to-hard quantization*.

3.2.1. Passing Gradient

In this approach the quantized values are modeled as the analog values corrupted by mutually independent i.i.d. noise [25,26,32], and thus quantization does not affect the back-propagation procedure. Since the quantization error is deterministically determined by the analog value [36], the resulting model is quite inaccurate. Specifically, while under some input distributions, the quantization noise can be modeled as being *uncorrelated* with the input [36], they are not mutually independent. In fact, in order for the quantization error to be independent of the input, one should use substractive dithered quantization [41], which does not represent the operation of practical ADCs. Consequently, using this model for quantization during training results in a mismatch between the trained system and the tested one.

Under this model, the continuous-to-discrete mapping is fixed, representing, e.g., uniform quantization, and the training algorithm back-propagates the gradient value intact through the quantization layer. An illustration of this approach is depicted in Figure 3a. We expect the resulting system to obtain poor performance when non-negligible distortion is induced by the quantizers. In our numerical study presented in Section 4.2, it is illustrated that this method achieves relatively poor performance at low quantization rates, where scalar quantization induces an error term which is non-negligible and depends on the analog input. It is, therefore, desirable to formulate a network structure which accounts for the presence of scalar quantizers during training, and is not restricted to fixed uniform quantizers.

3.2.2. Soft-to-Hard Quantization

An alternative approach is based on approximating the non-differentiable quantization mapping by a differentiable one. Here, we replace the continuous-to-discrete transformation with a non-linear activation function which has approximately the same behavior as

the quantizer, as illustrated in Figure 3b. Specifically, we use a sum of shifted hyperbolic tangents, which are known to closely resemble step functions in the presence of large magnitude inputs. The resulting scalar quantization mapping is given by:

$$\tilde{q}_{\tilde{M}}(x) = \sum_{i=1}^{\tilde{M}-1} a_i \tanh(c_i \cdot x - b_i), \tag{4}$$

where $\{a_i, b_i, c_i\}$ are a set of real-valued parameters. Please 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 [23], we refer to this method as *soft-to-hard quantization*.
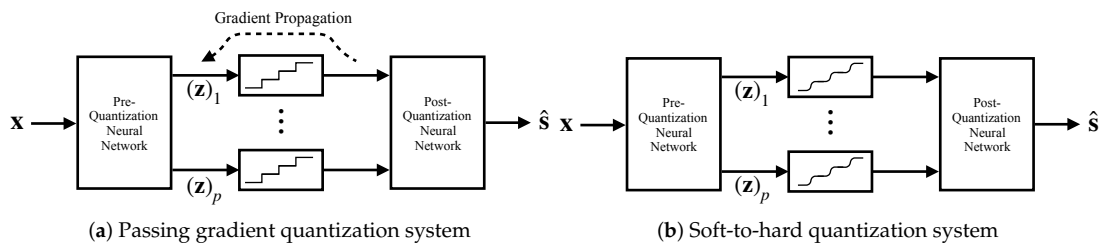

**Figure 3.** Task-based deep quantization architectures.

In addition to learning the weights of the analog and digital DNNs, this soft-to-hard approach allows the network to learn its quantization 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 scalar quantizer, resulting in a learned quantization mapping. The parameters $\{c_i\}$, which essentially control the resemblance of (4) to an actual continuous-to-discrete mapping, do not reflect on the quantization decision regions (controlled by $\{b_i\}$) and their associated digital values (determined by $\{a_i\}$), and are thus not learned from training. The set $\{c_i\}$ can be either set according to the quantization resolution $\tilde{M}$, or alternatively, modified using annealing-based optimization [42], where $\{c_i\}$ are manually increased during training. The proposed optimization is achieved by including the parameters $\{a_i, b_i\}$ as part of the network trainable parameters $\boldsymbol{\theta}$. Due to the differentiability of (4), one can now apply standard SGD to optimize the overall network, including the analog and digital DNNs as well as the quantization rule, in an end-to-end manner.

Once training is concluded, we replace the learned $\tilde{q}_{\tilde{M}}(\cdot)$ activation (4) 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 regions of the quantizer, and set 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 $\frac{b_0}{c_0} \leq \frac{b_1}{c_1} \leq \ldots \leq \frac{b_{\tilde{M}-1}}{c_{\tilde{M}-1}}$ (when this conditions is not satisfied, the parameters are sorted and re-indexed accordingly). The resulting quantizer is given by

$$Q_{\tilde{M}}^1(x) = \begin{cases} -\sum_{i=1}^{\tilde{M}-1} a_i & x \leq \frac{b_0}{c_0} \\ \tilde{q}_{\tilde{M}}\left(\frac{b_i}{2c_i} + \frac{b_{i+1}}{2c_{i+1}}\right) & \frac{b_i}{c_i} < x \leq \frac{b_{i+1}}{c_{i+1}} \\ \sum_{i=1}^{\tilde{M}-1} a_i & \frac{b_{\tilde{M}-1}}{c_{\tilde{M}-1}} < x. \end{cases} \tag{5}$$

An illustration of how the differentiable mapping (4) is converted into a continuous-to-discrete quantization rule via (5) is depicted in Figure 4. The dashed smooth curve in Figure 4 represents the differentiable function after training is concluded, and the straight curve is the resulting scalar quantizer.
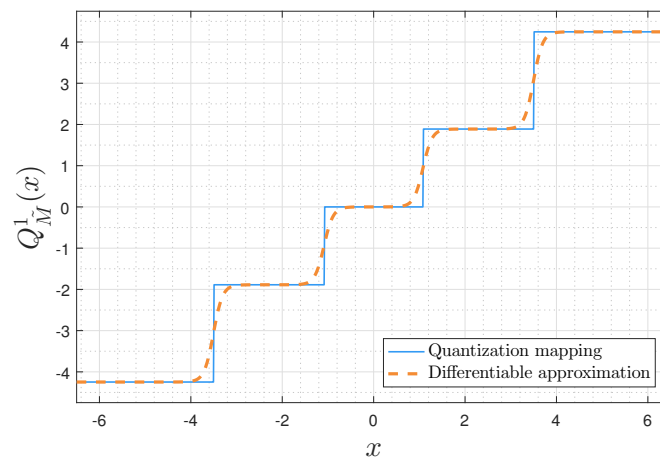
Figure 4. Soft-to-hard quantization rule illustration.

In the simulations study presented in Section 4.1, it is illustrated that the proposed method, which faithfully represents the presence of scalar quantizers during training and is capable of optimizing their decision regions, can outperform the model-based MSE minimizing task-based quantizer with scalar ADCs of [5], which requires complete knowledge of the underlying model, yet is restricted to uniform quantizers.

3.3. Discussion

The deep task-based quantizer proposed in Section 3.1 enables implementing quantization systems which learn how to map a continuous-amplitude input into a finite-bit representation in a manner which account for the task for which the signal is acquired. In the following we discuss some possible following directions which can facilitate transforming the proposed concept into DNN-aided task-based acquisition devices.

The system proposed in Section 3.1 implements hybrid multivariate acquisition using a set of identical scalar ADCs with learned decision regions, combined with DNN-based analog and digital transformations. While realizing DNNs in digital can be done in software, analog DNNs requires dedicated tunable hardware weights and activations. Such hardware networks, commonly referred to as neuromorphic electronic systems [37], implement configurable DNNs as analog components. Recent advances in memristors technology substantially facilitate the implementation of these hardware devices [43], contributing to the feasibility of our proposed deep task-based quantizer.

It is noted that in some applications, constrained analog structures may be preferable. For example, in MIMO receivers with a large number of antennas, i.e., massive MIMO, pre-quantization analog processing is commonly limited to phase shifting [21] or linear mappings [44]. Alternative hybrid acquisitions architectures for RF analog signals which include some level of controllable analog processing are dynamic metasurface antennas [45–47], which provide tunable analog combining in the form of Lorentzian filters. As such hybrid systems offer a limited range of analog mappings, one may prefer to use dedicated neural network circuitry for implementing deep task-based quantizers. In this case, the analog DNN is replaced with a single layer whose weights are restricted to have a unit magnitude, and this constraint has to be accounted for in training. Here we focus on generic analog DNNs, in which the weights are not constrained.

Our task-based quantizer can thus be implemented as a system consisting of adjustable analog hardware, configurable scalar quantizers, and software. The natural approach to set the parameters of the network would be to train the system model offline in software using an a priori acquired training set. The network weights and quantization decision regions obtained from this trained model can be then configured into the hardware components and the tunable ADCs, resulting in the desired task-based quantization system.

One can also envision an online trainable task-based quantizer, which is capable of further tuning its trainable parameters in real time to track dynamic environments, as in, e.g., [48–50]. For example, a communication receiver using a deep task-based quantizer for symbol detection, can exploit a priori knowledge of pilot sequences as labels corresponding to inputs acquired in real time. A major challenge in implementing such a system stems from the fact that both the labels $\{s^{(j)}\}$ as well as the inputs $\{x^{(j)}\}$ are required in order to update the network coefficients using conventional training algorithms, e.g., SGD. However, in our system the digital processor does not have direct access to the analog signal, but only to its quantized digital representation. Consequently, if the processor only uses digital values, it can only train the digital DNN using SGD. This challenge may be handled by allowing access to a high resolution quantized version of the analog signals, acquired in the specific time instances for which labels are available. An alternative approach is to use an error-correction-based update algorithm [51] instead of SGD, or reinforcement learning methods [52], since these techniques typically do not require direct access to the network input.

4. Application to MIMO Receivers

While the generic deep task-based quantizer proposed in Section 3 is applicable to a multitude of different setups, our main target application, studied in this section, is uplink multi-user MIMO communications. The problem of MIMO communications with low-resolution quantization is the focus of many recent works, including, e.g., [9,15,16,22,34]. Here, we consider a single cell multi-user MIMO system, in which n_u single antenna users are served by a base station (BS) with n_t antennas, which operates under quantization constraints. We focus on two tasks encountered in such setups: The first is channel estimation detailed in Section 4.1, for which we are capable of quantifying the performance gap of our system from optimality as well as comparing it to model-based designs. Then, in Section 4.2 we focus on symbol detection, which we treat as a classification task.

4.1. Channel Estimation Task

We first consider channel recovery, which is an estimation task commonly encountered in MIMO systems. We focus on a specific scenario for which we can compute both the fundamental performance limits, namely a lower bound on the achievable recovery accuracy which holds for any bit constrained system, as well as the performance of the best hybrid system restricted to using linear operations and uniform quantization, derived in [5]. These performance measures, which correspond to model-based systems, are used as a basis for comparison to evaluate our proposed data-driven task-based quantizer. The main motivation for the study detailed in this subsection is thus to compare the performance achievable using our proposed deep task-based quantizer to model-based techniques and the fundamental performance limits in a specific scenario where these values are computable.

In the following, we consider a channel estimation task carried out in a time diversity duplexing manner as in [9], using orthogonal pilot sequences of length $\tau_p \geq n_u$. We use $\boldsymbol{\Phi} \in \mathcal{C}^{\tau_p \times n_u}$ to denote the known pilot sequence matrix, where the orthogonality of the pilots implies that $\boldsymbol{\Phi}^H \boldsymbol{\Phi} = \tau_p \cdot \boldsymbol{I}_{n_u}$, and P is the SNR. Additionally, let $\boldsymbol{h} \in \mathcal{C}^{n_u \cdot n_t}$ be a random vector whose entires are 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], Equation (4)):

$$\boldsymbol{y} = \sqrt{P}(\boldsymbol{\Phi} \otimes \boldsymbol{I}_{n_t})\boldsymbol{h} + \boldsymbol{w}, \tag{6}$$

where $\otimes$ is the Kronecker product.

To put the setup in (6) in the framework of our problem formulation, which considers real-valued signals, we write the observations as $\boldsymbol{x} = \left[\text{Re}(\boldsymbol{y})^T, \text{Im}(\boldsymbol{y})^T\right]^T$ and the

unknown channel as $s = \left[\text{Re}(h)^T, \text{Im}(h)^T\right]^T$. 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 $\rho = \frac{\tau_p}{n_u}$, which is not smaller than one.

The performance measure for evaluating the quantization systems here is the average MSE, namely $\eta = \frac{1}{k}\mathbb{E}\left\{\|s - \hat{s}\|^2\right\}$. For the above model, 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 [39], and the resulting average distortion is obtained from rate-distortion theory ([3], Ch. 10.3) as

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

Please 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*.

We now numerically evaluate our proposed deep task-based quantizer, 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}. 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, namely the SNR as well as the distribution of h and w.

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 $\boldsymbol{\Phi}$ to be the first n_u columns of the $\tau_p \times \tau_p$ discrete Fourier transform matrix. In the implementation of 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. Furthermore, this setting guarantees fair comparison with the model-based system of [5], which focused on linear analog and digital processing. 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 $\{s^{(i)}, x^{(i)}\}_{i=1}^{t}$ are sampled from the true joint distribution of s, x; In the second training set, representing *SNR uncertainty*, $\{s^{(i)}, x^{(i)}\}_{i=1}^{t}$ are sampled from the joint distribution of s, x with different values of P, uniformly randomized over the set $[1, 10]$ for each realization. At the end of the training session, we fix the quantizer to implement the continuous-to-discrete rule in (5). We numerically evaluate our trained proposed deep quantizer using 2^{10} independent channel realizations.

In Figure 5 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 three theoretical measures: the MMSE $\tilde{\eta}$; the fundamental performance limits of channel estimation from quantized measurements, given by η_{opt} in (7); and the performance of the analytically derived task-based quantizer with scalar ADCs [5], denoted η_{sc}. Since [5] requires perfect knowledge of the underlying model, and particularly of the SNR P, and as this information may not be available accurately in practice, we also consider the case where [5] uses an estimation of P corrupted by zero-mean Gaussian noise with variance $\frac{P}{4}$. Finally, we compute the average MSE of the BLMMSE estimator proposed in [15] via ([15], Equation (15)). Since the BLMMSE estimator quantizes the observed signal without analog pre-processing, it is applicable only for $R \geq 1$.

Observing Figure 5, we note that the performance of our soft-to-hard deep quantizer is within a relatively 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 the model-based 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 the ability of deep task-based quantization to implement a feasible and optimal-approaching quantization system in a data-driven fashion.
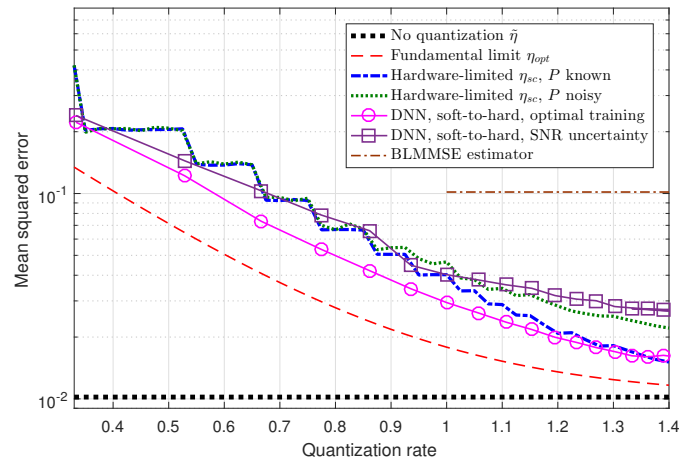

Figure 5. Numerical MSE versus theoretical measures.

4.2. Symbol Detection Task

The main task of a communication receiver is to recover the transmitted messages. Channel estimation, studied in the previous subsection, is intended to facilitate the recovery of the unknown symbols. Consequently, we next consider the task of symbol recovery, in which the receiver learns to recover a set of constellation points from its quantized channel output.

As shown in the previous subsection, multivariate complex-values (baseband) can be represented as real vector channels of extended dimensions. Therefore, here we focus on communications over a real-valued MIMO channel. In particular, we consider a BS equipped with $n_t = 12$ antennas, serving $n_u = 4$ users. The users transmit i.i.d. binary phase shift keying (BPSK) symbols, represented via the vector $s \in \{-1, 1\}^{n_u}$. The received signal at the BS, denoted $x \in \mathcal{R}^{n_t}$, is given by

$$x = Hs + w, \tag{8}$$

where $H \in \mathcal{R}^{n_t \times n_u}$ is the channel matrix with entries $(H)_{i,j} = e^{-|i-j|}$, representing spatial exponential decay, and $w \in \mathcal{R}^{n_t}$ is additive Gaussian noise with zero-mean i.i.d. entries of variance $\sigma_w^2 > 0$.

Here, the task of the BS is to recover the transmitted symbols vector s from the channel output x, i.e., in this scenario the input dimension is $n = n_t$ and the task dimension is $k = n_u$. We use a DNN architecture consisting of two fully connected layers in analog and two fully connected layers in digital. As this is a classification task, the output layer is a softmax function with 2^k probabilities, and the overall network is trained to minimize the cross-entropy loss. An illustration of the DNN structure is depicted in Figure 6. Unlike the scenario considered in the previous subsection, for which the number of quantizers p can be set according to the analytical results in [5], here this value was determined based on empirical evaluations. In particular, we use $p = \lfloor kR \rfloor$, resulting in each scalar quantizer using at least $n/k = 3$ bits in the hybrid system.
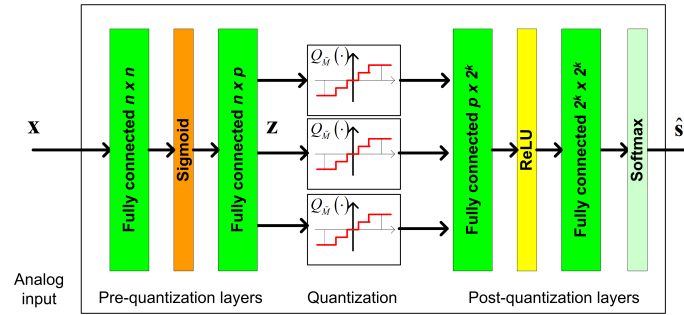
Figure 6. Task-based symbol detector DNN.

We compare the achievable bit error rate (BER) of our proposed deep task-based quantizer with soft-to-hard training to using the same architecture with passing gradient training, namely where the quantizers are replaced with additive i.i.d. noise uniformly distributed over the decision regions during training, which is the approach used to train neural networks with intermediate quantization in [25,26,32]. In particular, for the passing gradient method we used a uniform quantization rule over the support $[-2, 2]$. The DNNs are trained using a relatively small training set consisting of $t = 5000$ realizations sampled from the joint distribution of s, x.

The aforementioned data-driven systems are compared to two model-based symbol detectors, which require accurate CSI, i.e., knowledge of H or σ_w^2 from which the joint distribution of s and x can be inferred using (8):

1. The MAP rule for recovering s from x *without quantization constraints*, i.e.,

$$\hat{s}_{\text{MAP}} \triangleq \underset{s' \in \{-1,1\}^k}{\arg\max} \ \Pr(s = s'|x). \tag{9}$$

    The performance of the MAP detector with perfect CSI constitutes a lower bound on the achievable BER of any recovery scheme.

2. The MAP rule for recovering s from a uniformly quantized x with rate R, namely

$$\hat{s}_{\text{QMAP}} \triangleq \underset{s' \in \{-1,1\}^k}{\arg\max} \ \Pr\left(s = s'|Q_{\lfloor 2^R \rfloor}(x)\right), \tag{10}$$

    where $Q_{M'}(\cdot)$ represents the element-wise uniform quantization rule over the interval $[-2, 2]$ using M' decision regions. The performance of the quantized MAP detector represents the achievable BER when processing is carried out solely in the digital domain, i.e., without using analog processing and/or tunning the quantization mapping in light of the task.

Unlike the detectors based on the MAP rule in (9) and (10), data-driven task-based quantizers do not require CSI, namely no a priori knowledge of H or σ_w^2 is used in the detection procedure. Instead, a set of training samples are needed. In order to study the resiliency of our deep task-based quantizer to inaccurate training, we also compute the BER under CSI uncertainty, namely when the training samples are randomized from a joint distribution of s, x in which the entries of the matrix H in (8) are corrupted by additive i.i.d. Gaussian noise, whose variance is 20% the magnitude of the corresponding entry. For comparison, we also evaluate the BER of the MAP rule (9) with the same level of CSI uncertainty. The numerically computed BER values are averaged over 20,000 Monte Carlo simulations.

The simulated BER values versus SNR, defined here as $1/\sigma_w^2$, in the range of $[6, 14]$ dB, are depicted in Figures 7 and 8 for quantization rates $R = 1$ and $R = 2$, respectively. Observing Figures 7 and 8, we note that in the presence of accurate CSI, the BER of our deep task-based quantizer is comparable to that achievable using the MAP rule operating without quantization constraints. In particular, while the MAP detector, which is indepen-

dent of the quantization rate, achieves BER of 10^{-3} at SNR of 10 dB, the deep task-based quantizer obtains such BER values at SNRs of 13 and 12 dB, respectively, for quantization rates $R = 1$ and $R = 2$, respectively, namely SNR gaps of 3 and 2 dB. For comparison, the quantized MAP rule, which operates only in the digital domain, does not achieve BER values below 10^{-2} at $R = 1$ and requires SNR of over 13 dB to achieve BER of 10^{-3} at rate $R = 2$, i.e., with twice the number of bits used by the deep task-based quantizer to achieve the same error rate. This demonstrates the benefit of applying pre-quantization processing in the analog domain, which reduces the dimensionality of the input to the scalar quantizers, thus allowing using more accurate quantization while keeping the semantic information required to classify the symbols from the channel output.

The performance gain of the hybrid DNN architecture stems from the ability to properly model the scalar quantizers during training using our soft-to-hard approach. This model allows to jointly train both the analog and digital DNNs as well as the decision regions of the quantizers, while accurately reflecting the quantization mapping. For comparison, it is observed in Figures 7 and 8 that using the passing gradient approach, i.e., replacing quantization with additive uniformly distributed i.i.d. noise as was done in [25,26,32], leads to substantially deteriorated BER values compared to the proposed soft-to-hard approach. To understand whether the improved gains of soft-to-hard modeling over passing gradient stems from the better approximation of the continuous-to-discrete mapping or from the ability to use non-uniform quantizers, we compare in Figure 9 the performance of the task-based quantizers with soft-to-hard modeling and with passing gradient modeling for the scenario of Figure 7 when using a fixed uniform quantizer with soft-to-hard modeling. In particular, for the uniform soft-to-hard quantizer we used the model in (4) during training with the parameters $\{a_i, b_i\}$ being fixed to uniform partition of the interval $[-2, 2]$, i.e., not optimized during training. It is clearly observed in Figure 9 that most of the gain follows from the usage of an accurate differentiable approximation of the continuous-to-discrete quantization mapping, which allows to train the system in an end-to-end manner while faithfully representing quantization. The gains due to optimizing the decision regions are rather small, indicating that our proposed approach can also lead to substantial improvements when restricted to using uniform scalar quantizers.
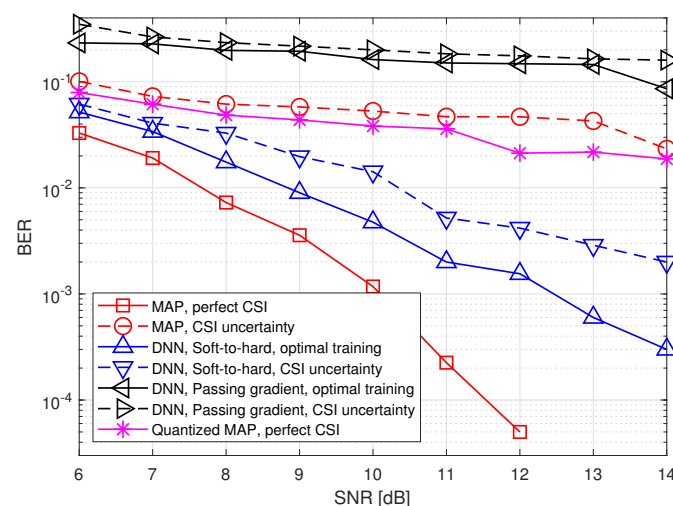

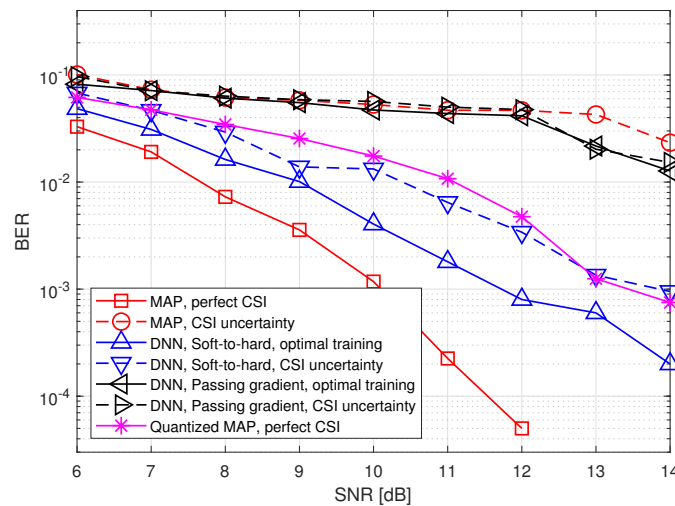**Figure 7.** BER versus SNR at rate $R = 1$.

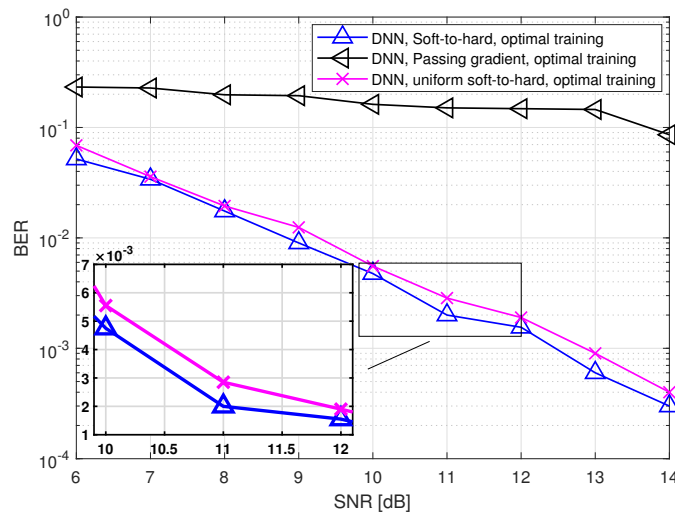**Figure 8.** BER versus SNR at rate $R = 2$.

Figure 9. BER versus SNR at rate $R = 1$.

The results in Figures 7 and 8 also demonstrate the improved robustness to inaccurate CSI. The performance of the model-based MAP detector is very sensitive to CSI uncertainty, resulting in a notable increase in BER due to the model mismatch. However, the performance of the deep task-based quantizer trained under CSI uncertainty is within an SNR gap of approximately 0.5–2 dB from its achievable performance when trained using accurate CSI. Furthermore, the deep task-based quantizer with CSI uncertainty substantially outperforms the MAP rule without quantization constraints with the same level of uncertainty for all considered scenarios, and outperforms the quantized MAP with accurate CSI for quantization rate of $R = 1$. This demonstrates the gains of using DNNs, with their established generalization properties, for overcoming the sensitivity of model-based approaches to inaccurate knowledge of the underlying parameters.

Next, we evaluate the BER of the considered quantization systems versus the quantization rate $R \in [1, 3]$. The results are depicted in Figures 10 and 11 for SNR values of 8 dB and 12 dB, respectively. Observing Figures 10 and 11, we note that the gain of the proposed deep task-based quantizer is more dominant when operating with low quantization rates. As the quantization rate approaches three bits per channel input, the BER of applying the MAP in the digital domain via (10) is only within a small gap of the hybrid quantizer with soft-to-hard training. However, for lower quantization rates, as well as in the presence of CSI uncertainty, the proposed deep task-based quantizer maintains its superiority observed in Figures 7 and 8. Furthermore, it is noted that when using the passing gradient training

approach, there is a very small gap between the performance achievable with and without CSI uncertainty. This observation is likely due to the fact that when modeling quantization as additive independent noise during training, the network is trained on a mismatched model, regardless of whether the training samples are taken from the same distribution as the test samples. Consequently, such data-driven quantizers operate under some level of uncertainty even when trained using an optimal training set.
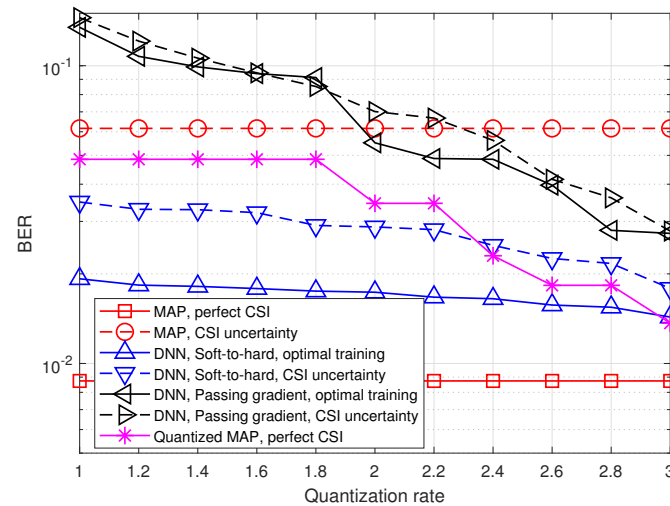

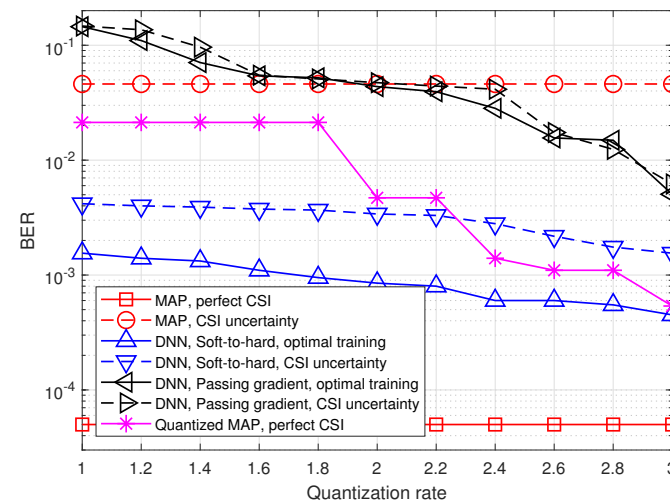**Figure 10.** BER versus quantization rate at 8 dB SNR.

Figure 11. BER versus quantization rate at 12 dB SNR.

Finally, we note that the DNNs used in this subsection were trained using a relatively small training set, consisting of $t = 5000$ samples. This indicates that such architectures can be used to realize an online trainable dynamic ADC, as discussed in Section 3.3.

5. Conclusions

In this work we designed a data-driven task-based quantization system, operating with scalar ADCs, using DNNs. We proposed a method for handling the non-differentiability of quantization by approximating its mapping as a smooth function. Our proposed model faithfully represents such continuous-to-discrete mappings while allowing to learn the quantization rule from training. We discussed how this strategy can be used for designing dynamic machine-learning-based ADCs for various tasks. Our numerical results, which considered channel estimation and symbol recovery in bit-constrained MIMO systems, demonstrate that the performance achievable with the proposed deep task-based quantizer is

comparable with the fundamental limits for this setup, achievable using optimal vector quantizers. Furthermore, we showed that by using a soft-to-hard approximation of the quantization procedure when training the network in an end-to-end fashion allows the system to be accurately trained with a relatively small training set, and that it notably outperforms the common approach for training DNNs with intermediate quantization.

References

1. Gray, R.M.; Neuhoff, D.L. Quantization. *IEEE Trans. Inform. Theory* **1998**, *44*, 2325–2383. [CrossRef]
2. Berger, T.; Gibson, J.D. Lossy source coding. *IEEE Trans. Inform. Theory* **1998**, *44*, 2693–2723. [CrossRef]
3. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley Press: Hoboken, NJ, USA, 2006.
4. Rodrigues, M.R.D.; Deligiannis, N.; Lai, L.; Eldar, Y.C. Rate-distortion trade-offs in acquisition of signal parameters. In Proceedings of the IEEE ICASSP, New-Orleans, LA, USA, 5–9 March 2017; pp. 6105–6109.
5. Shlezinger, N.; Eldar, Y.C.; Rodrigues, M.R.D. Hardware-limited task-based quantization. *IEEE Trans. Signal Process.* **2019**, *67*, 5223–5238. [CrossRef]
6. Salamatian, S.; Shlezinger, N.; Eldar, Y.C.; Medard, M. Task-based quantization for recovering quadratic functions using principal inertia components. In Proceedings of the IEEE ISIT, Paris, France, 7–12 July 2019.
7. Neuhaus, P.; Shlezinger, N.; Dorpinghaus, M.; Eldar, Y.C.; Fettweis, G. Task-based analog-to-digital converters. *arXiv* **2020**, arXiv:2009.14088.
8. Shlezinger, N.; Eldar, Y.C. Task-based quantization with application to MIMO receivers. *arXiv* **2019**, arXiv:1908.06845.
9. Shlezinger, N.; Eldar, Y.C.; Rodrigues, M.R.D. Asymptotic task-based quantization with application to massive MIMO. *IEEE Trans. Signal Process.* **2019**, *67*, 3995–4012. [CrossRef]
10. Xi, F.; Shlezinger, N.; Eldar, Y.C. BiLiMO: Bit-limited MIMO radar via task-based quantization. *arXiv* **2020**, arXiv:2010.00195.
11. Polyanskiy, Y.; Wu, Y. *Lecture Notes on Information Theory*; MIT: Cambridge, MA, USA, 2015.
12. Walden, R.H. Analog-to-digital converter survey and analysis. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1909–1935. [CrossRef]
13. Xiao, M.; Mumtaz, S.; Huang, Y.; Dai, L.; Li, Y.; Matthaiou, M.; Karagiannidis, G.K.; Bjornson, E.; Yang, K.; Chih-Lin, I. Millimeter wave communications for future mobile networks. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1909–1935. [CrossRef]
14. Mo, J.; Schniter, P.; Heath, R.W. Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs. *IEEE Trans. Signal Process.* **2018**, *66*, 1141–1154. [CrossRef]
15. Li, Y.; Tao, C.; Seco-Granados, G.; Mezghani, A.; Swindlehurst, A.L.; Liu, L. Channel estimation and performance analysis of one-bit massive MIMO systems. *IEEE Trans. Signal Process.* **2017**, *65*, 4075–4089. [CrossRef]
16. Choi, J.; Mo, J.; Heath, R.W. Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs. *IEEE Trans. Commun.* **2016**, *64*, 2005–2018. [CrossRef]
17. Jacobsson, S.; Durisi, G.; Coldrey, M.; Gustavsson, U.; Studer, C. Throughput analysis of massive MIMO uplink with low-resolution ADCs. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 4038–4051. [CrossRef]
18. Pirzadeh, H.; Swindlehurst, A.L. Spectral efficiency of mixed-ADC massive MIMO. *IEEE Trans. Signal Process.* **2018**, *66*, 3599–3613. [CrossRef]
19. Mollen, C.; Choi, J.; Larsson, E.G.; Heath, R.W. Uplink performance of wideband massive MIMO with one-bit ADCs. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 87–100. [CrossRef]
20. Studer, C.; Durisi, G. Quantized massive MU-MIMO-OFDM uplink. *IEEE Trans. Commun.* **2016**, *64*, 2387–2399. [CrossRef]
21. Stein, S.; Eldar, Y.C. A family of hybrid analog digital beamforming methods for massive MIMO systems. *IEEE Trans. Signal Process.* **2019**, *67*, 3243–3257.
22. Mo, J.; Alkhateeb, A.; Abu-Surra, S.; Heath, R.W. Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate tradeoffs. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 2274–2287. [CrossRef]
23. Agustsson, E.; Mentzer, F.; Tschannen, M.; Cavigelli, L.; Timofte, R.; Benini, L.; van Gool, L. Soft-to-hard vector quantization for end-to-end learning compressible representations. In Proceedings of the NIPS, Long Beach, CA, USA, 4–9 December 2017; pp. 1141–1151.

24. Toderici, G.; Vincent, D.; Johnston, N.; Hwang, S.J.; Minnen, D.; Shor, J.; Covell, M. Full resolution image compression with recurrent neural networks. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017.
25. Balle, J.; Laparra, V.; Simoncelli, E.P. End-to-end optimized image compression. *arXiv* **2016**, arXiv:1611.01704.
26. Balle, J.; Minnen, D.; Singh, S.; Hwang, S.J.; Johnston, N. Variational image compression with a scale hyperprior. *arXiv* **2018**, arXiv:1802.01436.
27. Johnston, N.; Vincent, D.; Minnen, D.; Covell, M.; Singh, S.; Chinen, T.; Hwang, S.J.; Shor, J.; Toderici, G. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. *arXiv* **2017**, arXiv:1703.10114.
28. Torfason, R.; Mentzer, F.; Agustsson, E.; Tschannen, M.; Timofte, R.; van Gool, L. Towards image understanding from deep compression without decoding. *arXiv* **2018**, arXiv:1803.06131.
29. Han, S.; Mao, H.; Dally, W.J. Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv* **2015**, arXiv:1510.00149.
30. Ullrich, K.; Meeds, E.; Welling, M. Soft weight-sharing for neural network compression. *arXiv* **2017**, arXiv:1702.04008.
31. Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Quantized neural networks: Training neural networks with low precision weights and activationss. *J. Mach. Learn. Res.* **2018**, *187*, 1–30.
32. Yang, Q.; Mashhadi, M.B.; Gunduz, D. Deep convolutional compression for massive MIMO CSI feedback. *arXiv* **2019**, arXiv:1907.02942.
33. Khobahi, S.; Naimipour, N.; Soltanalian, M.; Eldar, Y.C. Deep signal recovery with one-bit quantization. *arXiv* **2018**, arXiv:1812.00797.
34. Balevi, E.; Andrews, J.G. One-bit OFDM receivers via deep learning. *IEEE Trans. Commun.* **2019**, *67*, 4326–4336. [CrossRef]
35. Choi, J.; Cho, Y.; Evans, B.L.; Gatherer, A. Robust learning-based ML detection for massive MIMO systems with one-bit quantized signals. *arXiv* **2018**, arXiv:1811.12645.
36. Widrow, B.; Kollar, I.; Liu, M.C. Statistical theory of quantization. *IEEE Trans. Instrum. Meas.* **1996**, *45*, 353–361. [CrossRef]
37. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **1990**, *78*, 1629–1636. [CrossRef]
38. Li, J.; Chaddha, N.; Gray, R.M. Asymptotic performance of vector quantizers with a perceptual distortion measure. *IEEE Trans. Inform. Theory* **1999**, *45*, 1082–1091.
39. Wolf, J.K.; Ziv, J. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Trans. Inform. Theory* **1970**, *16*, 406–411. [CrossRef]
40. Liao, Y.; Shlezinger, N.F.N.; Eldar, Y.C.; Goldsmith, A. Deep symbol detection for millimeter wave communications. In Proceedings of the GLOBECOM, Waikola, HI, USA, 9–13 December 2019.
41. Wannamaker, R.A.; Lipshitz, S.P.; Vanderkooy, J.; Wright, J.N. A theory of nonsubtractive dither. *IEEE Trans. Signal Process.* **2000**, *48*, 499–516. [CrossRef]
42. Rose, K.; Gurewitz, E.; Fox, G.C. Vector quantization by deterministic annealing. *IEEE Trans. Inform. Theory* **1992**, *38*, 1249–1257. [CrossRef]
43. Danial, L.; Wainstein, N.; Kraus, S.; Kvatinsky, S. Breaking through the speed-power-accuracy tradeoff in ADCs using a memristive neuromorphic architecture. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *2*, 396–409.
44. Gong, T.; Shlezinger, N.; Ioushua, S.S.; Namer, M.; Yang, Z.; Eldar, Y.C. RF chain reduction for MIMO systems: A hardware prototype. *IEEE Syst. J.* **2020**, *14*, 5296–5307. [CrossRef]
45. Shlezinger, N.; Eldar, O.D.Y.C.; Yoo, I.; Imani, M.F.; Smith, D.R. Dynamic metasurface antennas for uplink massive MIMO systems. *IEEE Trans. Commun.* **2019**, *67*, 6829–6843. [CrossRef]
46. Wang, H.; Shlezinger, N.; Eldar, Y.C.; Jin, S.; Imani, M.F.; Yoo, I.; Smith, D.R. Dynamic metasurface antennas for MIMO-OFDM receivers with bit-limited ADCs. *arXiv* **2019**, arXiv:1912.06917.
47. Shlezinger, N.; Alexandropoulos, G.C.; Imani, M.F.; Eldar, Y.C.; Smith, D.R. Dynamic metasurface antennas for 6G extreme massive MIMO communications. *arXiv* **2020**, arXiv:2006.07838.
48. Shlezinger, N.; Farsad, N.; Eldar, Y.C.; Goldsmith, A.J. ViterbiNet: A deep learning based Viterbi algorithm for symbol detection. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3319–3331. [CrossRef]
49. Shlezinger, N.; Fu, R.; Eldar, Y.C. DeepSIC: Deep soft interference cancellation for multiuser MIMO detection. *arXiv* **2020**, arXiv:2002.03214.
50. Shlezinger, N.; Farsad, N.; Eldar, Y.C.; Goldsmith, A.J. Learned factor graphs for inference from stationary time sequences. *arXiv* **2020**, arXiv:2006.03258.
51. Widrow, B.; Lehr, M.A. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proc. IEEE* **1990**, *78*, 1415–1442. [CrossRef]
52. Sutton, R.G.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.