Communications in Information and Systems
Volume 20, Number 2, 131–162, 2020

Task-based quantization with application to MIMO receivers

Nir Shlezinger and Yonina C. Eldar*

Multiple-input multiple-output (MIMO) systems are required to communicate reliably at high spectral bands using a large number of antennas, while operating under strict power and cost constraints. In order to meet these constraints, future MIMO receivers are expected to operate with low resolution quantizers, namely, utilize a limited number of bits for representing their observed measurements, inherently distorting the digital representation of the acquired signals. The fact that MIMO receivers use their measurements for some task, such as symbol detection and channel estimation, other than recovering the underlying analog signal, indicates that the distortion induced by bit-constrained quantization can be reduced by designing the acquisition scheme in light of the system task, i.e., by *task-based quantization*. In this work we survey the theory and design approaches to task-based quantization, presenting model-aware designs as well as data-driven implementations. Such task-based quantizers are shown to notably outperform conventional approaches which the desired information from low-resolution measurements solely in the digital domain. Then, we show how one can implement a task-based bit-constrained MIMO receiver, presenting approaches ranging from conventional hybrid receiver architectures to structures exploiting the dynamic nature of metasurface antennas. This survey narrows the gap between theoretical task-based quantization and its implementation in practice, providing concrete algorithmic and hardware design principles for realizing task-based MIMO receivers.

1. Introduction

Modern wireless communications systems face a growing set of demands and challenges. Cellular base stations (BSs) are required to reliably provide high

*This work received funding from the Benoziyo Endowment Fund for the Advancement of Science, the Estate of Olga Klein – Astrachan, the European Union's Horizon 2020 research and innovation program under grant No. 646804-ERC-COG-BNYQ, and the Israel Science Foundation under grant No. 0100101.

throughput to an increasing number of user terminals (UTs), while maintaining feasible cost and power consumption. An emerging technology to meet these demands is to equip the wireless BSs with a large number of antenna elements, realizing *massive multiple-input multiple-output (MIMO) communications*. Theoretical studies indicate that substantial gains in spectral efficiency can be achieved by letting the number of BS antennas grow arbitrarily large [1, 2]. An additional method to increase the network throughput is to explore the millimeter wave (mmWave) frequency range [3], thus overcoming the spectral congestion of traditional wireless bands. Such mmWave communications is particularly suitable for massive MIMO systems: The short wavelengths of mmWave signals allows packing a large number of antenna elements at a small physical size, and the massive number of elements facilitates directed beamforming which is essential at mmWave bands.

While the theoretical gains of massive MIMO systems, particularly when combined with mmWave transmission, are clear, implementing such systems in practice under strict cost and power constraints is a challenging task. A major source of this increased cost are the analog-to-digital convertor (ADC) components, which allow the analog signals observed by each antenna element to be processed in digital. The power consumption of an ADC is directly related to the signal bandwidth and the number of bits used for digital representation [4]. Consequently, in massive MIMO systems, where the number of antennas and ADCs operating at high frequency bands is large, limiting the number of bits, thus operating under quantization constraints, is crucial to keep cost and power consumption feasible [3].

Focusing on uplink communications, i.e., when the BS acts as the receiver, quantization constraints imply that the BS cannot process the channel output directly but rather only an inaccurate distorted digital representation of it. The distortion induced by continuous-to-discrete quantization mappings degrades the ability to extract information, such as the underlying channel coefficients or the transmitted signal, from the observed channel output. Consequently, methods for channel estimation and symbol detection from quantized outputs are the focus of a large body of work, including, e.g., [5, 6, 7]. These schemes are carried out in the digital domain, i.e., they are *digital-only methods*, assuming a fixed quantization system.

An alternative emerging approach to processing only in the digital domain, which is the focus of the current survey, is to jointly design the quantization system along with the digital processing in light of the task as proposed in [8]. Such *task-based quantization* systems convert their received analog signal into a digital representation in a manner which preserves the semantic information required to carry out the task, rather than recovering

the analog signal, thus allowing to operate efficiently with standard ADCs under relatively tight bit constraints [8, 9, 10, 11]. Task-based quantization was shown to achieve notably improved accuracy in recovering the desired task compared to conventional digital-only methods operating under the same bit budget. Consequently, task-based quantizers, originally derived for generic digital signal processing applications in [8], bear the potential of significantly facilitating the design of massive MIMO receivers operating under bit constraints [9]. This follows since in MIMO systems, acquisition is carried out for specific tasks, most commonly channel estimation and symbol detection. These can be treated as recovering information embedded in the received signals, which in turn can be accurately and compactly extracted in digital using task-based quantization.

In this work we survey recent results in task-based quantization. We focus on its application for bit-constrained MIMO receivers, although task-based quantization is relevant in many other applications including sensor arrays, radar, medical imaging, and essentially any system which acquires physical signals for some task while operating under bit constraints. We begin by detailing model-aware methods for designing task-based quantizers. These methods jointly design the overall acquisition system along with the digital processing based on prior knowledge of the statistical model relating the observed analog signal and the desired task information to be extracted in digital. Our model-aware analysis characterizes the achievable accuracy in recovering the desired information under bit constraints for tasks which can be modeled as a linear function of the measurements as in, e.g., Rayleigh fading MIMO channel estimation [9]. Then, we show how the proposed approach can be extended to more involved tasks by utilizing the mathematical tool of principal inertia compoenents (PICs) [12]. Specifically, we show that PICs can facilitate identifying a proper transformation of the measurements from which the task can be treated as approximately linear, allowing to use the proposed task-based quantizer. We specialize the derivation for tasks where the desired information is encapsulated in quadratic functions of the measurements, which is the case in, e.g., covariance estimation [13] and direction of arrival (DOA) recovery [14]. For all considered tasks, task-based quantization is shown to achieve substantially improved accuracy in recovering the desired information compared to digital-only methods.

Next, we show how task-based quantization systems can be designed without explicitly specifying the statistical relationship between the observations and the desired task information, by tuning the overall acquisition system in a data-driven manner. We demonstrate how by combining machine learning (ML) tools with an accurate differentiable approximation of

the quantization rule, one can learn task-based quantization mappings from a set of labeled data. We demonstrate that learned task-based quantizers facilitates the recovery of information embedded in the observations in a complex manner when operating under tight bit constraints, while notably outperforming purely digital approaches with the same bit budget. Finally, we show how to implement MIMO receivers capable of dynamically adjusting their acquisition system in light of the task, thus realizing tunable task-based quantization. Our proposed design builds upon either conventional hybrid receiver architectures [15, 16], dedicated pre-acquisition hardware [17], or on exploiting the inherent configurability of receivers equipped with metasurface antennas [18, 19, 20], and we present hardware prototypes built in our lab, demonstrating the feasibility of task-based quantization in MIMO receivers.

The rest of this paper is organized as follows: Section 2 formulates the system model and reviews some basics in quantization theory. Methods for designing task-based quantizers based on prior model knowledge are detailed in Section 3. Section 4 presents data-driven design strategies. In Section 5 we show how one can implement task-based quantization in bit-constrained MIMO receivers, reviewing several candidate architectures and hardware prototypes. Section 6 provides some concluding remarks.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$, where the ith element of $\boldsymbol{x}$ is written as $(\boldsymbol{x})_i$. Boldface upper-case letters are used for matrices, e.g., $\boldsymbol{M}$, where $(\boldsymbol{M})_{i,j}$ denotes its (i,j)th element. 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, Kronecker product, and stochastic expectation are written as $(\cdot)^T$, $\|\cdot\|$, $\otimes$, and $\mathbb{E}\{\cdot\}$, respectively, and $\mathcal{R}$ is the set of real numbers. All logarithms are taken to basis two.

2. Preliminaries and problem formulation

2.1. Preliminaries in quantization theory

We begin by briefly reviewing the standard quantization setup, and recall the definition of a quantizer:

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

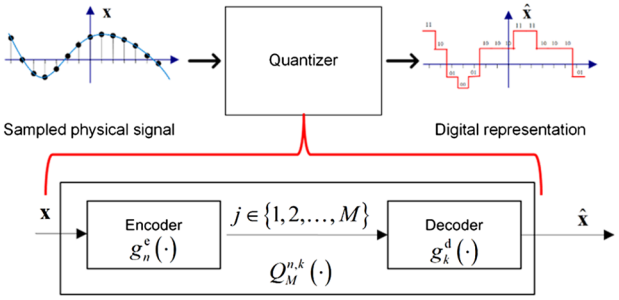

Figure 1: Quantizer illustration.

We write the output of the quantizer with input $\boldsymbol{x} \in \mathcal{X}^n$ as $\hat{\boldsymbol{x}} = g_k^{\mathrm{d}}\left(g_n^{\mathrm{e}}\left(\boldsymbol{x}\right)\right) \triangleq Q_M^{n,k}\left(\boldsymbol{x}\right)$. *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. An illustration of a quantization system is depicted in Fig. 1.

In the standard quantization problem, a $Q_M^{n,n}\left(\cdot\right)$ 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 its quantization rate $R \triangleq \frac{1}{n}\log M$, and the expected distortion $\mathbb{E}\{d\left(\boldsymbol{x}, \hat{\boldsymbol{x}}\right)\}$. For a fixed input size n and codebook size M, the optimal quantizer is $Q_M^{n,\mathrm{opt}}\left(\cdot\right) = \arg\min_{Q_M^{n,n}} \mathbb{E}\left\{d\left(\boldsymbol{x}, Q_M^{n,n}\left(\boldsymbol{x}\right)\right)\right\}$. Characterizing the optimal quantizer 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., [21], or asymptotically large inputs, namely, $n \to \infty$, via rate-distortion theory [22, Ch. 10].

2.2. Problem formulation

Here, we study *task-based quantization* [8], where 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 $\boldsymbol{s} \in \mathcal{S}^k \subseteq \mathcal{R}^k$ from a statistically dependent random vector $\boldsymbol{x} \in \mathcal{R}^n$ of larger dimensionality, i.e., $n \geq k$. The set $\mathcal{S}$ represents the possible values of the unknown vector: It can be continuous, representing an estimation task; or discrete, for classification tasks. This formulation accommodates a broad range of applications, including channel estimation and symbol detection, that are the common tasks considered in MIMO communications receivers [9], as well as covariance recovery [13] and DOA estimation [14]. The recovered estimate of $\boldsymbol{s}$, denoted $\hat{\boldsymbol{s}}$, is represented in digital using

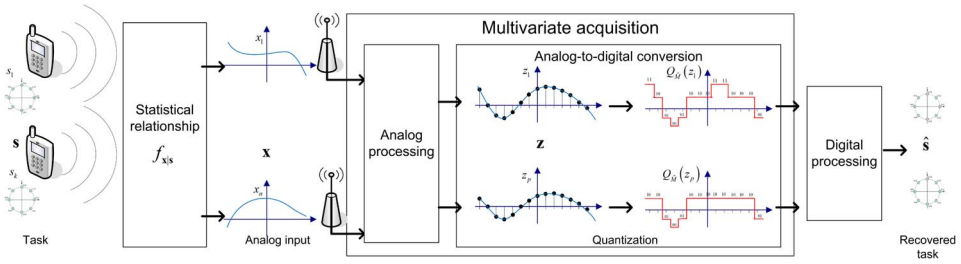

Figure 2: Hybrid quantization system model. For illustration, the task is recovering a set of constellation symbols in uplink MIMO communications.

up to $\log M$ bits, dictating the bit budget allowed for task-based quantization. The observed $\boldsymbol{x}$ is related to $\boldsymbol{s}$ via a conditional probability measure $f_{\boldsymbol{x}|\boldsymbol{s}}$. For example, in a communications setup. the conditional probability measure $f_{\boldsymbol{x}|\boldsymbol{s}}$ encapsulates the noisy channel.

The performance limits of task-based quantization with asymptotically large vectors, i.e., when $n \to \infty$ while $R = \frac{1}{n} \log M$ remains fixed, can be characterized using indirect rate-distortion theory [23]. Specifically, for estimation tasks with the mean-squared error (MSE) distortion objective, i.e., $d(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2$, the task-based quantization mapping which minimizes the MSE for a fixed quantization rate R was derived in [24] for fixed-size vectors. The resulting optimal strategy consists of applying vector quantization to the minimum MSE (MMSE) estimate of $\boldsymbol{s}$ from $\boldsymbol{x}$.

While vector quantizers allow to achieve more accurate digital representations of the acquired analog signal compared to their scalar counterparts [25, Ch. 23], practical ADCs typically utilize scalar quantizers. In particular, ADCs often apply the same continuous-to-discrete mapping to each sample, which is most commonly based on a uniform partition of the real line, i.e., scalar uniform quantization [4]. Nonetheless, in the presence of a task, one is not interested in recovering the analog signal, but rather estimate some underlying information embedded in it. This motivates the analysis of how to incorporate the presence of a task in the design of a quantization system utilizing scalar ADCs, and whether the distortion induced by conventional scalar quantization can be mitigated when recovering the task.

2.3. Hardware-limited task-based quantization

As discussed in the previous section, practical digital signal processing systems typically obtain a discrete representation of physical analog signals

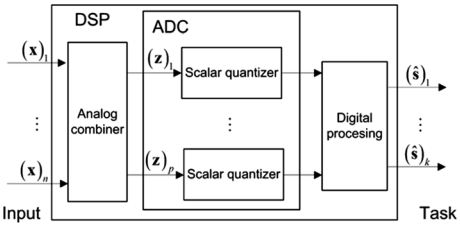

Figure 3: Block diagram of considered task-based quantization systems.

using scalar ADCs. In such systems, each continuous-amplitude sample is converted into a discrete representation using a single quantization rule. Therefore, in order to be able to account for the presence of a task in acquisition while operating with scalar ADCs, one must introduce some level of processing, in addition to that carried out in digital. We therefore consider hybrid acquisition systems as illustrated in Fig. 2, which is a common model in MIMO communication receivers [15, 16]. Hybrid architectures were originally proposed as a method to reduce the number of costly RF chains in MIMO receivers [15, 16], while here we exploit these structures to allow quantization under bit constraints for tasks. In such hybrid systems, a set of analog signals can be combined in analog prior to being converted to digital, a property which we exploit in order to facilitate extracting 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, we henceforth focus only the quantization aspect assuming a fixed sampling mechanism. The joint design of sampling and quantization in light of a task is left for future studies; initial results can be found in [26].

In the proposed hybrid architecture, the input to the ADC, denoted $\boldsymbol{z} \in \mathcal{R}^p$, where p denotes the number of scalar quantizers, is obtained from $\boldsymbol{x}$ using a pre-quantization mapping referred to as *analog combining*. Then, $\boldsymbol{z}$ is quantized using p identical scalar quantizers with resolution $\tilde{M} \triangleq \lfloor M^{1/p} \rfloor$ into a digital vector $Q(\boldsymbol{z})$. The overall number of bits is $p \cdot \log \tilde{M} \leq \log M$. The ADC output is processed in digital to obtain the estimate $\hat{\boldsymbol{s}} \in \mathcal{S}^k$. A schematic block diagram of the quantization system is depicted in Fig. 3. Designing task-based quantizers can be formulated as the joint optimization of the analog combining mapping, the scalar quantization rule, and the digital processing, such that the output $\hat{\boldsymbol{s}}$ will be an accurate estimate of the task vector $\boldsymbol{s}$, while operating under a fixed budget of up to $\log M$ bits.

The characterization of task-based quantization systems of the form of Fig. 3 consists of two complementary studies: First, we study in Section 3 how the overall system can be designed based on knowledge of the conditional distribution $f_{\boldsymbol{x}|\boldsymbol{s}}$ relating the observations and the task in a model-based fashion. Then, we discuss how task-based quantization mappings can be learned from labeled data building upon ML tools, and in particular, by utilizing deep neural networks (DNNs) to adapt task-based quantization mappings, in Section 4. Our results demonstrate that by properly tuning the hybrid architecture of task-based quantizers, one can approach the performance limits dictated by indirect rate-distortion theory, achievable using complex vector quantizers, while using conventional scalar ADCs operating as part of an acquisition system of feasible hardware requirements. These studies, which consider either purely model-based or purely data-driven designs, can be used as a basis for future research on hybrid model-based and data-driven systems, as in, e.g., [27], for task-based quantization.

3. Model-aware task-based quantization

In this section we detail how to design hybrid quantization systems to facilitate the recovery of the task vector $\boldsymbol{s}$ in the digital domain, based on prior knowledge of the underlying statistical model. In particular, we discuss how the analog combining, quantization rule, and digital processing components of the system in Fig. 3 can be jointly optimized based on knowledge of the conditional distribution relating the input $\boldsymbol{x}$ to the task vector $\boldsymbol{s}$, denoted $f_{\boldsymbol{x}|\boldsymbol{s}}$. We begin by presenting the model assumptions under which the analysis is carried out in Section 3.1. After that we present the resulting task-based quantization systems for estimation tasks of linear and quadratic nature in Sections 3.2–3.3, respectively.

3.1. System model

In order to obtain a meaningful and tractable characterization of the task-based quantization system of Fig. 3, we henceforth introduce two model assumptions upon which we base our results in the remainder of this section:

A1 We consider the task of estimating the task $\boldsymbol{s}$ in the MSE sense, namely, our performance measure is the MSE $\mathbb{E}\{\|\boldsymbol{s}-\hat{\boldsymbol{s}}\|^2\}$.

A2 We focus on uniform ADCs, and model the their operation in our derivations as non-subtractive uniform dithered quantizers [28].

Model assumption *A1* implies that the fidelity of an estimate $\hat{\boldsymbol{s}}$ can be represented as a sum of the MMSE and the excess MSE with respect to the MMSE estimate $\tilde{\boldsymbol{s}} = \mathbb{E}\{\boldsymbol{s}|\boldsymbol{x}\}$, as $\mathbb{E}\{\|\boldsymbol{s}-\hat{\boldsymbol{s}}\|^2\} = \mathbb{E}\{\|\boldsymbol{s}-\tilde{\boldsymbol{s}}\|^2\} + \mathbb{E}\{\|\tilde{\boldsymbol{s}}-\hat{\boldsymbol{s}}\|^2\}$. Consequently, in the following we characterize the performance in terms of the excess MSE $\mathbb{E}\{\|\tilde{\boldsymbol{s}}-\hat{\boldsymbol{s}}\|^2\}$. Since $\tilde{\boldsymbol{s}}$ is a function of $\boldsymbol{x}$, we divide our analysis based on the nature of this function, considering linear functions in Section 3.2, extending to quadratic and more general forms in Section 3.3.

Model assumption *A2* imposes a structure on the scalar quantization mapping. To formulate the resulting input-output relationship of the ADCs, let γ denote the support of the quantizer, and define $\Delta \triangleq \frac{2\gamma}{\tilde{M}}$ as the quantization spacing. The output of the uniform ADC with input sequence $z_1, z_2, \ldots, z_p$ can be written as $Q(z_i) = q(z_i + u_i)$, where $u_1, u_2, \ldots, u_p$ are i.i.d. random variables (RVs) uniformly distributed over $\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]$, mutually independent of the input, representing the dither signal. The function $q(\cdot)$, which implements the uniform quantization, is given by

$$(1) \quad q(z) = \begin{cases} -\gamma + \Delta\left(l - \frac{1}{2}\right) & z - l\Delta \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right], l \in \{0, 1, \ldots, \tilde{M}-1\} \\ \operatorname{sign}(z)\left(\gamma - \frac{\Delta}{2}\right) & |z| > \gamma. \end{cases}$$

When $\tilde{M} = 2$, the resulting quantizer is a standard one-bit sign quantizer of the form $q(z) = c \cdot \operatorname{sign}(z)$, where $c > 0$ is determined by the support γ.

Dithered quantizers significantly facilitate the analysis, due to the following favorable property: When operating within the support, the output can be written as the sum of the input and an additive zero-mean white quantization noise signal uncorrelated with the input. The drawback of adding dither is that it increases the energy of the quantization noise, namely, it results in increased distortion [28]. Nonetheless, the favorable property of dithered quantization is also satisfied in uniform quantization *without dithering* for inputs with bandlimited characteristic functions, and is approximately satisfied for various families of input distributions [29]. Consequently, while our analysis assumes dithered quantization, exploiting the resulting statistical properties of the quantization noise, the proposed system is applicable without dithering, as we demonstrate in our numerical study.

3.2. Linear estimation tasks

We begin by focusing on scenarios in which the stochastic relationship between the vector of interest $\boldsymbol{s}$ and the observations $\boldsymbol{x}$ are such that the MMSE estimate of $\boldsymbol{s}$ from $\boldsymbol{x}$ is a linear function of $\boldsymbol{x}$, i.e., $\exists \boldsymbol{\Gamma} \in \mathcal{R}^{k\times n}$ such that $\tilde{\boldsymbol{s}} = \boldsymbol{\Gamma}\boldsymbol{x}$. Accordingly, we restrict the analog combining and the digital

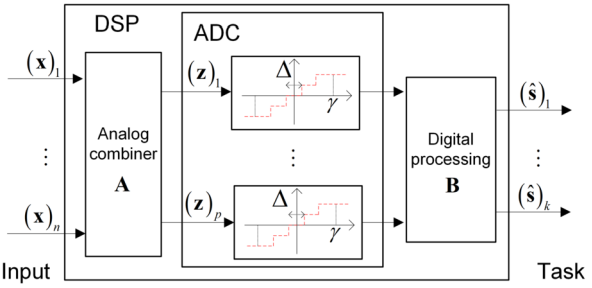

Figure 4: Model-aware task-based quantization for linear tasks illustration.

mapping components in Fig. 3 to be linear, namely, $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$ and $\hat{\boldsymbol{s}} = \boldsymbol{B}Q(\boldsymbol{z})$, for some $\boldsymbol{A} \in \mathcal{R}^{p\times n}$ and $\boldsymbol{B} \in \mathcal{R}^{k\times p}$. An illustration of the considered system architecture is depicted in Fig. 4. By focusing on these setups, we are able to explicitly derive the achievable distortion and to characterize the system which minimizes the MSE. This derivation reveals some non-trivial insights. For example, we show that the optimal approach when using vector quantizers, namely, to quantize the MMSE estimate [24], is no longer optimal when using standard scalar ADCs. Furthermore, as detailed in Section 3.3, our analysis provides guidelines for designing task-based quantization systems which can be used for more general relationships between $\boldsymbol{s}$ and $\boldsymbol{x}$, such as the recovery of quadratic tasks.

Let $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ be the covariance matrix of $\boldsymbol{x}$, assumed to be non-singular. Before we study the overall task-based quantization system, we first derive the digital processing matrix which minimizes the MSE for a given analog combiner $\boldsymbol{A}$ and the resulting MSE, stated in the following lemma [8, Lem. 1]:

Lemma 2. *For any analog combining matrix $\boldsymbol{A}$ and support γ such that the quantizers operate within their support, i.e.,* $\Pr\left(\left|(\boldsymbol{A}\boldsymbol{x})_l + u_l\right| > \gamma\right) = 0$*, the digital processing matrix which minimizes the MSE is given by*

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

and the achievable excess MSE, denoted $\mathrm{MSE}(\boldsymbol{A}) = \min_{\boldsymbol{B}} \mathbb{E}\left\{\left\|\tilde{\boldsymbol{s}} - \hat{\boldsymbol{s}}\right\|^2\right\}$*, is*

$$\mathrm{MSE}(\boldsymbol{A}) = \mathrm{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Gamma}^T - \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T + \frac{2\gamma^2}{3\tilde{M}^2}\mathbf{I}_p\right)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Gamma}^T\right). \tag{2b}$$

The digital processing matrix in Lemma 2 is the linear MMSE estimator of $\boldsymbol{s}$ from the vector $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e}$ represents the quantization noise,

which is white and uncorrelated with $\boldsymbol{Ax}$. This stochastic representation is a result of the usage of non-overloaded dithered quantizers. Nonetheless, in the following we use the model on which Lemma 2 is based to design task-based quantizers operating with small yet non-zero probability of overloading, i.e., $\Pr\left(\left|(\boldsymbol{Ax})_l + u_l\right| > \gamma\right) \approx 0$ for each l. In such cases modeling $\boldsymbol{Ax}$ and $\boldsymbol{e}$ as uncorrelated becomes a reliable approximation. Therefore, in order to use Lemma 2 to design task-based quantizers, we explicitly require to avoid overloading with high probability. This is achieved by fixing γ to be some multiple η of the maximal standard deviation of the input, allowing to bound the overload probability via Chebyshev's inequality [22, Pg. 64].

We now use Lemma 2 to obtain the analog combining matrix $\boldsymbol{A}^{\mathrm{o}}$ which minimizes the MSE and the resulting system. Define the matrix $\tilde{\boldsymbol{\Gamma}} \triangleq \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$, let $\{\lambda_{\tilde{\boldsymbol{\Gamma}},i}\}$ be its singular values arranged in a descending order, and set $\kappa \triangleq \eta^2\left(1-\frac{\eta^2}{3\tilde{M}^2}\right)^{-1}$. Note that for $i > \mathrm{rank}\left(\tilde{\boldsymbol{\Gamma}}\right)$, $\lambda_{\tilde{\boldsymbol{\Gamma}},i} = 0$. The resulting task-based quantization system is stated in the following theorem [8, Thm. 1]:

Theorem 3. *For the task-based quantization system under linear estimation tasks, the analog combining matrix $\boldsymbol{A}^{\mathrm{o}}$ is given by $\boldsymbol{A}^{\mathrm{o}} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{V}_{\boldsymbol{A}}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1/2}$, where $\boldsymbol{V}_{\boldsymbol{A}} \in \mathcal{R}^{n\times n}$ is the right singular vectors matrix of $\tilde{\boldsymbol{\Gamma}}$; $\boldsymbol{\Lambda}_{\boldsymbol{A}} \in \mathcal{R}^{p\times n}$ is a diagonal matrix with diagonal entries*

$$(\boldsymbol{\Lambda}_{\boldsymbol{A}})_{i,i}^2 = \frac{2\kappa}{3\tilde{M}^2\cdot p}\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+, \tag{3a}$$

with ζ set such that $\frac{2\kappa}{3\tilde{M}^2\cdot p}\sum_{i=1}^{p}\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+ = 1$; and $\boldsymbol{U}_{\boldsymbol{A}} \in \mathcal{R}^{p\times p}$ is a unitary matrix which guarantees that $\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T\boldsymbol{U}_{\boldsymbol{A}}^T$ is weakly majorized by all possible rotations of $\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T$. The support of the ADC is given by $\gamma^2 = \frac{\kappa}{p}$, and the digital processing matrix is equal to

$$\boldsymbol{B}^{\mathrm{o}}\left(\boldsymbol{A}^{\mathrm{o}}\right) = \tilde{\boldsymbol{\Gamma}}\boldsymbol{V}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T\left(\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}^T + \frac{2\gamma^2}{3\tilde{M}^2}\boldsymbol{I}_p\right)^{-1}\boldsymbol{U}_{\boldsymbol{A}}^T. \tag{3b}$$

The resulting minimal achievable excess MSE is

$$\mathbb{E}\left\{\left\|\tilde{\boldsymbol{s}}-\hat{\boldsymbol{s}}\right\|^2\right\} = \begin{cases} \sum_{i=1}^{k}\frac{\lambda_{\tilde{\boldsymbol{\Gamma}},i}^2}{\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+ +1}, & p\geq k \\ \sum_{i=1}^{p}\frac{\lambda_{\tilde{\boldsymbol{\Gamma}},i}^2}{\left(\zeta\cdot\lambda_{\tilde{\boldsymbol{\Gamma}},i}-1\right)^+ +1} + \sum_{i=p+1}^{k}\lambda_{\tilde{\boldsymbol{\Gamma}},i}^2, & p<k. \end{cases} \tag{3c}$$

Since the design objective is the MSE by *A1*, the optimal quantization system utilizing vector quantizers is known to recover $\tilde{\boldsymbol{s}} = \boldsymbol{\Gamma}\boldsymbol{x}$ in the analog

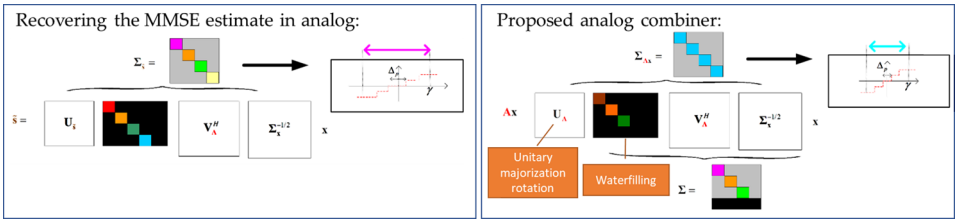

Figure 5: An illustration of the ADC input, its covariance, and the resulting quantization mapping when quantizing the MMSE estimate (left) and for the proposed combiner of Theorem 3 (right).

domain [24]. In the presence of scalar ADCs, Theorem 3 reveals two main differences in the desired pre-quantization mapping: First, the analog combiner essentially nullifies the weak eigenmodes of the correlation matrix of the MMSE estimate in (3a), as these eigenmodes are likely to become indistinguishable by finite resolution uniform scalar quantization. Then, the unitary rotation matrix $\boldsymbol{U_A}$, which guarantees that the entries of $\boldsymbol{z}$ have the same variance, minimizes the maximal variance of the quantized variables, allowing to use relatively fine quantization at a given resolution without risking high overloading probability. This combined operation of the analog mapping trades estimation error and quantization accuracy, allowing to optimize the digital representation in light of the task. An illustration of this analog combiner and its quantization rule compared to recovering $\tilde{\boldsymbol{s}}$ in analog is depicted in Fig. 5.

The characterization of the task-based quantization system in Theorem 3 gives rise to the following non-trivial insights: 1) In order to minimize the MSE, p must not be larger than the rank of the covariance matrix of $\tilde{\boldsymbol{s}}$ [8, Cor. 1]. This implies that reducing the dimensionality of the input prior to quantization contributes to recovering the task vector as higher resolution quantizers can be used without violating the overall bit constraint; and 2) When the covariance matrix of $\tilde{\boldsymbol{s}}$ is non-singular, quantizing the MMSE estimate minimizes the MSE if and only if the covariance matrix of $\tilde{\boldsymbol{s}}$ equals $\boldsymbol{I}_k$ up to a constant factor [8, Cor. 4]. This indicates that, except for very specific statistical models, quantizing the entries of the MMSE estimate vector, which is the optimal strategy when using vector quantizers [24], does not minimize the MSE when using uniform scalar ADCs.

To illustrate the gains of the task-based quantization system design which arises from Theorem 3, we next numerically evaluate its achievable MSE in a simulation study. We consider the estimation of a scalar intersymbol interference (ISI) channel from quantized observations. In this scenario,

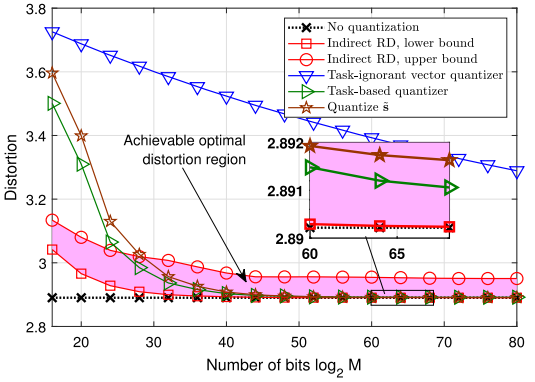

Figure 6: ISI channel recovery.

the parameter vector $\boldsymbol{s}$ represents the coefficients of a multipath channel with k taps. The channel is estimated from a set of $n = 120$ noisy observations $\boldsymbol{x}$, given by $(\boldsymbol{x})_i = \sum_{l=1}^{k} (\boldsymbol{s})_l \, a_{i-l+1} + v_i$, where a_i is a deterministic known training sequence, and $\{v_i\}_{i=1}^{n}$ are samples from an i.i.d. zero-mean unit variance Gaussian noise process independent of $\boldsymbol{s}$. In particular, the channel $\boldsymbol{s}$ is modeled as a $k = 8$ tap zero-mean Gaussian vector with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}}$, given by $\left(\boldsymbol{\Sigma}_{\boldsymbol{s}}\right)_{i,j} = e^{-|i-j|}$, $i, j \in \{1, 2, \ldots, k\}$, and $a_i = \cos\left(\frac{2\pi i}{n}\right)$ for $i > 0$ and $a_i = 0$ otherwise. Since $\boldsymbol{s}$ and $\boldsymbol{x}$ are jointly Gaussian, the MMSE estimate is a linear function of $\boldsymbol{x}$.

The MSE achievable by the task-based quantization system designed via Theorem 3 operating with conventional non-dithered uniform quantizers is compared to the MSE in recovering the MMSE estimate in analog prior to quantization, i.e., setting $\boldsymbol{A} = \boldsymbol{\Gamma}$. We also numerically evaluate upper and lower bounds on the minimal MSE under quantization constraints, achievable via indirect rate-distortion theory by applying the rate-distortion optimal source code to $\tilde{\boldsymbol{s}}$ (and thus given explicitly only in the limit $k \to \infty$ [30]), computed via [8, Prop. 1]. Finally, we evaluate the achievable MSE in applying a vector quantizer designed to accurately represent $\boldsymbol{x}$, from which $\boldsymbol{s}$ is estimated in digital, computed via [8, Prop. 2]. The latter intuitively represents the vector quantization system one would design without prior knowledge of the task for which $\boldsymbol{x}$ is acquired, and is thus referred to as task-ignorant vector quantizer. The MSE values are depicted in Fig. 6.

Observing Fig. 6, we note that the task-based quantizer substantially outperforms task-ignorant vector quantization, and approaches the optimal performance as M increases. In particular, when each scalar quantizer uses at least five bits, i.e., $\log M \geq 5k$, the quantization error becomes negligible

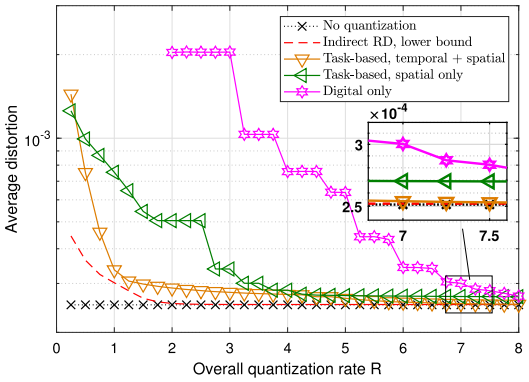

Figure 7: MIMO channel recovery.

and the overall distortion is effectively the minimum achievable estimation error, i.e., the MMSE. Furthermore, we note that task-based quantization outperforms recovering $\tilde{\boldsymbol{s}}$ in analog, and the gain is most notable at small values of M. These results demonstrate that by accounting for the presence of a task via joint optimization of the analog combiner, quantization rule, and digital processing, one can approach the optimal performance, dictated by indirect rate-distortion theory, using standard uniform ADCs commonly used in digital signal processing systems.

To evaluate the performance of task-based quantization in massive MIMO systems, we consider the recovery of multi-cell MIMO channel based on the setup detailed in [9, Sec. V]. Here, the system consists of 7 cells with 10 single-antenna UTs in each cell, and the receiver, that is equipped with 100 antennas, estimates its intra-cell 100×10 channel matrix from the channel output, which is corrupted by intercell interference and Gaussian noise with variance of 10^{-3}. The UTs are uniformly distributed in a hexagonal cell of radius 400 m, following the model in [1], with receive side correlation dictated by Jakes model with 0.4 wavelength element spacing [31]. Estimation is carried out based on 40 pilot symbols determined by the first 10 rows of the 40×40 discrete Fourier transform (DFT) matrix.

The average MSE of the proposed task-based quantizer compared to the indirect rate-distortion bound and the MMSE achievable without quantization constraints is depicted in Fig. 7. The input vector $\boldsymbol{x}$ here represents the channel outputs corresponding to all transmitted pilot symbols, and thus the system designed via Theorem 3 combines samples received at different time instances, which may be difficult to implement in practice. Therefore, we also depict in Fig. 7 the MSE when the analog processing is restricted

to combine only samples received at the same time instance using the same linear mapping, i.e., spatial only combining, obtained using [9, Prop. 4]. Finally, we depict the MSE without analog combining, i.e., a digital only receiver, in which the digital processing is based on the linear MMSE channel estimator from quantized measurements, and thus consists a bound on the performance achievable using approximations of the linear MMSE estimator, such the channel estimator proposed in [5].

Observing Fig. 7 we note that, similarly to the ISI channel in Fig. 6, the MSE achievable using task-based quantization is within a very small gap from the indirect rate-distortion curve for quantization rates larger than $R = 1.5$. The task-based quantizer with spatial combining is capable of achieving near-optimal performance for $R > 3$, due to its ability to exploit the spatial correlation. It is also observed that the average MSE of estimating the channel only in the digital domain is notably higher compared to task-based quantization, which jointly operates in both analog and digital while tuning the quantization rule accordingly, demonstrating the gains of task-based quantization over digital-only designs.

3.3. Quadratic estimation tasks

In the previous section we showed that allowing the analog mapping to reduce dimensionality and rotate the quantized signal can contribute to the overall recovery performance by balancing estimation and quantization errors. However, this analysis was carried out only for scenarios in which $\tilde{\boldsymbol{s}}$ is a linear function of $\boldsymbol{x}$, resulting in $\mathbb{E}\{\boldsymbol{s}|\boldsymbol{z}\}$ being a linear function of the input to the quantizers $\boldsymbol{z}$. In many scenarios of interest, such as covariance estimation [13] and DOA recovery [14] from quantized measurements, the desired information can be extracted from a quadratic function of the measurements, i.e., functions $\{\boldsymbol{x}^T\boldsymbol{C}_i\boldsymbol{x}\}_{i=1}^k$, where each $\boldsymbol{C}_i \in \mathcal{R}^{n\times n}$ is symmetric.

Here, we show how the analysis of the previous section can be applied for designing task-based quantizers for the task of recovering non-linear functions of $\boldsymbol{x}$ under quantization constraints, focusing on quadratic functions and Gaussian inputs. Our strategy is based on identifying a family of analog mappings $h(\cdot)$ for which $\boldsymbol{z}$ corresponds to the scenario studied in Section 3.2. To that aim, we use PIC-based analysis [12], which provides a decomposition of the statistical relationship between two RVs, that is directly related to MMSE estimation. In particular, for a pair of RVs (x, y), the principal inertia functions $\{f_i(\cdot)\}$ and $\{g_i(\cdot)\}$ formulate an orthonormal basis spanning the Hilbert space of functions of x and y, respectively, which diagonalize MMSE estimation, i.e., there exists a set of scalar coefficients $\{\rho_i\}$ such that

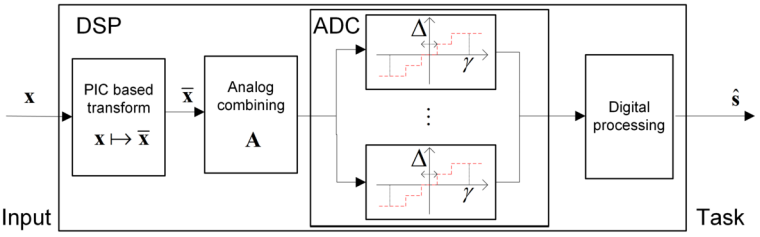

Figure 8: Quantization System of Fig. 2 with quadratic analog mapping.

$\mathbb{E}\{f_i(x)|y\} = \rho_i g + i(y)$ and $\mathbb{E}\{g_i(y)|x\} = \rho_i f_i(x)$. The benefit of using PICs in our context is their ability to decompose functions of the observations in a manner which reflects on the structure of the MMSE estimate. In particular, here we use this tool to identify a transformation of the input $\boldsymbol{x}$ under which recovering quadratic functions of it is converted to a linear manipulation. Defining $\bar{\boldsymbol{x}} \triangleq \text{vec}(\boldsymbol{x}\boldsymbol{x}^T)$, this results in the following theorem [10, Thm. 1]:

Theorem 4. *For any $p \times n^2$ matrix $\boldsymbol{A}$ with $p \leq n^2$, the MMSE estimate of $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x}$ from the vector $\boldsymbol{z} = \boldsymbol{A}(\bar{\boldsymbol{x}} - \mathbb{E}\{\bar{\boldsymbol{x}}\})$ can be written as*

$$\mathbb{E}\big\{f(\boldsymbol{x})|\boldsymbol{z}\big\} = \boldsymbol{d}^T\boldsymbol{z} + \mathbb{E}\{f(\boldsymbol{x})\}, \tag{4}$$

for some $p \times 1$ vector $\boldsymbol{d}$, which depends on $\boldsymbol{C}$, $\boldsymbol{A}$, and the covariance of $\boldsymbol{x}$.

Theorem 4 implies that the task-based quantization system design guidelines proposed in Theorem 3 can be utilized to facilitate the recovery of quadratic functions from quantized measurements by applying analog mappings of the form $\boldsymbol{z} = \boldsymbol{A}h(\boldsymbol{x}) = \boldsymbol{A}(\bar{\boldsymbol{x}} - \mathbb{E}\{\bar{\boldsymbol{x}}\})$. Here the matrix $\boldsymbol{A} \in \mathcal{R}^{p\times n^2}$ encapsulates the ability to reduce the dimensionality and to rotate the quantized vector, and can be designed via Theorem 3 by replacing the input $\boldsymbol{x}$ with $\bar{\boldsymbol{x}} - \mathbb{E}\{\bar{\boldsymbol{x}}\}$. The resulting quantization system is depicted in Fig. 8.

Although Theorem 4 specifically considers functionals $f(\boldsymbol{x})$ of a quadratic form, analogous schemes could be constructed for broader classes of functions. The main feature of Theorem 4 is the ability to represent $\mathbb{E}\{f(\boldsymbol{x})|\boldsymbol{z}\}$ either exactly, or possibly approximately, as a linear function of $\boldsymbol{z} = \boldsymbol{A}h(\boldsymbol{x})$ for some transformation $h(\cdot)$. Once the analog mapping satisfies this request, Theorem 3 can be applied to optimize the overall recovery accuracy of the quantization system. Formulated in terms of PICs, the choice of $h(\cdot)$ imposes structure on the joint distribution $(\boldsymbol{x}, \boldsymbol{z})$. Consequently, when the task is to recover a function $f(\boldsymbol{x})$ which can be decomposed using PICs as

$f(\boldsymbol{x}) = \sum \alpha_i f_i(\boldsymbol{x})$, any analog processing which results in $\boldsymbol{z}$ such that

$$\mathbb{E}\{f(\boldsymbol{x})|\boldsymbol{z}\} \approx \sum_{i=1}^{l} \alpha_i \rho_i(\boldsymbol{z})_i + \mathbb{E}\{f(\boldsymbol{x})\}, \tag{5}$$

would allow to design the analog pre-quantization step using existing tools derived for setups in which the MMSE estimate is linear. This implies that when recovering some function $f(\boldsymbol{x})$, the structure of the analog mapping should be designed as to yield linear basis functions $g_i(\boldsymbol{z})$, allowing the resulting system to be optimized using Theorem 3.

To demonstrate the ability of the proposed design to yield accurate task-based quantizers, we simulate an empirical covariance estimation scenario. Here, the input is given by $\boldsymbol{x} = [\boldsymbol{v}_1^T, \dots \boldsymbol{v}_4^T]^T$, where $\{\boldsymbol{v}_i\}_{i=1}^4$ are i.i.d. 3×1 zero-mean Gaussian random vectors, i.e., $n = 12$. The entries of the covariance matrix of $\boldsymbol{v}_i$, denoted $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, are $(\boldsymbol{\Sigma}_{\boldsymbol{v}})_{i,j} = e^{-|i-j|}$. The parameter of interest is the 3×3 empirical covariance matrix $\frac{1}{4}\sum_{i=1}^4 \boldsymbol{v}_i \boldsymbol{v}_i^T$, which is completely determined by its upper triangular matrix, stacked as the desired vector $\tilde{\boldsymbol{s}}$, thus $k = 6$. For the considered scenario, we evaluate the MSE achievable by the task-based quanziation system of Fig. 8 where the analog combiner, quantization support, and digital processing are obtained via Theorem 3. The task-based quantizer is compared to recovering the empirical covariance in analog, as well as to directly quantizing $\boldsymbol{x}$, i.e., a task-ignorant scalar quantizer, and a hybrid system utilizing linear analog combiners based on [8, Sec. V]. For all the above systems, in order to avoid overloading the quantizers, the support is set to η times the maximal sum of the standard deviation and absolute mean value of the entries of the input to the ADC, where we let η increase linearly with the number of bits in the range $[3, 6.5]$. The achievable MSEs versus the number of bits are depicted in Fig. 9. Observing Fig. 9, we note that the task-based quantizer, which is designed to balance the quantization and estimation errors, achieves the best MSE performance. Quantizing $\tilde{\boldsymbol{s}}$ directly results in notable quantization errors when operating with a small number of bits, due to the need to set the support to a relatively large value resulting in coarse quantization. This demonstrates how the task-based quantization design proposed in Section 3.2 for linear tasks can be extended to apply for recovering non-linear functions.

4. Deep task-based quantization

In Section 3 we designed hybrid analog-digital acquisition systems, which consist of analog combining, scalar quantization, and digital processing, to

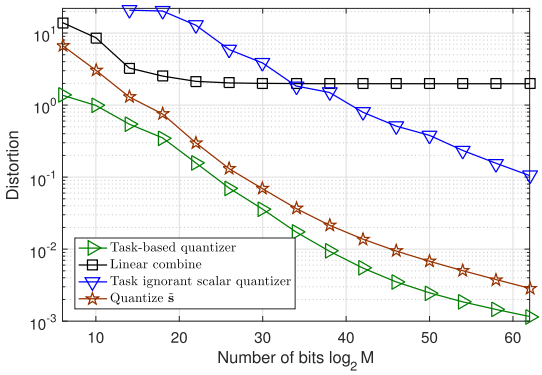

Figure 9: Empirical covariance recovery.

accurately recover some underlying information embedded in the observed analog signal. The systems proposed in Section 3 are model-aware, requiring accurate knowledge of the statistical relationship between the observations and the task, i.e., $f_{\boldsymbol{x}|\boldsymbol{s}}$. Two notable challenges are associated with such model-aware designs: 1) Accurate knowledge of the statistical model $f_{\boldsymbol{x}|\boldsymbol{s}}$ may be unavailable in practice; 2) Even when $f_{\boldsymbol{x}|\boldsymbol{s}}$ is perfectly known, analytically tractable characterizations are obtained only for tasks of relatively simple form, e.g., linear and quadratic functions, under the model assumptions *A1*–*A2*. This limits the design to estimation tasks *A1*, does not explore arbitrary quantization rules *A2*, and may not lead to analytically tractable systems when operating under complex statistical relationships.

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 utilizing ML methods, one can implement task-based quantizers without the need to explicitly know the underlying model and to analytically derive the proper quantization rule. Furthermore, when the parameters of the hybrid analog-digital system are learned from data and not specified analytically, the quantization mapping can be optimized along with the system parameters instead of fixing a uniform rule as in (1). Finally, additional families of tasks, such as classification, can be considered by properly setting the loss function utilized in the learning process.

In this section we present a generic DNNs architecture which utilizes ML for task-based quantization with scalar ADCs, referred to as *deep task-based quantization* [11]. We begin with the system architecture in Section 4.1, after which we present how the quantization mapping is learned in Section 4.2. We provide numerical results along with a discussion in Section 4.3.

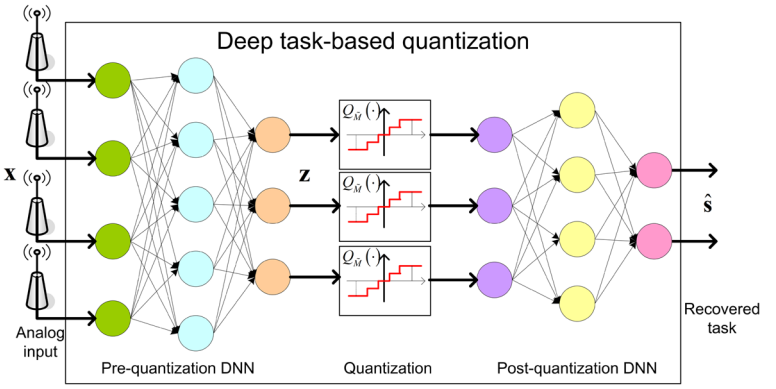

Figure 10: Deep task-based quantization system architecture.

4.1. System architecture

Deep task-based quantization operates in a data-driven manner, learning the analog transformation, quantization mapping, and digital processing, from a training data set, consisting of t independent realizations of $\boldsymbol{s}$ and $\boldsymbol{x}$, denoted $\{\boldsymbol{s}^{(i)}, \boldsymbol{x}^{(i)}\}_{i=1}^{t}$. In general, the training samples may be taken from a set of joint distributions, and not only from the true (unknown) joint distribution of $\boldsymbol{s}$ and $\boldsymbol{x}$. Here, the analog pre-quantization mapping and the digital post-quantization processing are parameterized as layers of a DNN, as illustrated in Fig. 10. By doing so, the overall task-based quantization system, including the analog combining, quantization rule, and digital processing, can be trained from data in an end-to-end manner using e.g., stochastic gradient descent (SGD). While the proposed system focuses only on the quantization aspect of ADCs, the resulting design approach can be extended to account also for sampling in addition to quantization, as considered in [26].

In the proposed architecture, the scalar ADC, which implements the continuous-to-discrete mapping, is modeled as an activation function between two intermediate layers, interfacing the analog processing and the digital part. 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. While the digital DNN can be implemented in software, the analog

DNN requires dedicated configurable hardware. Such analog networks can be implemented using neuromorphic electronic systems [32], or alternatively, the analog processing can be constrained to represent the possible configurations of the circuitry connecting the analog inputs and the ADCs, as we discuss in Section 5 in the context of MIMO receivers. The system input is the observed $\boldsymbol{x}$, and $\boldsymbol{\theta}$ denotes the network parameters. Two families of tasks are considered:

- **Estimation**: Here, the system 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 $\boldsymbol{s}$. The loss function is the empirical MSE:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{t}\sum_{j=1}^{t} \left\| \boldsymbol{s}^{(j)} - \psi_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(j)}\right) \right\|_2^2. \tag{6}$$

- **Classification**: In such tasks, the system should decide between a finite number of options. 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:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{t}\sum_{j=1}^{t} -\log \psi_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(j)};\boldsymbol{s}^{(j)}\right). \tag{7}$$

4.2. Learned quantization mappings

The proposed architecture implements scalar quantization as an intermediate activation in a joint analog-digital hybrid DNN. The overall network is trained in an end-to-end manner using some variant of SGD with backpropagation to minimize the loss function $\mathcal{L}(\cdot)$. Such end-to-end training can be either carried out offline on a separate machine, applying the learned weights to tune the analog hardware, or alternatively, the system itself can tune its parameters, given direct access to the analog signals and their corresponding labels during training. The quantization layer, which interfaces

the analog and digital domains, converts its continuous-amplitude input into a discrete quantity. The non-differentiable nature of such continuous-to-discrete mappings induces a challenge in applying SGD for optimizing the network parameters. 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. Thus, straight-forward application of SGD with back-propagation fails to properly set the pre-quantization network.

This challenge can be tackled by approximating the non-differentiable quantization mapping by a differentiable one, as proposed in [33]. This is achieved by replacing the continuous-to-discrete transformation with a non-linear activation function which has approximately the same behavior as the quantizer. 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:

$$\tilde{q}(z) = \sum_{i=1}^{\tilde{M}-1} a_i \tanh\left(c_i \cdot z - b_i\right), \tag{8}$$

where $\{a_i, b_i, c_i\}$ are real-valued parameters. When the parameters $\{c_i\}$ increase, the corresponding hyperbolic tangents approach step functions.

In addition to learning the weights of the analog and digital DNNs, this approach allows to learn the quantization function, and particularly, the best suitable constants $\{a_i\}$ and $\{b_i\}$. These tunable parameters are later used to determine the decision regions of the scalar quantizer, where the set $\{b_i\}$ is used for the decision regions limits while $\{a_i\}$ determines the corresponding discrete values assigned to each decision region. The parameters $\{c_i\}$, which essentially control the resemblance of (8) to an actual continuous-to-discrete mapping, do not reflect on the quantization rule, and are thus not learned from 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 (8), one can now apply standard SGD with back-propagation 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, the learned $\tilde{q}(\cdot)$ activation (8) is replaced with a scalar quantization mapping dictated by the tunable parameters $\{a_i, b_i\}$. An illustration of how the differentiable mapping (8) is converted into a continuous-to-discrete quantization rule is depicted in Fig. 11. The dashed smooth curve in Fig. 11 represents the differentiable function after training is concluded, and the straight curve is the resulting scalar quantizer.

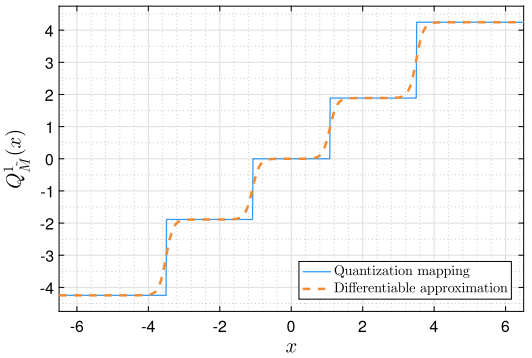

Figure 11: Differentiable approximation of the quantization rule illustration.

4.3. Numerical results

We next numerically demonstrate the achievable performance of deep task-based quantization. In the following, we model the relationship between the observed $\boldsymbol{x}$ and the task $\boldsymbol{s}$ as

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{w}, \tag{9}$$

for some fixed $\boldsymbol{H} \in \mathcal{R}^{n\times k}$, where $\boldsymbol{w} \in \mathcal{R}^n$ is a zero-mean Gaussian vector with i.i.d. entries of variance $\sigma_w^2 > 0$.

We begin with an estimation task for which we can compare the data-driven task-based system to its model-aware counterpart detailed in Section 3.2. Here, we set $\sigma_w^2 = 0.25$, $n = 120$, $k = 40$, while $\boldsymbol{s}$ is a zero-mean Gaussian vector with i.i.d. unit variance entries. The matrix $\boldsymbol{H}$ is set to

$$\boldsymbol{H} = \begin{bmatrix} \mathrm{Re}\big(\boldsymbol{\Phi} \otimes \boldsymbol{I}_{10}\big) & \mathrm{Im}\big(\boldsymbol{\Phi} \otimes \boldsymbol{I}_{10}\big) \\ -\mathrm{Im}\big(\boldsymbol{\Phi} \otimes \boldsymbol{I}_{10}\big) & \mathrm{Re}\big(\boldsymbol{\Phi} \otimes \boldsymbol{I}_{10}\big) \end{bmatrix},$$

where $\boldsymbol{\Phi}$ is the first 4 columns of the 12×12 DFT matrix. This setting represents channel estimation in Rayleigh fading MIMO channels using orthogonal pilots [11, Sec. IV]. In Fig. 12 we numerically evaluate the average MSE versus the quantization rate R of deep task-based quantization compared to the fundamental performance limit dictated by indirect rate-distortion theory, as well as to the performance of the model-aware task-based quantizer discussed in Section 3. To guarantee fair comparison with the model-aware system we set the pre and post quantization DNNs to consist of linear layers. Following [8, Prop. 2], we set the number of scalar quanizers to $p = k$ for

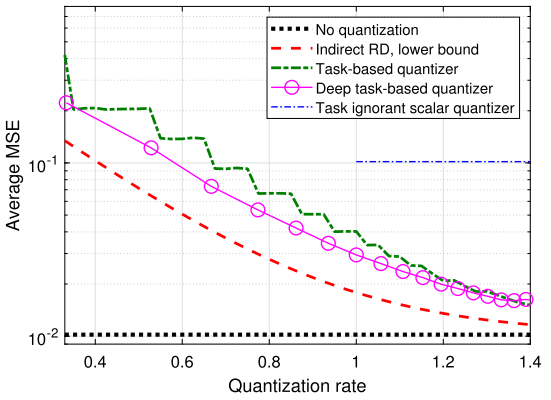

Figure 12: Estimation task.

both task-based quantizers. The data-driven system is trained using $t = 2^{15}$ labeled pairs, and all systems are tested using 2^{10} test samples. We also depict in Fig. 12 the average MSE of a task-ignorant system in which estimation is carried out only in the digital domain, using the method for channel estimation from quantized measurements proposed in [5].

Observing Fig. 12, we note that the fact that data-driven quantizer is not restricted to uniform quantizers allows it to outperform the model-aware system of Section 3 especially in lower quantization rates. Furthermore, the performance of both task-based quantizers is within a relatively small gap of the fundamental performance limits. These results demonstrate the ability of deep task-based quantization to implement a feasible and optimal-approaching quantization system in a data-driven fashion.

Next, we consider a classification task, where the objective is to minimize the bit error rate (BER) in recovering symbols generated from a discrete constellation. For such tasks the model-aware system of Section 3 is not applicable as Assumption *A1* does not hold. Again, the observations $\boldsymbol{x}$ are related to the task vector $\boldsymbol{s}$ via (9). However, here the entries of $\boldsymbol{s}$ are i.i.d. uniformly distributed over $\mathcal{S} = \{-1, 1\}$ representing, e.g., symbol detection in MIMO communications. In particular, we use $n = 12$, $k = 4$, and set the entries of $\boldsymbol{H}$ to $(\boldsymbol{H})_{i,j} = e^{-|i-j|}$. For the deep task-based quantizer we use two layers in analog and two layers in digital. The output layer is a softmax function with 2^k probabilities, and the overall network is trained to minimize the cross-entropy loss (7) using $t = 5000$ labeled samples. Unlike the estimation task for which the number of quantizers p can be set according to the analytical results in [8], here this value was determined based on

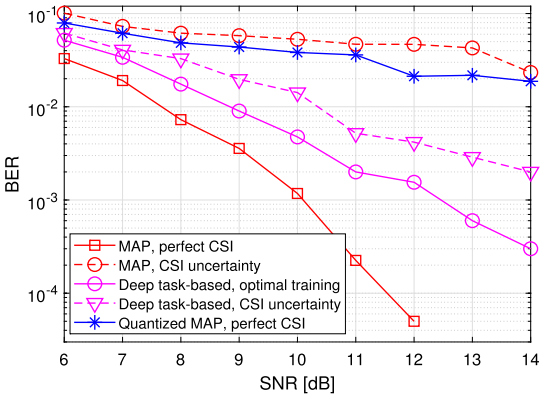

Figure 13: Classification task.

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.

The numerically computed BER averaged over 20000 trials versus the signal-to-noise ratio (SNR) defined as $1/\sigma_w^2$ of the deep task-based quantizer with quantization rate $R = 1$ is depicted in Fig. 13 compared to the maximum a-posteriori probability (MAP) rule operating for recovering $\boldsymbol{s}$ from $\boldsymbol{x}$, i.e., without quantization constraints, as well as the MAP rule for recovering $\boldsymbol{s}$ from a uniformly quantized $\boldsymbol{x}$ with rate $R = 1$, representing a task-ignorant digital only system. It is noted that the MAP detectors require prior knowledge of $\boldsymbol{H}$ or σ_w^2, while the data-driven quantizer is invariant of the underlying model and learns its mapping from training. In order to study the resiliency of deep task-based quantization to inaccurate training, we also compute the BER under channel state information (CSI) uncertainty, namely, when the training samples are randomized from a joint distribution of $\boldsymbol{s}, \boldsymbol{x}$ in which the entries of the matrix $\boldsymbol{H}$ in (9) 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 with the same level of CSI uncertainty.

Observing Fig. 13, 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. For comparison, the quantized MAP rule, which operates only in the digital domain, achieves significantly worse BER performance compared to the hybrid deep task-based quatizer, demonstrating the benefit of applying pre-quantization processing in the analog domain in order to utilize more accurate quantization while

keeping the semantic information required to carry out the task. The results in Fig. 13 also demonstrate the improved robustness of the data-driven system 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 performance when trained using accurate CSI. This demonstrates the gains of using DNNs for overcoming the sensitivity of model-based approaches to inaccurate model knowledge.

5. Hardware implementation for MIMO receivers

In the previous sections we presented the concept of task-based quantization, in which the components of a hybrid analog-digital system are jointly optimized to facilitate the recovery of some underlying information under bit constraints. We considered two complementary strategies for tuning task-based quantizers: a model-aware approach and a data-driven method. Here, we discuss how the systems designed using either of the aforementioned strategies can be realized, as well as which additional practical considerations must be taken into account and how they can be incorporated in the design. We focus here on task-based quantization for MIMO receivers, in which multiple signals are acquired for some task other than recovering them in digital, and where quantization constraints play an important role.

Conventional MIMO receivers obtain their observations using a set of antennas, where each antenna is connected to a dedicated scalar ADC, typically implementing a uniform quantization mapping. Consequently, the main challenge in realizing hybrid task-based quantizers for MIMO receivers stems from the need to introduce additional processing in analog prior to quantization. Furthermore, this analog combining is required to be dynamically configurable, allowing it to be adapted when operating in dynamic environments. In the following we elaborate on two strategies for implementing such hybrid MIMO receivers: First, in Section 5.1 we discuss hybrid receivers with dedicated analog combining hardware. Then, we present how the emerging technology of dynamic metasurface antennas (DMAs) can be exploited to introduce controllable analog combining in Section 5.2.

5.1. Dedicated analog combiner hardware

A common strategy to implement MIMO receivers, particularly when equipped with a large number of antennas and when operating in high spectral bands, is to introduce dedicated analog circuitry between the antennas

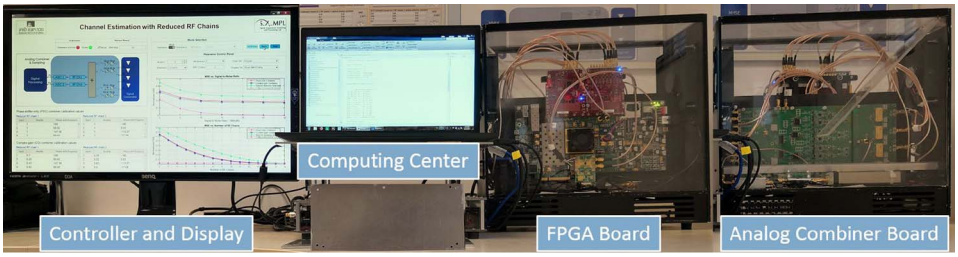

Figure 14: Analog combiner prototype demonstration setup.

and the ADCs. The original motivation for implementing such hybrid receivers is to reduce the number of costly RF chains, namely, the main purpose of the analog combiner is to reduce the dimensionality of the acquired signals allowing the receiver to operate with less RF chains than antennas [15, 16]. The typical implementation of such analog combiners is based on an inter-connection of phase shifters and adders, either connecting a controllable phase-shifted version of the signal observed at each antenna to each ADC, resulting in a *fully-connected phase shifter network*, or alternatively, by dividing the antennas into subsets, each phase shifted and connected to a distinct ADC, referred to as *partially-connected phase shifter network* [16].

The resulting system model of a hybrid MIMO receiver thus includes an additional linear processing prior to acquisition, similarly to the model used in our derivation in Section 3.2, and can thus be exploited for realizing task-based quantization. In particular, for a hybrid receiver with a fully-connected phase shifter network, the resulting matrix $\boldsymbol{A}$ in Section 3.2 is subject to an additional constraint which stems from the usage of adjustable phase shifters, that only the phase of its entries can be configured, i.e., $|(\boldsymbol{A})_{i,j}| = 1$ for each $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, n\}$. This constraint can accounted for by identifying the unconstrained analog combining matrix via, e.g., Theorem 3, and projecting it to the feasible set of fully-connected phase shifter networks, similarly to [16, Alg. 2]. Alternatively, when using a data-driven design as proposed in Section 4, one can account for the additional design constraints by letting the trainable parameters of the analog network to be the phases of the entries of the matrix $\boldsymbol{A}$.

The difficulties associated with using phase shifter networks as analog combiners for task-based quantization can be mitigated by introducing adjustable gains into the analog circuitry. For example, the prototype proposed in [17], depicted in Fig. 14, implements a complex-gain analog combiner operating in the sub-6 GHz band using digitally controllable vector multipliers. A controllable gain analog combiner operating in the 25–30 GHz

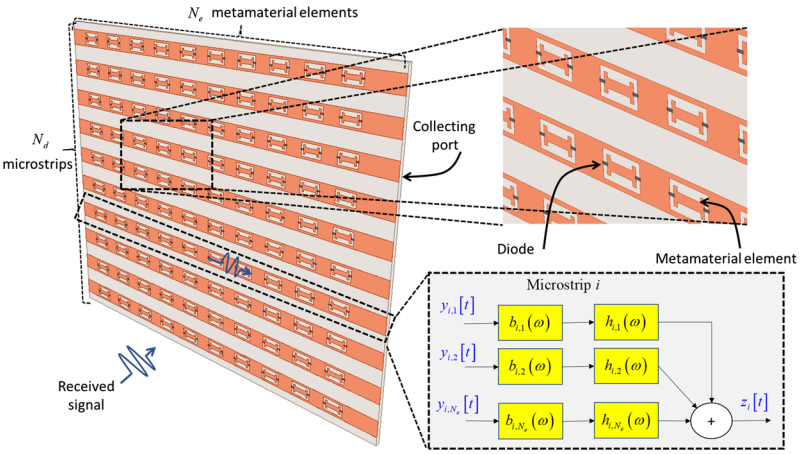

Figure 15: DMA system model illustration.

band based on RF integrated circuits was proposed in [34]. The resulting model of a hybrid receiver equipped with such analog combiners effectively allows to control both the gain and phase of each entry of the matrix $\boldsymbol{A}$ individually in run-time, thus allowing to implement the task-based quantization systems proposed in the previous sections. The main drawback of such implementations compared to phase shifter networks is the cost and complexity associated with controllable complex gain analog circuits.

5.2. Analog combining via dynamic metasurface antennas

The analog combiners discussed in the previous section require the MIMO receiver to be equipped with a dedicated analog combining hardware interfacing its antenna elements and the ADCs. An alternative strategy to achieve configurable analog combining without additional dedicated circuitry implements the pre-quantization processing as part of the antenna architecture, by using DMAs. The conventional gains of DMAs over standard antenna arrays stem from the fact metasurfaces typically use much less power and cost less [35], while facilitating the implementation of a large number of elements in a given physical area. An additional gain of DMAs is their ability to implement tunable combining as an inherent byproduct of their architecture.

In particular, DMAs consist of a set of microstrips, each embedded with configurable radiating metamaterial elements [36]. When used as a receive antenna, the signals observed by the elements are captured at a single output port for each microstrip, feeding an ADC. The relationship between these

signals and the micropstip output is dictated by two main properties: 1) Each element of index l of microstrip i acts as resonant electrical circuit, whose frequency response is described by the Lorentzian form [36]

$$b_{i,l}(\omega) = \frac{F_{i,l}\omega^2}{(\omega_{i,l}^R)^2 - \omega^2 - j\omega\chi_{i,l}}, \tag{10}$$

where $F_{i,l}$, $\chi_{i,l}$, and $\omega_{i,l}^R$ are the oscillator strength, damping factor, and angular resonance frequency, respectively, which are all externally configurable parameters. 2) Each signal which propagates from an element to the output port undergoes a different path, and thus accumulates a different delay. The delay accumulated by the signal captured at the lth element of the ith micropstrip can be modeled as a filter with frequency response $h_{i,l}(\omega)$. The signal observed at the output port of the ith micrtopstrip can thus be written as the sum of outputs of the filters $b_{i,l}(\omega)h_{i,l}(\omega)$ whose inputs are the signals observed by the corresponding elements, as illustrated in Fig. 15.

The resulting model relating the observed signals and the DMA output ports, which are the signals fed to the ADCs, represents a form of *frequency-selective analog combining*. Specifically, the fact that the parameters of the Lorentzian response in (10) can be modified element-wise, indicates that the inherent processing carried out inside each micropstrip can be tuned to facilitate acquisition under bit constraints by tuning the resulting combining as part of a task-based quantizer, see, e.g., [19]. Consequently, when using a MIMO receiver with a DMA-based antenna array, one can implement a form of task-based quantization without requiring additional dedicated analog combining hardware by properly tuning the frequency response of each element along with the quantization mapping and the digital processing utilizing either of the methods discussed in Sections 3–4.

The architectures detailed in this section can all be used to realize task-based quantization in MIMO receivers, by exploiting either the model-aware design guidelines proposed in Section 3, or alternatively, by learning the task-based quantization mapping from labeled data as suggested in Section 4. Combining the architectures detailed in this section with the design methods proposed in the previous sections thus narrows the gap between the theory of task-based quantization and its concrete implementation in MIMO receivers.

6. Conclusion

In this paper we reviewed the theory and design methods for task-based quantization systems. Such systems carry out acquisition using simple bit-limited scalar ADCs. The associated distortion is mitigated by accounting

for the system task in acquisition, via jointly optimizing some level of analog pre-processing along with the quantization rule and digital post-processing in light of the system task. We first presented model-aware design methods which infer the operation of the system components based on prior knowledge of the statistical model relating the observations and the information of interest to be recovered in digital. We then proposed an alternative design approach which does not require knowledge of the underlying model, and learns its task-based quantization mapping from a set of labeled samples using ML tools. Finally, we presented several hardware architectures which can facilitate the implementation of task-based quantization mechanisms in MIMO receivers. The combined results detailed in this survey pave the way to the realization of MIMO receivers operating accurately and efficiently under strict bit constraints by using task-based quantization techniques.

References

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, p. 3590, 2010.

[2] N. Shlezinger and Y. C. Eldar, "On the spectral efficiency of noncooperative uplink massive MIMO systems," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1956–1971, 2019.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[4] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, 1999.

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

[6] H. Pirzadeh and A. L. Swindlehurst, "Spectral efficiency of mixed-ADC massive MIMO," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3599–3613, 2018. MR3832393

[7] S. Khobahi, N. Naimipour, M. Soltanalian, and Y. C. Eldar, "Deep signal recovery with one-bit quantization," in *Proc. IEEE ICASSP*, 2019.

[8] N. Shlezinger, Y. C. Eldar, and M. R. Rodrigues, "Hardware-limited task-based quantization," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5223–5238, 2019. MR4016282

[9] N. Shlezinger, Y. C. Eldar, and M. R. Rodrigues, "Asymptotic task-based quantization with application to massive MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 3995–4012, 2019. MR3993401

[10] S. Salamatian, N. Shlezinger, Y. C. Eldar, and M. Médard, "Task-based quantization for recovering quadratic functions using principal inertia components," in *Proc. IEEE ISIT*, 2019. MR3683549

[11] N. Shlezinger and Y. C. Eldar, "Deep task-based quantization," *arXiv preprint arXiv:1908.06845*, 2019.

[12] F. P. Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5011–5038, 2017. MR3683549

[13] M. R. Rodrigues, N. Deligiannis, L. Lai, and Y. C. Eldar, "Rate-distortion trade-offs in acquisition of signal parameters," in *Proc. IEEE ICASSP*, 2017, pp. 6105–6109.

[14] K. Yu, Y. D. Zhang, M. Bao, Y.-H. Hu, and Z. Wang, "DOA estimation from one-bit compressed array data via joint sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1279–1283, 2016.

[15] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.

[16] S. S. Ioushua and Y. C. Eldar, "A family of hybrid analog–digital beamforming methods for massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3243–3257, 2019. MR3968421

[17] T. Gong, N. Shlezinger, S. S. Ioushua, M. Namer, Z. Yang, and Y. C. Eldar, "RF chain reduction for MIMO systems: A hardware prototype," *IEEE Syst. J.*, 2020.

[18] N. Shlezinger, O. Dicker, Y. C. Eldar, I. Yoo, M. F. Imani, and D. R. Smith, "Dynamic metasurface antennas for uplink massive MIMO systems," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6829–6843, 2019.

[19] H. Wang, N. Shlezinger, Y. C. Eldar, S. Jin, M. F. Imani, I. Yoo, and D. R. Smith, "Dynamic metasurface antennas for MIMO-OFDM receivers with bit-limited ADCs," *arXiv preprint arXiv:1912.06917*, 2019.

[20] N. Shlezinger, G. C. Alexandropoulos, M. F. Imani, Y. C. Eldar, and D. R. Smith, "Dynamic metasurface antennas for 6G extreme massive MIMO communications," *arXiv preprint arXiv:2006.07838*, 2020.

[21] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998. MR1658787

[22] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012. MR2239987

[23] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. 26, no. 5, pp. 518–521, 1980. MR0583937

[24] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. 16, no. 4, pp. 406–411, 1970. MR0272503

[25] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," 2015.

[26] N. Shlezinger, R. J. G. van Sloun, I. A. M. Hujiben, G. Tsintsadze, and Y. C. Eldar, "Learning task-based analog-to-digital conversion for MIMO receivers," in *Proc. IEEE ICASSP*, 2020.

[27] N. Farsad, N. Shlezinger, A. J. Goldsmith, and Y. C. Eldar, "Data-driven symbol detection via model-based machine learning," *arXiv preprint arXiv:2002.07806*, 2020.

[28] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[29] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical theory of quantization," *IEEE Trans. Instrum. Meas.*, vol. 45, no. 2, pp. 353–361, 1996.

[30] V. Kostina and S. Verdú, "Nonasymptotic noisy lossy source coding," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6111–6123, 2016. MR3565102

[31] W. C. Jakes and D. C. Cox, *Microwave mobile communications*. Wiley-IEEE Press, 1994.

[32] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[33] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 1141–1151.

[34] S. Mondal, R. Singh, A. I. Hussein, and J. Paramesh, "A 25–30 GHz fully-connected hybrid beamforming receiver for MIMO communication," *IEEE J. Solid-State Circuits*, vol. 53, no. 5, pp. 1275–1287, 2018.

[35] M. C. Johnson, S. L. Brunton, N. B. Kundtz, and J. N. Kutz, "Sidelobe canceling for reconfigurable holographic metamaterial antennas," *IEEE Trans. Antennas Propag.*, vol. 63, no. 4, pp. 1881–1886, Apr. 2015. MR3340892

[36] D. R. Smith, O. Yurduseven, L. Pulido-Mancera, P. Bowen, and N. B. Kundtz, "Analysis of a waveguide-fed metasurface antenna," *Phys. Rev. Applied*, vol. 8, no. 5, Nov. 2017.

Nir Shlezinger
School of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Be'er-Sheva
Israel
E-mail address: nirshl@bgu.ac.il

Yonina C. Eldar
Faculty of Mathematics and Computer Science
Weizmann Institute of Science
Rehovot
Israel
E-mail address: yonina@weizmann.ac.il

Received February 7, 2020