Design and Analysis of Hardware-limited Non-uniform Task-based Quantizers

Neil Irwin Bernardo, *Graduate Student Member, IEEE*, Jingge Zhu, *Member, IEEE*, Yonina C. Eldar, *Fellow, IEEE*, and Jamie Evans, *Senior Member, IEEE*

Abstract—**Hardware-limited task-based quantization is a new design paradigm for data acquisition systems equipped with serial scalar analog-to-digital converters using a small number of bits. By taking into account the underlying system task, task-based quantizers can efficiently recover the desired parameters from the low-bit quantized observation. Current design and analysis frameworks for hardware-limited task-based quantization are only applicable to inputs with bounded support and uniform quantizers with non-subtractive dithering. Here, we propose a new framework based on generalized Bussgang decomposition that enables the design and analysis of hardware-limited task-based quantizers that are equipped with non-uniform scalar quantizers or that have inputs with unbounded support. We first consider the scenario in which the task is linear. Under this scenario, we derive new pre-quantization and post-quantization linear mappings for task-based quantizers with mean squared error (MSE) that closely matches the theoretical MSE. Next, we extend the proposed analysis framework to quadratic tasks. We demonstrate that our derived analytical expression for the MSE accurately predicts the performance of task-based quantizers with quadratic tasks.**

Index Terms—**Quantization, Analog-to-digital conversion**

I. INTRODUCTION

DIGITAL systems are equipped with quantizers to facilitate the processing, storage, and communication of information embedded in continuous-amplitude samples. In principle, the most accurate digital representation of a sampled signal is obtained by jointly mapping the samples to the digital domain via vector quantization [1], [2]. The optimal trade-off between compression and fidelity is fundamentally described by rate-distortion theory [3]. However, in practice, the quantization process is performed by analog-to-digital converters (ADC) which typically operate in a serial scalar manner [4]. Under this setup, the incoming continuous-time analog signal is first sampled and the samples are sequentially mapped by the quantizer in digital form using a finite number of quantization bits [5]. A linear increase in the number of quantization bits corresponds to an exponential increase in power consumption [6]. Therefore, there is growing interest in the use of low-resolution data converters. For instance, recent

N.I. Bernardo is with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia and also with the Electrical and Electronics Engineering Institute, University of the Philippines Diliman, Quezon City 1101, Philippines (e-mail: bernardon@student.unimelb.edu.au).

J. Zhu and J. Evans are with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia (e-mail: jingge.zhu@unimelb.edu.au; jse@unimelb.edu.au).

Y. C. Eldar is with the Faculty of Math and CS, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

works on low-power communication receivers, such as [7]–[12], have focused on investigating the performance limits of low-resolution receiver architectures and designing novel methods that enable various receiver functionalities (detection, channel estimation, and synchronization) to work in the low-resolution regime.

Data acquisition systems are often designed such that the input and output of the quantizers are close with respect to some pre-defined distortion measure [13, Ch. 10][14]. This design approach, however, does not take into account the underlying system task. In several signal processing and communication applications, the objective is not to faithfully recover the input signal, but rather to extract some low-dimensional parameters/features embedded in the quantized measurements. Such systems that take into account the underlying task are generally referred to as *task-based quantization*, and task-based quantization systems equipped with serial scalar ADCs are specifically referred to as *hardware-limited task-based quantization* [15].

Previous works [15]–[18] have shown that, by exploiting the *a priori* knowledge regarding the system task, hardware-limited task-based quantizers can outperform digital systems that simply extract the desired parameters from the quantized measurements. Performance gain in task-based quantization is achieved by employing a hybrid analog/digital (A/D) architecture and jointly designing the analog pre-quantization mapping and digital post-quantization mapping in light of the underlying system task. The task-based quantization framework has been applied in various tasks such as channel estimation [15], empirical covariance estimation [18], multiple-input multiple-output (MIMO) radar receivers [19], task-specific beamforming [20], MIMO communication [21]–[23], symbol detection [24], [25], and graph signal compression [26]. Moreover, the combined effect of sampling and quantization in hardware-limited task-based systems has been studied in [27], [28] and it is shown that the optimal performance of bitrate-constrained data acquisition systems is generally achieved by sampling below the Nyquist rate. The optimal sampling and quantization scheme for task-based data acquisition may also be learned via data-driven approaches if the input distribution is not known [29].

Despite the aforementioned benefits and the wide range of applications of task-based quantization, the existing framework for analyzing hardware-limited task-based quantization is only applicable to scalar uniform ADCs with non-subtractive dithering. While dithering offers analytical tractability, dithered quantizers generally have subpar performance compared to
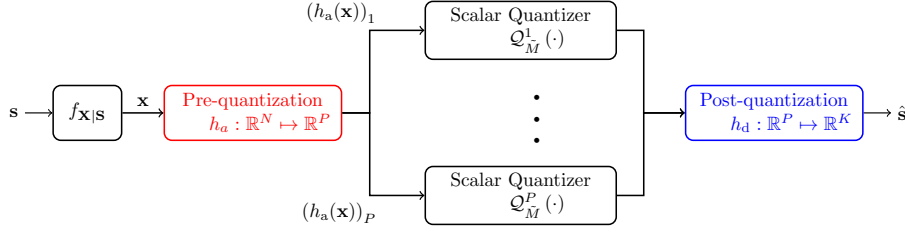
Figure 1: System model for task-based quantization.

their non-dithered counterpart. Also, the theory of nonsubtractive dithering [30] only applies to uniform quantizers. Mathematical tools [31] for analyzing dithered non-uniform quantization exist but only for subtractive dithering, which is often impractical. As such, the current framework does not fully capture the actual performance of task-based quantization. Simulation results of [15] depict large performance gaps between the dithered and non-dithered case when the number of quantization levels per scalar quantizer is low. Furthermore, the analysis framework relies on the assumption that the overload probability (i.e. the probability that the ADC input does not exceed the specified dynamic range of the ADC) is zero. This assumption can be quite restrictive so that the analysis framework only holds approximately for input signals with infinite support. Guidelines on how to set the overload probability for a given number of quantization levels are provided in [27]. Still, the simulated distortion of dithered task-based quantization is approximately 5% higher than what the analytical expression predicts in the numerical results.

Here, we provide a new approach to design and analyze hardware-limited task-based quantization systems with analog pre-quantization and digital post-quantization linear mappings based on generalized Bussgang decomposition [32]. In contrast to the state-of-the-art (SOTA) analysis framework [15], the proposed framework does not rely on the zero overload probability assumption and is also applicable to non-uniform scalar quantizers and non-dithered settings. Our proposed framework restricts the pre-quantization mapping to be within the class of linear mappings that make the inputs of the scalar quantizers uncorrelated. While this restriction may lead to suboptimal performance, our numerical results show that the proposed framework can achieve lower distortion than previous results when the quantization budget is limited. More importantly, a crucial advantage of our analysis is that the simulated distortion of task-based quantizers designed using our method fits well with the predictions of our theoretical framework, even if the underlying system task is nonlinear. This is in contrast to previous results which only hold approximately. The main contributions of our work are the following:

- We provide descriptions of the analog and digital linear mappings of task-based quantization under a linear task assumption (i.e. the task is a linear function of the observations). The derived linear mappings are conceptually different from the linear mappings in previous works. We present numerical results showing that, in some cases, task-based quantizers designed using our approach can outperform task-based quantizers designed using the

SOTA analysis framework [15].
- We show that the actual mean squared error (MSE) of the task-based system under the derived analog and digital linear mappings fits the theoretical MSE in contrast to previous results. Moreover, the proposed analysis framework also enables a model-based analysis of task-based quantization with non-uniform quantizers. To the best of our knowledge, there is no framework in the literature that facilitates model-based analysis of task-based quantization with non-uniform quantizers.
- We show how to extend the proposed framework to non-linear tasks. More specifically, we consider the quadratic task problem of empirical covariance estimation and show that the task-based quantization system designed using our proposed framework achieves lower MSE than the simulated MSE of the task-based system designed using the framework presented in [18].

The rest of the paper is organized as follows: Section II formulates the system model and states the model assumptions for the linear task scenario. Section III presents the new analysis framework. Section IV provides numerical results and analysis for the proposed framework in Section III. Section V extends the developed framework to quadratic tasks. Finally, Section VI concludes the paper.

II. PROBLEM FORMULATION AND ANALYSIS TOOLS

A. Problem Setup and Model Assumptions

The system model of the task-based quantization with hardware constraints is illustrated in Figure 1. The *task vector* $\mathbf{s} \in \mathbb{R}^{K \times 1}$ contains the parameters we aim to recover. However, the input to the task-based quantizer is not $\mathbf{s}$ but the *measurement vector* $\mathbf{x} \in \mathbb{R}^{N \times 1}$. The statistical relationship between $\mathbf{s}$ and $\mathbf{x}$ is described by the conditional probability $f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s})$. With slight abuse of notation, we simply write the conditional probability as $f_{\mathbf{X}|\mathbf{S}}$. We also assume that $\mathbf{s}$ and $\mathbf{x}$ are both zero-mean random vectors and have covariance matrices given by $\mathbf{\Sigma_s}$ and $\mathbf{\Sigma_x}$, respectively.

The measurement vector $\mathbf{x}$ is projected to $\mathbb{R}^{P \times 1}$, where $P \leq N$, using an analog pre-quantization mapping, denoted h_{a}. The P outputs of the analog pre-quantization mapping are fed to P scalar quantizers. From [15, Corollary 1], the optimal choice of P must not exceed K. Each scalar quantizer has $\tilde{M} = \lfloor M^{\frac{1}{P}} \rfloor$ number of quantization levels, where M is a constraint on the overall number of quantization levels. As pointed out in [15], the parameter $\tilde{M}$ is directly related to the power consumption of an ADC. We allow the quantization
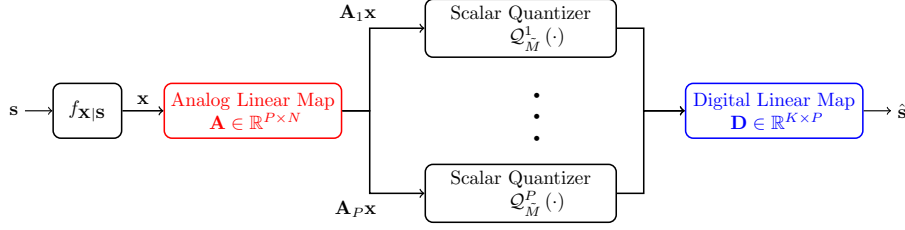
Figure 2: Task-based quantization with analog combining matrix and digital processing matrix.

levels of the scalar quantizers to have non-uniform and non-identical structure. In this work, we assume that the scalar quantizers are designed using the Lloyd-Max algorithm [33]. Note that quantizers designed using the Lloyd-Max algorithm satisfy $\mathbf{E}\{X_{\text{in}}|\mathcal{Q}(X_{\text{in}})\} = \mathcal{Q}(X_{\text{in}})$, where X_{in} is the input to the quantizer (i.e. each quantization interval is represented by its mean value). This property of the scalar quantizers is crucial in the derivation of our main results. Finally, the P outputs of the scalar quantizers, denoted $\mathbf{z} \in \mathbb{R}^{P \times 1}$, are fed to a digital post-processing function to estimate the task vector $\mathbf{s}$. We represent the estimate of the task vector as $\hat{\mathbf{s}} \in \mathbb{R}^{K \times 1}$.

The goal is to recover $\mathbf{s}$ from quantized measurements $\mathbf{z}$. This problem setup is referred in the information theory community as the *indirect source coding problem*[1], and was first introduced in [34]. We design h_{a} and h_{d} such that the MSE between $\mathbf{s}$ and $\hat{\mathbf{s}}$ is minimized. Mathematically, we have the following optimization problem:

$$\min_{h_{\text{a}}, h_{\text{d}}} \mathbb{E}\left\{||\mathbf{s} - \hat{\mathbf{s}}||^2\right\}$$
$$= \mathbb{E}\left\{||\mathbf{s} - \tilde{\mathbf{s}}||^2\right\} + \min_{h_{\text{a}}, h_{\text{d}}} \mathbb{E}\left\{||\tilde{\mathbf{s}} - \hat{\mathbf{s}}||^2\right\}, \quad (1)$$

where $\tilde{\mathbf{s}} = \mathbb{E}\{\mathbf{s}|\mathbf{x}\}$ is the minimum MSE (MMSE) estimator of $\mathbf{s}$ given the measurement vector $\mathbf{x}$. The RHS of (1) shows that the MSE can be written as a sum of two terms. The first term quantifies the minimum estimation error of $\mathbf{s}$ from $\mathbf{x}$ whereas the second term accounts for the minimal distortion in quantizing the MMSE estimate. The first term is independent on the actual structure of the scalar quantizers and design of the pre- and post-quantization mappings [15]. Thus, we can focus our attention on minimizing the second MSE term. We shall refer to this MSE term as the *quantizer-dependent MSE*.

To facilitate recovery of the task vector $\mathbf{s}$ under practical hardware setting, we follow the approach of [15] which is to impose h_{a} and h_{d} to be linear mappings, as shown in Figure 2. That is, we introduce an analog combining matrix $\mathbf{A} \in \mathbb{R}^{P \times N}$ and a digital processing matrix $\mathbf{D} \in \mathbb{R}^{K \times P}$ to operate as h_{a} and h_{d}, respectively. The quantities $\{\mathbf{A}_p \mathbf{x}\}_{p=1}^P$ are the P outputs of the analog combining matrix. The use of linear mappings in the analog and digital domain offers a lot of benefits from a practical viewpoint and is already done in various hybrid A/D receiver architectures (see [35]–[38]).

Similar to previous works, we further relax the problem by considering linear tasks, i.e. $\tilde{\mathbf{s}} = \mathbf{\Gamma}\mathbf{x}$ for some $\mathbf{\Gamma} \in \mathbb{R}^{K \times N}$. Under the linear task scenario, we are able to derive closed-form expressions for $\mathbf{A}$, $\mathbf{D}$, and MSE than what were obtained

in the previous work. The framework developed for the linear task will be later extended to the more general nonlinear tasks.

B. Analysis Techniques for Task-based Quantization

The previous work[15] carried out the analysis assuming the system is equipped with non-subtractive uniform dithered quantizers. Whenever the input falls inside the dynamic range of a uniform dithered quantizer, the output can be written as a sum of the input and an additive zero-mean white quantization noise that is uncorrelated with the input. This simplication enables the derivation of the optimal linear mappings and MSE under a uniform dithered setting. Numerical results show that using these linear mappings on uniform undithered quantizers can further reduce the distortion. However, the theoretical framework established in [15] is unable to accurately predict the actual MSE of the task-based quantizer. We present numerical results in Sections IV and V to demonstrate this issue. Moreover, the non-subtractive dithering framework [30] does not apply to non-uniform quantization.

To avoid the shortcomings of the previous work, we consider a different analysis technique to represent the output of the scalar quantizers in a more analytically tractable form. More precisely, we use the generalized Bussgang decomposition to represent the quantization process as a noisy linear function of the input. That is,

$$\mathbf{z} = \mathcal{Q}_M^{1:P}(\mathbf{A}\mathbf{x})$$
$$= \mathbf{B}\mathbf{A}\mathbf{x} + \boldsymbol{\eta}. \quad (2)$$

Here, $\mathcal{Q}_M^{i:j}(\cdot)$ denotes the outputs of the scalar quantizers from index i to index j. The square matrix $\mathbf{B} \in \mathbb{R}^{P \times P}$ is called the Bussgang gain and $\boldsymbol{\eta} \in \mathbb{R}^{P \times 1}$ is the distortion vector uncorrelated with the quantizer input $\mathbf{A}\mathbf{x}$. The Bussgang gain matrix can be written as

$$\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{zx}} \mathbf{A}^T \left(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T\right)^{-1}, \quad (3)$$

where $\boldsymbol{\Sigma}_{\mathbf{zx}}$ is the cross-covariance between $\mathbf{z}$ and $\mathbf{x}$. The covariance of the distortion vector, denoted $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, can be expressed as

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{\Sigma}_{\mathbf{z}} - \mathbf{B}\left(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T\right)\mathbf{B}^T, \quad (4)$$

where $\boldsymbol{\Sigma}_{\mathbf{z}}$ is the covariance of $\mathbf{z}$.

The generalized Bussgang decomposition is exact; the intuition is that $\mathbf{B}\mathbf{A}\mathbf{x}$ is the linear MMSE estimate of $\mathbf{z}$ given the observation $\mathbf{A}\mathbf{x}$ (not necessarily Gaussian) [32]. However, the distribution of $\boldsymbol{\eta}$ is not known and the Bussgang gain matrix is, in general, not diagonal. Therefore, we introduce a restriction

---

[1]Other names used in the literature are remote source coding and noisy source coding.

on the structure of $\mathbf{A}$ that makes the Bussgang gain matrix $\mathbf{B}$ diagonal, regardless of the distribution of the measurement vector $\mathbf{x}$.

Restriction 1. *Suppose we denote $\mathbf{A}_i$ to be the i-th row of the analog combining matrix $\mathbf{A}$. Then, we pick $\mathbf{A}_i$ such that, for any $i \neq j$, we have $\mathbf{A}_i\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}} \perp \mathbf{A}_j\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}$, where $\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}$ is the matrix square root of $\boldsymbol{\Sigma}_{\mathbf{x}}$.*

Note that imposing Restriction 1 may yield sub-optimal task-based quantizer designs. In fact, we demonstrate in Section IV that our design framework, which is based on Restriction 1, does not always produce the task-based quantizer design with the lowest MSE. Nonetheless, our framework can achieve better performance than the current design and analysis frameworks when the scalar quantizers have very low resolution. Restriction 1 also forces the elements of $\boldsymbol{\eta}$ to be uncorrelated[2].

In the next section, we will show how imposing Restriction 1 forces $\mathbf{B}$ to be diagonal. We will then use Restriction 1 in conjunction with the generalized Bussgang decomposition to establish a new framework for analyzing and designing hardware-limited task-based quantizers.

III. MAIN RESULTS

We now characterize the hardware-limited task-based quantizer which minimizes (1) under Restriction 1. We first define a quantity that is crucial in stating the main results of the paper.

Definition 1. *Suppose $\mathbf{A}_p \in \mathbb{R}^{1 \times N}$ is the p-th row of $\mathbf{A}$. Then, the distortion factor of the p-th quantizer, denoted $\rho_{\mathrm{q}}^{(p)}$, accounts for the relative amount of distortion introduced by the p-th quantizer to its input and is expressed as*

$$\rho_{\mathrm{q}}^{(p)} = \frac{\mathbb{E}\{(Z_p - \mathbf{A}_p\mathbf{x})^2\}}{\mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\}}, \tag{5}$$

where Z_p is the output of the p-th scalar quantizer and the denominator term is the energy of the quantizer input.

A typical scenario in which the MMSE estimator $\tilde{\mathbf{s}}$ is a linear function of $\mathbf{x}$ is when the task vector $\mathbf{s}$ and the measurement vector $\mathbf{x}$ are jointly Gaussian [40, Section 3.2.7]. Conveniently, the quantizer inputs $\mathbf{A}\mathbf{x}$ are also Gaussian. The distortion factor $\rho_{\mathrm{q}}^{(p)}$ for a Gaussian input and Lloyd-Max scalar quantizer is tabulated in [41] for $\tilde{M} = 1$ up to $\tilde{M} = 36$ levels. For high-rate quantizers, the distortion factors of non-uniform and uniform quantizers under a Gaussian input are $\rho_{\mathrm{q}}^{(p)} \approx \frac{\pi\sqrt{3}}{2} \cdot \tilde{M}^{-2}$ and $\rho_{\mathrm{q}}^{(p)} \approx 1.47 \cdot \tilde{M}^{-1.74}$, respectively [2], [41].

The following proposition characterizes the diagonal entries of the Bussgang gain matrix under Restriction 1.

Proposition 1. *Under Restriction 1, the Bussgang gain matrix $\mathbf{B}$ is a diagonal matrix and can be expressed as*

$$\mathbf{B} = \mathrm{diag}\{\mathbf{1} - \boldsymbol{\rho}_{\mathrm{q}}\}, \tag{6}$$

where the $\mathrm{diag}\{\cdot\}$ operator generates a $P \times P$ diagonal matrix with entries coming from the $P \times 1$ vector $\{\cdot\}$, and $\boldsymbol{\rho}_{\mathrm{q}} = [\rho_{\mathrm{q}}^{(1)}, \rho_{\mathrm{q}}^{(2)}, \cdots, \rho_{\mathrm{q}}^{(P)}]^T$.

Proof. See Appendix A. □

We now present the main results of our work.

Proposition 2. *For any analog combining matrix $\mathbf{A}$ that satisfies Restriction 1, the optimal digital processing matrix, denoted $\mathbf{D}^\circ$, which minimizes the MSE is given by*

$$\mathbf{D}^\circ(\mathbf{A}) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T)^{-1}. \tag{7}$$

Consequently, the quantizer-dependent MSE can be expressed as

$$\mathbb{E}\{||\tilde{\mathbf{s}} - \hat{\mathbf{s}}||^2\}$$
$$= \mathrm{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma}^T\right)$$
$$- \mathrm{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T\mathbf{B}\left(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T\right)^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma}^T\right) \tag{8}$$

Proof. See Appendix B. □

Theorem 1. *Under Restriction 1, the optimal analog combining matrix, denoted $\mathbf{A}^\circ$, is*

$$\mathbf{A}^\circ = \mathbf{V}_{\mathrm{opt}}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}, \tag{9}$$

where the rows of $\mathbf{V}_{\mathrm{opt}}^T \in \mathbb{R}^{P \times N}$ are the P right singular vectors of $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}$ corresponding to the P largest singular values. The optimal digital processing matrix for a given $\mathbf{A} = \mathbf{A}^\circ$, denoted $\mathbf{D}^\circ(\mathbf{A}^\circ)$, is

$$\mathbf{D}^\circ(\mathbf{A}^\circ) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{V}_{\mathrm{opt}}. \tag{10}$$

Using $\mathbf{A}^\circ$ and $\mathbf{D}^\circ$ gives the following quantizer-dependent MSE:

$$\mathbb{E}\{||\tilde{\mathbf{s}} - \hat{\mathbf{s}}||^2\}$$
$$= \begin{cases} \sum_{i=1}^{K} \lambda_{\tilde{\boldsymbol{\Gamma}},i} \cdot \rho_{\mathrm{q}}^{(i)} & , \text{ if } P \geq K \\ \sum_{i=1}^{K} \lambda_{\tilde{\boldsymbol{\Gamma}},i} \cdot \rho_{\mathrm{q}}^{(i)} + \sum_{i=P+1}^{K} \lambda_{\tilde{\boldsymbol{\Gamma}},i} & , \text{ otherwise} \end{cases} \tag{11}$$

where $\lambda_{\tilde{\boldsymbol{\Gamma}},i}$ is the i-th eigenvalue of $\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^T$ (arranged in descending order).

Proof. See Appendix C. □

When the number of quantization levels per quantizer is sufficiently large and $P = K$, the distortion vector $\boldsymbol{\eta}$ becomes negligible, and the estimate of the task vector can be expressed as

$$\hat{\mathbf{s}} \approx \mathbf{D}^\circ\mathbf{A}^\circ\mathbf{x}$$
$$\approx \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{V}_{\mathrm{opt}}\mathbf{V}_{\mathrm{opt}}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{x}$$
$$\approx \boldsymbol{\Gamma}\mathbf{x} \approx \tilde{\mathbf{s}}$$

(i.e. our estimate of the task vector approaches the MMSE estimate). Consequently, the quantizer-dependent MSE term approaches zero since $\rho_{\mathrm{q}}^{(p)} \to 0$ as $\tilde{M} \to \infty$.

The design of the task-based quantizer in Theorem 1 has a nice intuition. The optimal analog combiner first applies a

---

[2]The requirements mentioned in [39, page 3] to make $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ diagonal are satisfied since $\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T$ and $\mathbf{B}$ are diagonal.

Figure 3: MSE vs overall number of bits of Systems A-D, task: channel estimation with $K = 2$ channel taps.
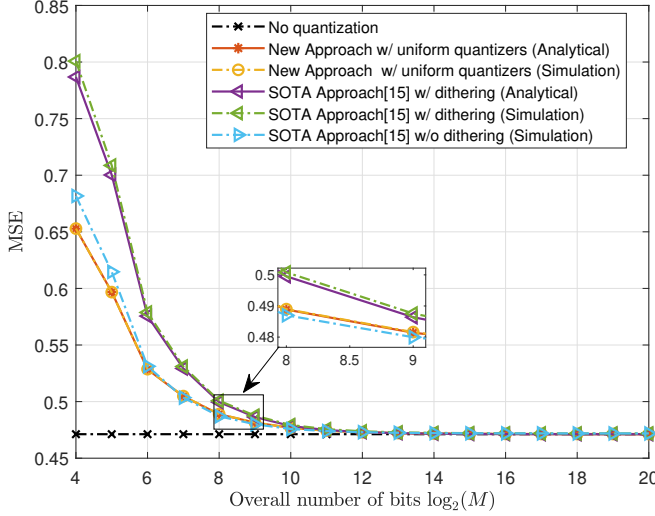

Figure 4: MSE vs overall number of bits of Systems A-D, task: channel estimation with $K = 8$ channel taps.

whitening filter to the measurement vector $\mathbf{x}$. Then, the matrix $\mathbf{V}_{\text{opt}}^T$ maps the "whitened" signal from the *measurement space* to a space with lower number of dimensions. We shall call this the *task space*. Quantization is performed in the task space to get $\mathbf{z}$. In the digital domain, the $\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{V}_{\text{opt}}$ term in $\mathbf{D}^\circ$ inverts the operation of $\mathbf{A}^\circ$ to get some intermediate result $\hat{\mathbf{x}}$, a linear estimate of $\mathbf{x}$ given $\mathbf{z}$. Finally, we compute $\hat{\mathbf{s}} = \boldsymbol{\Gamma}\hat{\mathbf{x}}$ to get an estimate of the task vector.

There are several differences between the linear mappings and MSE expressions derived in Theorem 1, and their counterparts in [15, Theorem 1]. First, the linear mappings we derived are independent of the actual structure of the scalar quantizers and their parameters. The linear mappings only depend on the underlying system task $\boldsymbol{\Gamma}$ and the statistics of the measurement vector $\mathbf{x}$. Thus, we get the same analog linear mappings for both uniform and non-uniform quantizers. In contrast, the optimal linear mappings in [15, Theorem 1] change as the quantizer parameters (e.g. M, dynamic range, spacing, etc) are varied. Second, we looked for the optimal $\mathbf{A}$ within the class of analog combiners that satisfy Restriction 1. However, we have not shown that there is no loss of optimality if we restrict the search space within this class. In fact, the analog linear mappings obtained using the SOTA approach do not necessarily satisfy this property. Third, the quantizer-dependent MSE expressions have different structures. The quantizer-dependent MSE expression in our new approach is a linear combination of the eigenvalues of $\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^T$, weighed by the distortion factors of the scalar quantizers. On the other hand, the quantizer-dependent MSE in [15, Theorem 1] and the eigenvalues of $\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^T$ exhibit a nonlinear relationship. We provide a more in-depth comparison of the two analysis frameworks in the next section.

IV. NUMERICAL STUDY FOR LINEAR TASK

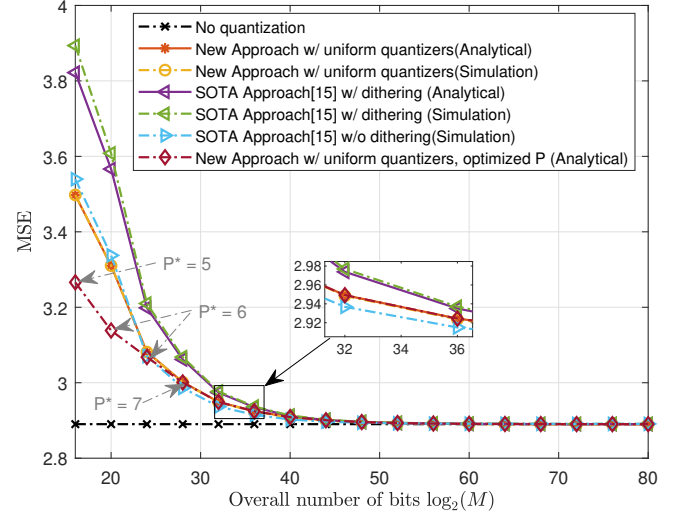We now apply our proposed analysis framework for the hardware-limited task-based quantization in a scenario which involves parameter acquisition from quantized observations. More precisely, we consider a scalar channel estimation problem where samples are corrupted by intersymbol interference (ISI) and noise, as in [15, Section VI-A]. The task vector $\mathbf{s}$ represents the coefficients of a K-taps multipath channel that we want to estimate. We aim to recover the task vector $\mathbf{s}$ from the $N = 120$ noisy observations contained in $\mathbf{x}$, where the n-th element of $\mathbf{x}$ is given by

$$x_n = \sum_{l=1}^{K} \mathbf{s}_l a_{n-l+1} + w_n \quad \forall n \in \{1, 2, \cdots, N\}. \quad (12)$$

The coefficients $\{a_l\}$ account for a deterministic training sequence that is known by the task-based quantizer. The quantities $\{w_n\}_{n=1}^{N}$ represent the i.i.d. zero-mean Gaussian noise process that has unit variance and is independent of $\mathbf{s}$. The channel $\mathbf{s}$ is modeled as a zero-mean Gaussian vector with the i-th row and j-th column of its covariance matrix is given by

$$\Sigma_{\mathbf{s}}^{(i,j)} = e^{-|i-j|}, \qquad \forall i, j \in \{1, 2, \cdots, K\}.$$

Effectively, $\mathbf{x}$ and $\mathbf{s}$ are jointly Gaussian so the linear task assumption $\tilde{\mathbf{s}} = \boldsymbol{\Gamma}\mathbf{x}$ is satisfied, where $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{\mathbf{sx}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$. Finally, we set the training sequence to be

$$a_l = \begin{cases} \cos\left(\frac{2\pi l}{N}\right), & l > 0 \\ 0, & \text{otherwise} \end{cases}. \quad (13)$$

Using the above setup, we evaluate the distortion of the hardware-limited task-based quantizer designed using our proposed analysis framework, and compare it to that of the hardware-limited task-based quantizer designed using the SOTA analysis framework. We consider two channels: (a) one with $K = 2$ channel taps, and (b) one with $K = 8$ channel taps. By default, we set $P = K$. However, we allow P to be optimized in some parts of the numerical study. For our proposed framework, we used the distortion factors
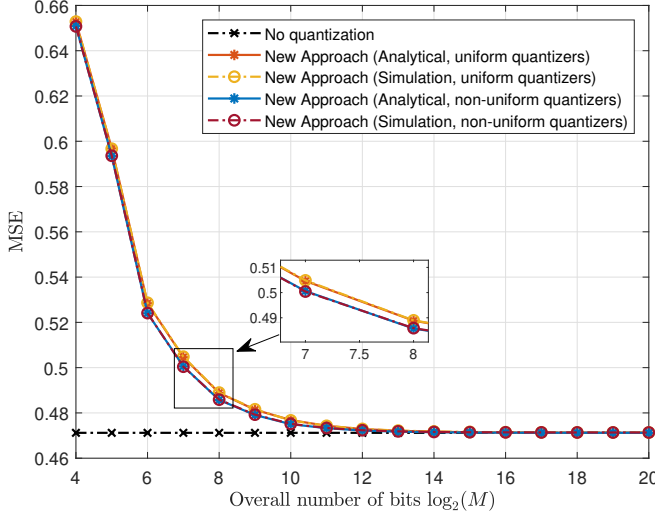
Figure 5: Comparison of task-based system with uniform and non-uniform quantization, channel estimation with $K = 2$.

for a Lloyd-Max non-uniform quantizer with Gaussian input. Since $\mathbf{x}$ is a Gaussian random vector, the quantizer input $\mathbf{A}_p\mathbf{x}$ for some $p \in \{1, 2, \cdots, P\}$ is a linear combination of N Gaussian random variables. Thus, $\mathbf{A}_p\mathbf{x}$ is also Gaussian. We also set the range of the overall quantization levels to be $\log_2 M \in [2 \cdot K, 10 \cdot K]$. Our numerical study will evaluate the distortions incurred by the following quantization systems:

- **System A (No quantization):** The MMSE $\mathbb{E}\{||\mathbf{s} - \tilde{\mathbf{s}}||^2\}$, which is the optimal distortion of an unquantized system. This quantity can be expressed as

$$\mathbb{E}\{||\mathbf{s} - \tilde{\mathbf{s}}||^2\} = \text{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{s}}\right) - \text{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{sx}}^T\right). \quad (14)$$

- **System B (SOTA Approach w/ dithering):** This is the distortion of the dithered hardware-limited task-based quantizer in which the analog and digital linear mappings $\mathbf{A}$ and $\mathbf{D}$ are designed using [15, Theorem 1]. Both simulated and theoretical distortions are evaluated. The simulated MSEs are computed empirically by averaging the MSE over 500,000 Monte Carlo runs.

- **System C (SOTA Approach w/o dithering):** Since dithering increases the energy of the quantization noise, we also simulate the MSE of the hardware-limited task-based quantizer without dithering ($\mathbf{A}$ and $\mathbf{D}$ are still designed using [15, Theorem 1]).

- **System D (New Approach, uniform quantizers):** This is the distortion of the hardware-limited task-based quantizer designed under our proposed analysis framework. The uniform quantizers are designed using the Lloyd-Max algorithm for equally-spaced level quantizers (See [41, Equation 8]). Both simulated and theoretical distortions are evaluated. The simulated MSEs are computed empirically by averaging the MSE over 500,000 Monte Carlo runs. Moreover, our analytical expression enables us to optimize the number of scalar quantizers. We also present the theoretical MSE using the optimal P,

denoted P^*. This is computed by trying all possible $P \in \{1, 2, \cdots, K\}$ in (11).

- **System E (New Approach, non-uniform quantizers):** This is the same as System D but we allow the quantizers to be non-uniform. The thresholds and representative levels of the scalar quantizers are designed using the Lloyd-Max algorithm [41]. Note that we did not change the configuration of the linear mappings since the derived linear mappings in Theorem 1 is agnostic of the actual structure of the scalar quantizers.

Figures 3 and 4 depict the distortions for System A to System D for $K = 2$ and $K = 8$ channel taps, respectively. In both cases, it can be observed that the MSE of System D is lower than that of System B. The performance gain is more pronounced in the low resolution regime but the gap between the MSEs of the two frameworks diminishes as the overall number of bits is increased. When all the scalar quantizers in the quantizer model have at least five bits, i.e. $\log_2 M \geq 5K$, the quantizer-dependent MSE is negligible and most of the overall MSE comes from (14). We also demonstrate in the $K = 8$ setup that using lower P may yield lower MSE when there is a tight quantization budget.

There is no clear winner between System C and System D. System D has lower MSE when the overall number of bits are limited but is slightly outperformed by System C at some values of $\log_2(M)$. We conjecture that its subpar performance at some cases is due to Restriction 1. That is, there is loss of optimality when restricting the search for the optimal $\mathbf{A}$ within a class of analog linear mappings that satisfy Restriction 1. Nonetheless, we point out that the simulated MSE and the theoretical MSE (i.e. Equation (14) + Equation (11)) of System D perfectly coincide in our numerical study. This is expected since the proposed framework is exact, provided the assumptions on the scalar quantizers and analog combining matrix are satisfied. On the other hand, we can see that the simulated MSE of the dithered task-based quantizer designed using the SOTA framework does not perfectly match the theoretical MSE. This is because the overload probabilities of the quantizers are nonzero. Thus, [15, Theorem 1] only holds approximately. Furthermore, the SOTA analysis framework is not capable of accurately predicting the simulated MSE of System C.

When the scalar quantizers in System D are replaced with non-uniform quantizers, i.e. System E, we observe in Figure 5 that the MSE of the task-based quantizer designed using our proposed analysis framework slightly improved. The use of non-uniform quantizers in our proposed framework is expected to provide performance gain, albeit small, since non-uniform quantizers generally have lower distortion factor compared to uniform quantizers. More importantly, we emphasize that the simulated MSE of the task-based system equipped with non-uniform quantizers coincides with our theoretical predictions. The proposed framework enables a model-based analysis of task-based quantization with non-uniform quantizers. To the best of our knowledge, only a data-driven approach [29] for task-based quantization with non-uniform quantizers is available in the literature.
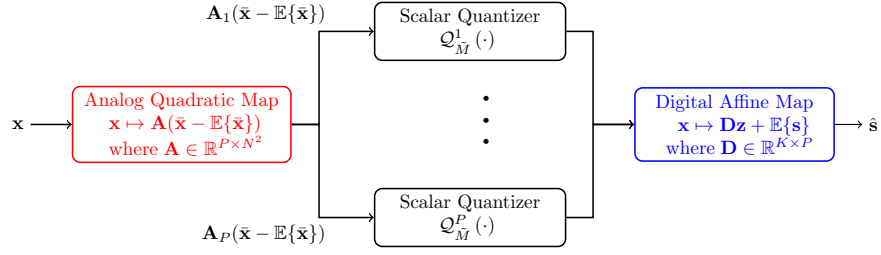
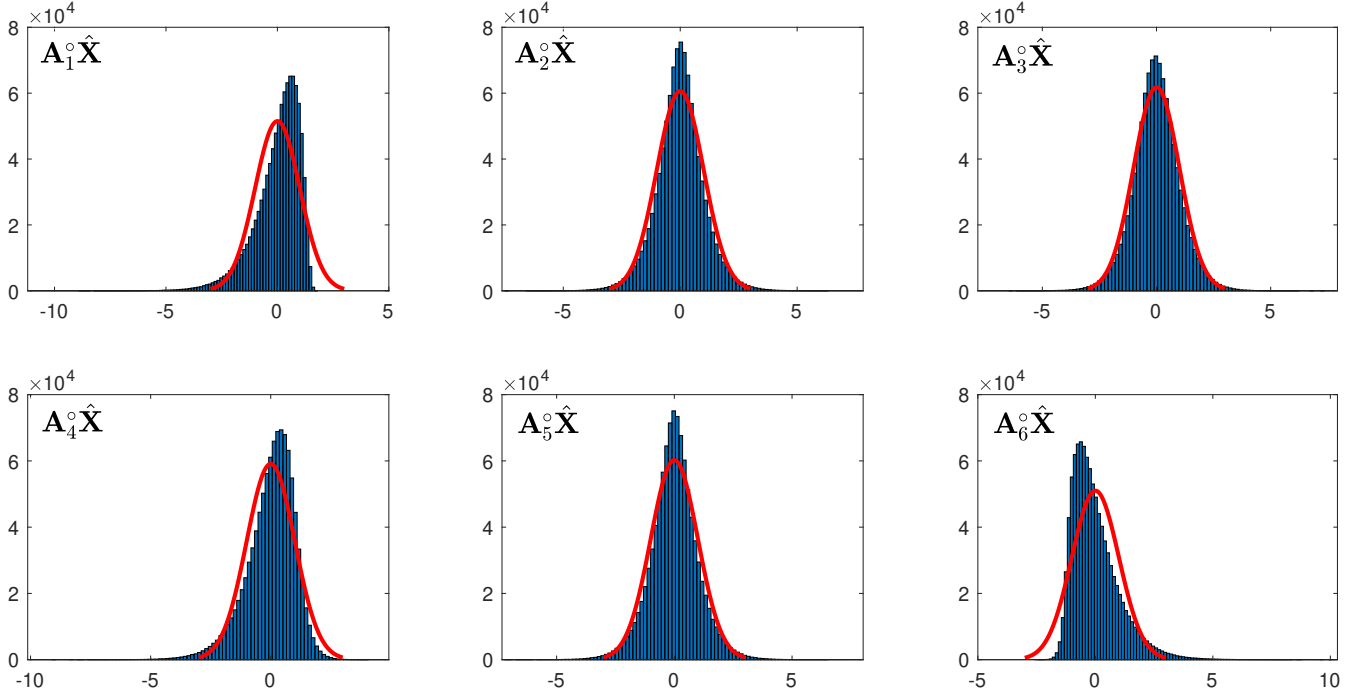Figure 6: System model of task-based quantization with analog quadratic mapping and digital affine mapping.

Figure 7: Empirical distributions of the six elements of $\mathbf{A}^{\circ}\bar{\mathbf{x}}$. The distributions are generated from 1,000,000 Monte Carlo runs. The red plots correspond to the best-fit Gaussian distributions.

In the next section, we extend the framework to quadratic tasks.

V. EXTENSION OF THE PROPOSED FRAMEWORK TO QUADRATIC TASK

A. Setup for the Quadratic Task Problem

To extend the proposed analysis framework to the quadratic task problem, we consider the task-based quantization model depicted in Figure 6 and follow the approach of [18]. The measurement vector $\mathbf{x} \in \mathbb{R}^{N \times 1}$ is a zero-mean Gaussian random vector and the task is to recover a set of quadratic functions $\{\mathbf{x}^T \mathbf{C}_k \mathbf{x}\}_{k=1}^K$, where each $\mathbf{C}_k \in \mathbb{R}^{N \times N}$ satisfies $\mathbb{E}\{\mathbf{x}^T \mathbf{C}_k \mathbf{x}\} < \infty$. We shall represent the results of these quadratic functions using a $K \times 1$ task vector $\mathbf{s}$ whose entries are given by $s_k = \mathbf{x}^T \mathbf{C}_k \mathbf{x}$.

Since we are interested in quadratic tasks, we introduce the quadratic measurement vector $\bar{\mathbf{x}} = \text{vec}\left(\mathbf{x}\mathbf{x}^T\right) \in \mathbb{R}^{N^2 \times 1}$, where $\text{vec}(\mathbf{C}) \in \mathbb{R}^{mn \times 1}$ is the vectorization of $\mathbf{C}^{m \times n}$, i.e.

the vector is a vertical stacking of the columns of $\mathbf{C}$. We also introduce $\mathbf{G} \in \mathbb{R}^{K \times N^2}$ whose k-th row is given by $\text{vec}^T(\mathbf{C}_k)$. Consequently, we can write the task vector as $\mathbf{s} = \mathbf{G}\bar{\mathbf{x}}$. This formulation "linearizes" the quadratic task problem.

Due to [18, Theorem 2], we can focus on analog mapping h_{a} of the form

$$h_{\mathrm{a}}: \quad \mathbf{x} \mapsto \mathbf{A}\left(\bar{\mathbf{x}} - \mathbb{E}\{\bar{\mathbf{x}}\}\right), \qquad (15)$$

where $\mathbf{A} \in \mathbb{R}^{P \times N^2}$ is a matrix that applies a rotation and dimensionality reduction to the shifted quadratic measurement vector $\bar{\mathbf{x}} - \mathbb{E}\{\bar{\mathbf{x}}\}$, and on digital mapping h_{d} of the form

$$h_{\mathrm{d}}: \quad \mathbf{x} \mapsto \mathbf{D}\mathbf{z} + \mathbb{E}\{\mathbf{s}\}, \qquad (16)$$

where $\mathbf{z} \in \mathbb{R}^{P \times 1}$ are the P outputs of the scalar quantizers. We find the matrices $\mathbf{A}$ and $\mathbf{D}$ that minimizes $\mathbb{E}\{\|\mathbf{s} - \hat{\mathbf{s}}\|^2\}$, where $\hat{\mathbf{s}}$ is the output of the task-based quantizer.

B. System Design

We now apply our proposed framework to the *linearized* quadratic task. First, we let $\boldsymbol{\Sigma}_{\bar{\mathbf{x}}} \in \mathbb{R}^{N^2 \times N^2}$ be the covariance matrix of $\bar{\mathbf{x}}$. Since $\mathbf{x}$ is a zero-mean Gaussian random vector, $\mathbf{x}\mathbf{x}^T$ is an $N \times N$ Wishart matrix of degree 1. Thus, the elements of $\boldsymbol{\Sigma}_{\bar{\mathbf{x}}} \in \mathbb{R}^{N^2 \times N^2}$ and $\mathbf{E}\{\bar{\mathbf{x}}\}$ can be obtained from [42]. Alternatively, these quantities can be computed empirically as done in [18]. The following corollary of Theorem 1 gives the quadratic task extension of the proposed framework.

Corollary 1. *Under Restriction 1, the optimal analog combining matrix, denoted $\mathbf{A}^\circ$, is*

$$\mathbf{A}^\circ = \bar{\mathbf{V}}_{\text{opt}}^T \boldsymbol{\Sigma}_{\bar{\mathbf{x}}}^{-\frac{1}{2}}, \tag{17}$$

where the rows of $\bar{\mathbf{V}}_{\text{opt}}^T \in \mathbb{R}^{P \times N^2}$ are the P right singular vectors of $\tilde{\mathbf{G}} = \mathbf{G}\boldsymbol{\Sigma}_{\bar{\mathbf{x}}}^{\frac{1}{2}}$ corresponding to the P largest singular values. The optimal digital processing matrix for a given $\mathbf{A} = \mathbf{A}^\circ$, denoted $\mathbf{D}^\circ(\mathbf{A}^\circ)$, is

$$\mathbf{D}^\circ(\mathbf{A}^\circ) = \mathbf{G}\boldsymbol{\Sigma}_{\bar{\mathbf{x}}}^{\frac{1}{2}}\bar{\mathbf{V}}_{\text{opt}}. \tag{18}$$

Using $\mathbf{A}^\circ$ and $\mathbf{D}^\circ$ gives the following MSE:

$$\mathbb{E}\{||\mathbf{s} - \hat{\mathbf{s}}||^2\}$$
$$= \begin{cases} \sum_{i=1}^{K} \lambda_{\tilde{\mathbf{G}},i} \cdot \rho_q^{(i)} & , \text{ if } P \geq K \\ \sum_{i=1}^{K} \lambda_{\tilde{\mathbf{G}},i} \cdot \rho_q^{(i)} + \sum_{i=P+1}^{K} \lambda_{\tilde{\mathbf{G}},i} & , \text{ otherwise} \end{cases} \tag{19}$$

where $\lambda_{\tilde{\mathbf{G}},i}$ is the i-th eigenvalue of $\tilde{\mathbf{G}}\tilde{\mathbf{G}}^T$ (arranged in descending order).

C. Numerical Results

We now demonstrate the effectiveness of the proposed framework on the quadratic task problem. We consider the empirical covariance estimation problem described in [18, Section V]. The input is given by $\mathbf{x} = [\mathbf{y}_1^T, \cdots, \mathbf{y}_4^T]^T$, where $\{\mathbf{y}_l\}_{l=1}^4$ are i.i.d. 3×1 zero mean Gaussian random vectors. Hence, the measurement vector $\mathbf{x}$ is a 12×1 vector. The i-th row and j-th column of the covariance matrix of each $\mathbf{y}_l$, denoted $\boldsymbol{\Sigma}_{\mathbf{y}}$, is given by

$$\boldsymbol{\Sigma}_{\mathbf{y}}^{(i,j)} = e^{-|i-j|}, \qquad \forall i, j \in \{1, 2, 3\}.$$

The parameter we want to recover is a 3×3 empirical covariance matrix $\frac{1}{4}\sum_{l=1}^4 \mathbf{y}_l\mathbf{y}_l^T$, which is completely determined by its upper triangular matrix. Thus, the task vector $\mathbf{s}$ has length $K = 6$.

Since $\bar{\mathbf{x}}$ is not a Gaussian random vector, we expect $\mathbf{A}^\circ\bar{\mathbf{x}}$ to be non-Gaussian as well. Indeed, as illustrated in Figure 7, the P outputs of the analog quadratic mapping are non-Gaussian. Therefore, we use the $\text{lloyds}(\cdot)$ function of MATLAB to get the P Lloyd-Max scalar quantizers and their corresponding distortion factors.

We evaluate the distortion incurred by the following task-based quantization systems:

- **System F (SOTA Approach for Quadratic Task [18]):** This is the distortion of the task-based quantizer where $\mathbf{A}$ and $\mathbf{D}$ are designed using [18, Theorem 3]. We present
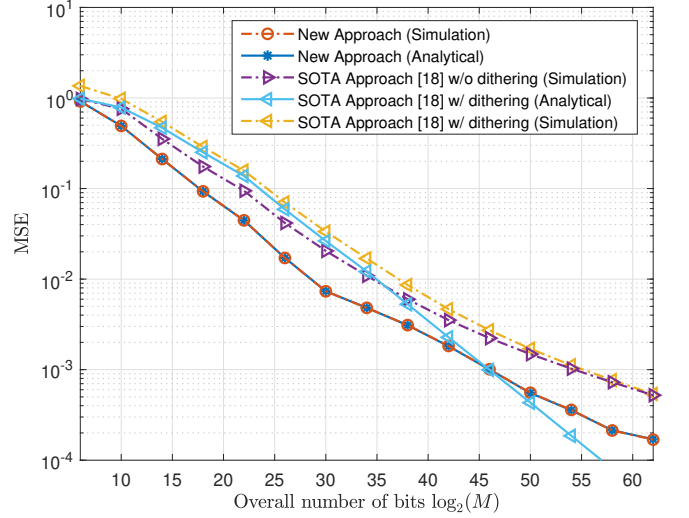

Figure 8: MSE vs overall number of bits, empirical covariance recovery.

the simulated MSEs of both the dithered and non-dithered case. The simulated MSEs are computed empirically by averaging the MSE over 1,000,000 Monte Carlo runs. The theoretical MSE [18, Theorem 3] is also evaluated.

- **System G (New Approach for Quadratic Task, non-uniform quantizers):** This is the distortion of the task-based quantizer where $\mathbf{A}$ and $\mathbf{D}$ are designed using Corollary 1. Both simulated and analytical MSEs are evaluated. Since the elements of $\mathbf{A}^\circ\bar{\mathbf{x}}$ are non-Gaussian and non-i.i.d., the distortion factors are first obtained empirically and then applied to equation (19) to produce the analytical MSE. The simulated MSE is computed empirically by averaging the MSE over 1,000,000 Monte Carlo runs.

Figure 8 shows the distortions of the quantization systems. It can be observed that the simulated MSE of System G is lower than that of non-dithered System F. Moreover, the analytical expression we derived closely matches the simulated MSE of our task-based quantizer. This numerical result demonstrates that our proposed analysis framework can be potentially applied to problems with nonlinear tasks. On the other hand, we see that while the analytical MSE expression in [18, Theorem 3] expects System F to yield the lowest MSE in certain scenarios, it does not accurately predict the actual MSE of the dithered task-based quantizer designed using the SOTA framework. In fact, the relative discrepancy gets worse as the number of quantization levels is increased.

VI. CONCLUSION

In this work, we formulated a new analysis framework based on the Bussgang decomposition for hardware-limited task-based quantization that overcomes limitations of the current SOTA framework. More precisely, our framework does not rely on the zero overload probability assumption and works for both uniform and non-uniform scalar quantizers without dithering. Our first contribution is a rigorous derivation of the

optimal linear mappings and analytical MSE in the linear task scenario under a restriction on the analog combiner. In contrast to the linear mappings derived under the SOTA framework, our mappings do not depend on the structure of the scalar quantizers. We then demonstrated in our numerical study that the simulated MSE of the task-based quantizer designed under our proposed framework coincides with our theoretical predictions. Additionally, we also extended our framework to quadratic tasks and showed that our analytical expression for the MSE continues to hold provided that the distortion factors are known or can be computed empirically.

One notable research direction is to investigate the applicability of the analysis framework to nonlinear tasks beyond the quadratic task problem. For instance, can we extend our proposed method to design and analyze hardware-limited task-based quantizers performing classification tasks? It is also interesting to see how the proposed framework can be utilized to design power-efficient analog mappings.

APPENDIX A
PROOF OF PROPOSITION 1

By Restriction 1, $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T$ becomes a diagonal matrix. That is,

$$\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T = \mathrm{diag}\{\mathbf{P_{Ax}}\}, \tag{20}$$

where $\mathbf{P_{Ax}} = [P_{\mathbf{Ax}}^{(1)}, \cdots, P_{\mathbf{Ax}}^{(P)}]^T$ and $P_{\mathbf{Ax}}^{(i)} = \mathbb{E}\{(\mathbf{A}_i\mathbf{x})^2\}$. In addition, it can be shown that

$$\begin{aligned}\mathbb{E}\{z_i\mathbf{x}^T\mathbf{A}_j^T\} &= \mathbb{E}\{(B_{ii}\mathbf{A}_i\mathbf{x} + \eta_i)\mathbf{x}^T\mathbf{A}_j^T\} \\ &= B_{ii}\mathbf{A}_i\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}_j^T + \mathbb{E}\{\eta_i\mathbf{x}^T\mathbf{A}_j^T\} \\ &= 0\end{aligned}$$

for $i \neq j$. The quantity B_{ii} is the i-th diagonal entry of $\mathbf{B}$. The first line follows by applying Bussgang decomposition at the output of the i-th quantizer. The first term in the second line is zero due to Restriction 1 while the second term in the second line is zero since the distortion at the i-th quantizer is uncorrelated with the input of the j-th quantizer. Thus, $\boldsymbol{\Sigma}_{\mathbf{zx}}\mathbf{A}^T$ is a diagonal matrix. Consequently, the Bussgang gain matrix in (3) is also diagonal.

To derive the diagonal elements of the Bussgang gain matrix, we expand (1):

$$\begin{aligned}\rho_{\mathrm{q}}^{(p)} &= \frac{\mathbb{E}\{(z_p - \mathbf{A}_p\mathbf{x})^2\}}{\mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\}} \\ &= \frac{\mathbb{E}\{z_p^2\} - 2\mathbb{E}\{z_p\mathbf{x}^T\mathbf{A}_p^T\} + \mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\}}{\mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\}}.\end{aligned}$$

Note that we considered scalar quantizers that satisfy the property $\mathbb{E}\{X_{\mathrm{in}}|\mathcal{Q}(X_{\mathrm{in}})\} = \mathcal{Q}(X_{\mathrm{in}})$. As such, we have

$$\begin{aligned}\mathbb{E}\{z_p\mathbf{x}^T\mathbf{A}_p^T\} &= \mathbb{E}\{\mathbb{E}\{z_p\mathbf{x}^T\mathbf{A}_p^T|z_p\}\} \\ &= \mathbb{E}\{z_p^2\}, \tag{21}\end{aligned}$$

where the first line follows from the law of iterated expectation, and the second line follows from $\mathbb{E}\{X_{\mathrm{in}}|\mathcal{Q}(X_{\mathrm{in}})\} = \mathcal{Q}(X_{\mathrm{in}})$. Effectively, the distortion factor becomes

$$\begin{aligned}\rho_{\mathrm{q}}^{(p)} &= \frac{\mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\} - \mathbb{E}\{z_p\mathbf{x}^T\mathbf{A}_p^T\}}{\mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\}} \\ &= 1 - \frac{\mathbb{E}\{z_p\mathbf{x}^T\mathbf{A}_p^T\}}{\mathbb{E}\{(\mathbf{A}_p\mathbf{x})^2\}} \\ &= 1 - B_{pp}.\end{aligned}$$

The claim is proven by doing the above analysis for all $p \in \{1, \cdots, P\}$.

APPENDIX B
PROOF OF PROPOSITION 2

From (1), we can simply focus on finding $\mathbf{D}^\circ$ that minimizes the quantizer-dependent MSE for a given $\mathbf{A}$. Under the assumption that $\tilde{\mathbf{s}}$ is a linear task, the optimal $\mathbf{D}$ which results in $\hat{\mathbf{s}}$ being the linear MMSE estimate of $\tilde{\mathbf{s}}$ given $\mathbf{z} = \mathbf{BAx} + \boldsymbol{\eta}$ is

$$\begin{aligned}\mathbf{D}^\circ(\mathbf{A}) &= \mathbb{E}\{\tilde{\mathbf{s}}\mathbf{z}^T\}\mathbb{E}\{\mathbf{zz}^T\}^{-1} \\ &= \mathbb{E}\{\tilde{\mathbf{s}}(\mathbf{BAx} + \boldsymbol{\eta})^T\}\mathbb{E}\{\mathbf{zz}^T\}^{-1} \\ &= \left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T\mathbf{B}^T + \boldsymbol{\Gamma}\mathbb{E}\{\mathbf{x}\boldsymbol{\eta}^T\}\right)\left(\boldsymbol{\Sigma}_{\mathbf{zx}}\mathbf{A}^T\right)^{-1} \\ &= \boldsymbol{\Gamma}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T\mathbf{B}^T(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T)^{-1}\mathbf{B}^{-1} \\ &= \boldsymbol{\Gamma}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T)^{-1}. \tag{22}\end{aligned}$$

The first line follows from the definition of a linear MMSE estimator. The second line follows from the generalized Bussgang decomposition. The third line follows from the linear task assumption $\tilde{\mathbf{s}} = \boldsymbol{\Gamma}\mathbf{x}$ and the relationship between $\boldsymbol{\Sigma}_\mathbf{z}$ and $\boldsymbol{\Sigma}_{\mathbf{zx}}$ established in equation (21). The fourth line is obtained from the Bussgang gain matrix expression in (3) and fact that $\mathbb{E}\{\mathbf{x}\boldsymbol{\eta}^T\} = \mathbf{0}$. To see this, we expand $\mathbb{E}\{\mathbf{x}\boldsymbol{\eta}^T\}$ as follows:

$$\begin{aligned}\mathbb{E}\{\mathbf{x}\boldsymbol{\eta}^T\} &= \mathbb{E}\{\mathbf{x}\left(\mathbf{z} - \mathbf{BAx}\right)^T\} \\ &= \boldsymbol{\Sigma}_{\mathbf{zx}}^T - \boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T\mathbf{B}^T \\ &= \mathbf{0},\end{aligned}$$

where the third line holds because of (3). Finally, the last line in (22) follows from the fact that $\mathbf{B}$ and $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T$ are diagonal matrices. As such,

$$\begin{aligned}\mathbf{B}^T(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T)^{-1}\mathbf{B}^{-1} &= \mathbf{B}^T\mathbf{B}^{-1}(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T)^{-1} \\ &= (\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T)^{-1}.\end{aligned}$$

Consequently, the quantizer-dependent MSE term of the linear MMSE estimator becomes

$$\begin{aligned}\mathbb{E}\{||\tilde{\mathbf{s}} - \hat{\mathbf{s}}||^2\} &= \mathbb{E}\{||\boldsymbol{\Gamma}\mathbf{x} - \mathbf{D}^\circ\mathbf{z}||^2\} \\ &= \mathrm{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\Gamma}^T\right) \\ &\quad - \mathrm{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T\mathbf{B}\left(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T\right)^{-1}\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\Gamma}^T\right),\end{aligned}$$

which proves the claim.

APPENDIX C
PROOF OF THEOREM 1

Let $\tilde{\mathbf{A}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}$. Then, the quantized-dependent MSE term becomes

$$
\begin{aligned}
\mathbf{E}\{||\tilde{\mathbf{s}} - \hat{\mathbf{s}}||^2\} =& \mathrm{Tr}\left(\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^T\right) \\
& - \mathrm{Tr}\left(\tilde{\boldsymbol{\Gamma}}\tilde{\mathbf{A}}^T\mathbf{B}\left(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\right)^{-1}\tilde{\mathbf{A}}\tilde{\boldsymbol{\Gamma}}^T\right) \\
=& \mathrm{Tr}\left(\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^T\right) \\
& - \mathrm{Tr}\left(\tilde{\mathbf{A}}\tilde{\boldsymbol{\Gamma}}^T\tilde{\boldsymbol{\Gamma}}\tilde{\mathbf{A}}^T\mathbf{B}\left(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\right)^{-1}\right), \quad (23)
\end{aligned}
$$

where the second line comes from the cyclic property of the trace function. Since the first term is independent of $\tilde{\mathbf{A}}$, the optimization problem simplifies to

$$
\tilde{\mathbf{A}}^\circ = \arg\max_{\tilde{\mathbf{A}}}\mathrm{Tr}\left(\tilde{\mathbf{A}}\tilde{\boldsymbol{\Gamma}}^T\tilde{\boldsymbol{\Gamma}}\tilde{\mathbf{A}}^T\mathbf{B}\left(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\right)^{-1}\right). \quad (24)
$$

Due to Restriction 1, $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$ is a diagonal matrix. Thus, we can also represent $\tilde{\mathbf{A}}$ as

$$
\tilde{\mathbf{A}} = \mathbf{H}_{\tilde{\mathbf{A}}}\mathbf{V}^T, \quad (25)
$$

where $\mathbf{H}_{\tilde{\mathbf{A}}} \in \mathbb{R}^{P \times N}$ is a scaling matrix whose off-diagonal entries are zero and whose p-th entry in the main diagonal corresponds to a scaling of the p-th output of the analog combining matrix. The matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$ is a unitary matrix. Under this setting, we can reduce the optimization problem to

$$
\mathbf{H}^\circ, \mathbf{V}^\circ = \arg\max_{\mathbf{H}_{\tilde{\mathbf{A}}}, \mathbf{V}} \mathrm{Tr}\left(\mathbf{H}_{\tilde{\mathbf{A}}}\mathbf{V}^T\tilde{\boldsymbol{\Gamma}}^T\tilde{\boldsymbol{\Gamma}}\mathbf{V}\mathbf{H}_{\tilde{\mathbf{A}}}^T\mathbf{B}\left(\mathbf{H}_{\tilde{\mathbf{A}}}\mathbf{H}_{\tilde{\mathbf{A}}}^T\right)^{-1}\right).
$$

Due to [43, Theorem II.1], $\mathbf{V}^\circ$ is the matrix containing the right singular vectors of $\tilde{\boldsymbol{\Gamma}}$. This further simplifies the optimization problem to

$$
\mathbf{H}^\circ = \arg\max_{\mathbf{H}_{\tilde{\mathbf{A}}}} \sum_{i=1}^{\min(P,K)} \lambda_{\tilde{\boldsymbol{\Gamma}},i}[1 - \rho_{\mathrm{q}}^{(i)}],
$$

where $\lambda_{\tilde{\boldsymbol{\Gamma}},i}$ is the i-th eigenvalue of $\tilde{\boldsymbol{\Gamma}}^T\tilde{\boldsymbol{\Gamma}}$ (arranged in descending order). It can be observed that the new objective function is independent of $\mathbf{H}_{\tilde{\mathbf{A}}}$ as long as the entries of main diagonal of $\mathbf{H}_{\tilde{\mathbf{A}}}$ are positive (otherwise, $\mathbf{H}_{\tilde{\mathbf{A}}}\mathbf{H}_{\tilde{\mathbf{A}}}^T$ will not be invertible). Without loss of generality, we set $\mathbf{H}^\circ = [\mathbf{I}_{P \times P} \ \mathbf{0}_{P \times (N-P)}]$. Consequently, we get

$$
\begin{aligned}
\mathbf{A}^\circ =& \mathbf{H}^\circ(\mathbf{V}^\circ)^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}} \\
=& \mathbf{V}_{\mathrm{opt}}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}},
\end{aligned}
$$

where the rows of $\mathbf{V}_{\mathrm{opt}}^T$ are the P right singular vectors of $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}$ corresponding to the P largest singular values. To verify that $\mathbf{A}^\circ$ satisfies Restriction 1, note that

$$
\begin{aligned}
\mathbf{A}^\circ\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{A}^\circ)^T =& \mathbf{V}_{\mathrm{opt}}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{V}_{\mathrm{opt}} \\
=& \mathbf{V}_{\mathrm{opt}}^T\mathbf{V}_{\mathrm{opt}}.
\end{aligned}
$$

Since the singular vectors are orthogonal to each other, then $\mathbf{V}_{\mathrm{opt}}^T\mathbf{V}_{\mathrm{opt}}$ is a diagonal matrix.

By plugging in $\mathbf{A}^\circ$ to $\mathbf{D}^\circ(\mathbf{A})$, we get

$$
\mathbf{D}^\circ(\mathbf{A}^\circ) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{V}_{\mathrm{opt}}. \quad (26)
$$

Finally, the quantizer dependent MSE can be written as

$$
\begin{aligned}
&\mathbb{E}\{||\tilde{\mathbf{s}} - \hat{\mathbf{s}}||^2\} \\
&= \sum_{i=1}^{K} \lambda_{\tilde{\boldsymbol{\Gamma}},i} - \sum_{i=1}^{\min(K,P)} \lambda_{\tilde{\boldsymbol{\Gamma}},i} \cdot [1 - \rho_{\mathrm{q}}^{(i)}] \\
&= \begin{cases} \sum_{i=1}^{K} \lambda_{\tilde{\boldsymbol{\Gamma}},i} \cdot \rho_{\mathrm{q}}^{(i)}, & , \text{if } P \geq K \\ \sum_{i=1}^{P} \lambda_{\tilde{\boldsymbol{\Gamma}},i} \cdot \rho_{\mathrm{q}}^{(i)} + \sum_{i=P+1}^{K} \lambda_{\tilde{\boldsymbol{\Gamma}},i}, & \text{otherwise.} \end{cases}
\end{aligned}
$$

REFERENCES

[1] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. USA: Kluwer Academic Publishers, 1991.

[2] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[3] T. Berger and J. Gibson, "Lossy Source Coding," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2693–2723, 1998.

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

[5] S. V. Kosonocky and P. H. Xiao, "Analog-to-digital conversion architectures," in *Digital Signal Processing Handbook* (V. K. Madisetti and D. B. Williams, eds.), CRC Press LLC, 1999.

[6] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 539–550, April 1999.

[7] N. I. Bernardo, J. Zhu, and J. Evans, "On Minimizing Symbol Error Rate Over Fading Channels With Low-Resolution Quantization," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7205–7221, 2021.

[8] N. I. Bernardo, J. Zhu, and J. Evans, "On the Capacity-Achieving Input of Channels with Phase Quantization," *IEEE Transactions on Information Theory*, pp. 1–1, 2022.

[9] N. I. Bernardo, J. Zhu, and J. Evans, "On the Capacity-Achieving Input of the Gaussian Channel with Polar Quantization," *IEEE Transactions on Communications*, pp. 1–1, 2022.

[10] J. Choi, G. Lee, A. Alkhateeb, A. Gatherer, N. Al-Dhahir, and B. L. Evans, "Advanced Receiver Architectures for Millimeter-Wave Communications with Low-Resolution ADCs," *IEEE Communications Magazine*, vol. 58, no. 8, pp. 42–48, 2020.

[11] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-Optimal Joint Channel-and-Data Estimation for Massive MIMO With Low-Precision ADCs," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2541–2556, 2015.

[12] M. Schlüter, M. Dörpinghaus, and G. P. Fettweis, "Bounds on Phase, Frequency, and Timing Synchronization in Fully Digital Receivers With 1-bit Quantization and Oversampling," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6499–6513, 2020.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[14] A. Kipnis, Y. C. Eldar, and A. J. Goldsmith, "Fundamental Distortion Limits of Analog-to-Digital Compression," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6013–6033, 2018.

[15] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Hardware-Limited Task-Based Quantization," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5223–5238, 2019.

[16] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Asymptotic task-based quantization with application to massive mimo," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 3995–4012, 2019.

[17] N. Shlezinger and Y. C. Eldar, "Task-Based Quantization with Application to MIMO Receivers," *Communications in Information and Systems*, vol. 20, pp. 131–162, 2020.

[18] S. Salamatian, N. Shlezinger, Y. C. Eldar, and M. Médard, "Task-Based Quantization for Recovering Quadratic Functions Using Principal Inertia Components," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 390–394, 2019.

[19] F. Xi, N. Shlezinger, and Y. C. Eldar, "BiLiMO: Bit-Limited MIMO Radar via Task-Based Quantization," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6267–6282, 2021.

[20] T. Zirtiloglu, N. Shlezinger, Y. C. Eldar, and R. Tugce Yazicigil, "Power-efficient hybrid mimo receiver with task-specific beamforming using low-resolution adcs," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5338–5342, 2022.

[21] S. Rini, L. Barletta, Y. C. Eldar, and E. Erkip, "A General Framework for MIMO Receivers with Low-Resolution Quantization," in *2017 IEEE Information Theory Workshop (ITW)*, pp. 599–603, 2017.

[22] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "MIMO Networks with One-Bit ADCs: Receiver Design and Communication Strategies," *IEEE Transactions on Communications*, pp. 1–1, 2021.

[23] N. I. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Capacity Bounds for One-Bit MIMO Gaussian Channels with Analog Combining," 2022. arXiv Preprint.

[24] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, "LoRD-Net: Unfolded Deep Detection Network With Low-Resolution Receivers," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5651–5664, 2021.

[25] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, "Model-Inspired Deep Detection with Low-Resolution Receivers," in *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 3349–3354, 2021.

[26] P. Li, N. Shlezinger, H. Zhang, B. Wang, and Y. C. Eldar, "Graph signal compression via task-based quantization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5514–5518, 2021.

[27] P. Neuhaus, N. Shlezinger, M. Dörpinghaus, Y. C. Eldar, and G. Fettweis, "Task-based analog-to-digital converters," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5403–5418, 2021.

[28] P. Neuhaus, N. Shlezinger, M. Dörpinghaus, Y. C. Eldar, and G. Fettweis, "Task-based analog-to-digital converters for bandlimited systems," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1985–1989, 2021.

[29] N. Shlezinger and Y. Eldar, "Deep task-based quantization," *Entropy*, vol. 23, p. 104, 01 2021.

[30] R. Wannamaker, S. Lipshitz, J. Vanderkooy, and J. Wright, "A Theory of Nonsubtractive Dither," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 499–516, 2000.

[31] E. Akyol and K. Rose, "On Constrained Randomized Quantization," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3291–3302, 2013.

[32] O. T. Demir and E. Bjornson, "The bussgang decomposition of nonlinear systems: Basic theory and mimo extensions [lecture notes]," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 131–136, 2021.

[33] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[34] R. Dobrushin and B. Tsybakov, "Information Transmission with Additional Noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.

[35] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, "Hybrid Architectures With Few-Bit ADC Receivers: Achievable Rates and Energy-Rate Tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2274–2287, 2017.

[36] K. Roth and J. A. Nossek, "Achievable Rate and Energy Efficiency of Hybrid and Digital Beamforming Receivers With Low Resolution ADC," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2056–2068, 2017.

[37] X. Song, T. Hälsig, W. Rave, B. Lankl, and G. Fettweis, "Analog Equalization and Low Resolution Quantization in Strong Line-of-sight MIMO Communication," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2016.

[38] S. S. Ioushua and Y. C. Eldar, "A Family of Hybrid Analog–Digital Beamforming Methods for Massive MIMO Systems," *IEEE Transactions on Signal Processing*, vol. 67, no. 12, pp. 3243–3257, 2019.

[39] E. Björnson, L. Sanguinetti, and J. Hoydis, "Hardware Distortion Correlation Has Negligible Impact on UL Massive MIMO Spectral Efficiency," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1085–1098, 2019.

[40] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. USA: John Wiley & Sons, Inc., 1st ed., 1996.

[41] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, 1960.

[42] S. W. Nydick, "The Wishart and Inverse Wishart Distributions," 2012.

[43] J. Lasserre, "A Trace Inequality for Matrix Product," *IEEE Transactions on Automatic Control*, vol. 40, no. 8, pp. 1500–1501, 1995.