Non-Linear Analog Processing Gains in Task-Based Quantization

Marian Temprana Alonso*, Farhad Shirani*, Neil Irwin Bernardo†, Yonina C. Eldar‡
*School of Computing and Information Sciences, Florida International University, Miami, FL, {mtempran,fshirani}@fiu.edu
†Electrical and Electronics Engineering Institute, University of the Philippines Diliman,
Quezon City, Philippines, neil.bernardo@eee.upd.edu.ph
‡ Department of Mathematics and Computer Science, Weizmann Institute of Science,
Rehovot, Israel, yonina.eldar@weizmann.ac.il

***Abstract*—In task-based quantization, a multivariate analog signal is transformed into a digital signal using a limited number of low-resolution analog-to-digital converters (ADCs). This process aims to minimize a fidelity criterion, which is assessed against an unobserved task variable that is correlated with the analog signal. The scenario models various applications of interest such as channel estimation, medical imaging applications, and object localization. This work explores the integration of analog processing components—such as analog delay elements, polynomial operators, and envelope detectors—prior to ADC quantization. Specifically, four scenarios, involving different collections of analog processing operators are considered: (i) arbitrary polynomial operators with analog delay elements, (ii) limited-degree polynomial operators, excluding delay elements, (iii) sequences of envelope detectors, and (iv) a combination of analog delay elements and linear combiners. For each scenario, the minimum achievable distortion is quantified through derivation of computable expressions in various statistical settings. It is shown that analog processing can significantly reduce the distortion in task reconstruction. Numerical simulations in a Gaussian example are provided to give further insights into the aforementioned analog processing gains.**

I. Introduction

Sensing, communication, and data compression systems utilize analog-to-digital converters (ADCs) to transform observed continuous-time analog signals into digital signals which can then be efficiently processed, communicated, and stored [1]–[12]. An ADC typically samples the signal at equally-spaced time intervals, and the amplitude of each sample is sequentially mapped onto a finite collection of quantization bins via comparison with pre-determined thresholds. The number of quantization bins is determined by the resolution of the ADC, and is quantified in terms of its output bits, e.g., a one-bit ADC has two quantization bins and its operation is parameterized by a single ADC threshold. Increasing the ADC resolution leads to reduced distortion. However, the ADC power consumption grows exponentially in the number of output bits. More precisely, in theory, the power consumption of an ADC is proportional to $f_s 2^{n_q}$, where f_s is the sampling rate and n_q is the number of output bits of the ADC [1], [13]. As an example, the power consumption of current commercial high-speed (≥ 20 GSample/s), high-resolution (e.g., 8-12 bits) ADCs is around 500 mW per ADC [14]. This has led to significant recent interest in the use of low-resolution ADCs in data acquisition and processing systems and the design of hardware architectures and algorithms which mitigate the resulting loss in distortion due to coarse quantization.

Task-based quantization has emerged as a promising solution to mitigate the aforementioned rate-loss due to coarse quantization using low resolution ADCs [5]–[9], [15]–[19]. The idea in task-based quantization is that the analog signal observed by the system is often digitized to be processed towards accomplishing a specific task, e.g., channel estimation, object localization, or pattern recognition in medical imaging [5], [20]–[22]. Consequently, the ADCs and their accompanying analog processing circuits may be designed in a way to extract the task-relevant bits of information from the analog signal, while filtering out the irrelevant information through the lossy quantization process. In other words, the analog processing components and ADC thresholds are designed so that the distortion between the task reconstruction and the ground-truth task is minimized, rather than minimizing the distortion between the original signal and its reconstruction in the digital domain [5], [6], [23]. Consequently, performance gains in task-based quantization are achieved by employing a hybrid analog/digital (A/D) architecture and jointly designing the analog pre-quantization mapping and digital post-quantization mapping with respect to the underlying task.

Prior design frameworks for task-based quantization have focused on linear processing in the analog domain. In this work, we consider the use of non-linear analog processing operators using implementable collections of analog components — consisting of analog delay elements, polynomial operators, and envelope detectors prior to ADC quantization — to further mitigate the coarse quantization distortion loss when using low resolution ADCs. This builds upon recent works [10], [11], [24], where the design and implementation of such circuit components for high frequency applications was considered in the context of wireless communications. It was shown that the power consumption of these analog processing components is negligible compared to that of the ADCs, hence justifying their application in such scenarios. Particularly, we consider four scenarios using analog operators consisting of: (i) arbitrary polynomial operators with analog

This work was supported in part by NSF grants CCF-2241057 and CCSS-2242700.

delay elements, (ii) limited-degree polynomial operators, excluding delay elements, (iii) sequences of envelope detectors, and (iv) a combination of analog delay elements and linear combiners. In each scenario, we quantify the fundamental performance limits, in terms of achievable distortion in task reconstruction under general statistical assumptions on the task statistics. Furthermore, given a fixed ADC power budget — using a fixed number and resolution of ADCs — we show that the resulting task-reconstruction distortion decreases compared to the prior approach of using linear analog processing.

Notation: The set $\{1,2,\cdots,n\}, n \in \mathbb{N}$ is represented by $[n]$. The vector $(x_1,x_2,\dots,x_n)$ is written as $x(1{:}n)$ and x^n, interchangeably. The ith element is written as $x(i)$ and x_i, interchangeably. An $n\times m$ matrix is written as $h(1{:}n,1{:}m) = [h_{i,j}]_{i,j\in[n]\times[m]}$. Sets are denoted by calligraphic letters such as $\mathcal{X}$.

II. Problem Formulation

The task-based quantization setting considered in this work is shown in Figure 1. In the following, we describe the general problem formulation, and provide examples in the context of channel estimation as a motivating application.

Task vector: The (unobserved) sequence of task vectors $S^{n\times\ell} = (S^n(1), S^n(2),\cdots,S^n(\ell))$ are independently and identically distributed according to an underlying probability distribution $P_{S^n}(\cdot)$ defined on $\mathbb{R}^{n\times\ell}$, where $n\in\mathbb{N}$ is the dimension of the task vector and $\ell\in\mathbb{N}$ is the blocklength. The vector $S^n(j)$ is the task vector at *time* j, $j\in[\ell]$. The objective in task-based quantization is to produce an accurate reconstruction of the task vector based on a sequence of coarsely quantized indirect observations. As an example, in the context of channel estimation, the task vector $S^n(j)$ represents the channel coefficient matrix at time j, and the objective is to produce an accurate channel estimate via indirect observations acquired by sending a sequence of pilot signals over the channel.

Measurement Vector: The (observed) sequence of measurements is a sequence of real-valued vectors $X^{m\times\ell} = (X^m(1), X^m(2),\cdots,X^m(\ell)$, where $m\in\mathbb{N}$. Each $X^m(j), j\in[\ell]$ is produced conditioned on the realization of the task-vector $s^n(j)$ according to the conditional distribution $P_{X^m|S^n}(\cdot|s^n(j))$. For instance, in the context of channel estimation, the measurement vector at time j models the analog channel output when a pilot signal is sent over the channel.

Analog Processing Functions: The measurement vectors $X^{m\times\ell}$ are fed sequentially to a collection of analog processing functions $f^{(a)}_{i,j}:\mathbb{R}^{m\times j}\to\mathbb{R}, i\in[n_q], j\in[\ell]$, where $n_q\in\mathbb{N}$, and the choice of $f^{(a)}_{i,j}, i\in[n_q], j\in[\ell]$ is restricted by the limitations of the analog circuit design as discussed in the sequel. In general, we assume that the analog processing functions at time j are chosen from a set $\mathcal{F}_{a,j}$ of *implementable analog functions.* The output of the analog processing functions is denoted by $W^{n_q\times\ell}$, where $W_{i,j}\triangleq f^{(a)}_{i,j}(X^{m\times i}), i\in[n_q], j\in[\ell]$, $f^{(a)}_{i,j}\in\mathcal{F}_{a,j}$, and $n_q\in\mathbb{N}$. Note that in the general scenario described here, $W_{i,j}$ may casually depend on the past realizations of the measurement vector. The analog processing functions may consist of linear combiners, delay elements, non-linear operators such as low degree polynomial operators, and envelope detectors [5], [10], [25], [26]. For a fixed number and resolution of ADCs, our objective is to quantify the gains due to the use of each of the aforementioned classes of non-linear analog processing functions, in terms of achievable distortion, in comparison with linear analog processing.

ADC Module. At time $j\in[\ell]$, the processed signal vector $W^{n_q}(j)$ is fed to a set of n_q ADCs each with $\kappa\in\mathbb{N}$ output levels. The quantization output is defined as $\widehat{W}^{n_q}(j)$, where

$$\widehat{W}(j,k) = k \iff W(j,i)\in[t_j(i,k), t_j(i,k+1)], \qquad (1)$$

$k\in[0,\kappa-1]$, $i\in[n_q]$, and we have defined $t_j(i,0)\triangleq-\infty$ and $t_j(i,\kappa)\triangleq\infty$. We call $t_j^{n_q\times(\kappa)}$ the *threshold matrix* at time j. $\widehat{W}(j,i)$ is called the quantization output of the κ-level ADC with thresholds $t_j^{n_q}(i)$ for input $W_j(i)$.

Digital Processing Function: A digital processing function $f_d:\mathbb{R}^{n_q\times\ell}\to\mathbb{R}^{n\times\ell}$ acts on the sequence of quantized vectors $\widehat{W}^{n_q\times\ell}$ to produce the task reconstruction $\widehat{S}^{n\times\ell}$. There are no restrictions on the choice of the digital processing function.

Distortion Function: Given $d:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^+$, the ℓ-shot distortion is defined as:

$$d_\ell \triangleq \frac{1}{\ell}\sum_{j=1}^{\ell}\mathbb{E}_{S^{n\times\ell},X^{m\times\ell}}(d(S^n(j),\widehat{S}^n(j))).$$

In summary, a task-based quantization setup is characterized by the tuple $(n,m,P_{S^n},P_{X^m|S^n},(\mathcal{F}_{a,j})_{j\in\mathbb{N}},n_q,\kappa,d(\cdot,\cdot))$.

Given a collection of analog processing functions $(f^{(a)}_{i,j})_{i\in[n_q],j\in[\ell]}$ and thresholds $t_j^{n_q\times\kappa}$, $j\in[\ell]$, the digital processing function minimizing distortion is given by:

$$f^*_d = \operatorname*{arg\,min}_{f_d:\mathbb{R}^{n_q\times\ell}\to\mathbb{R}^{n\times\ell}} \mathbb{E}_{S^{n\times\ell},X^{m\times\ell}}(d(f_d(W^{n_q\times\ell}),S^{n\times\ell})).$$

For instance, if $d(\cdot,\cdot)$ is the square error distortion function, then by the orthogonality principle, we have $f^*_d(W^{n_q\times\ell}) = \mathbb{E}(S^{n\times\ell}|W^{n_q\times\ell})$. Since there are no restrictions on the choice of the digital processing functions, in the sequel, we always assume that the optimal digital processing function is used for reconstruction, i.e., $\widehat{S}^{n\times\ell} = f^*_d(W^{n_q\times\ell})$. Consequently, we focus on the optimization problem for the choice of analog processing functions and quantization thresholds.

System Objective: The objective in task-based quantization is to find the optimal choice of system parameters which minimize the achievable distortion given a fixed number and resolution of ADCs and a fixed collection of implementable analog processing functions $\mathcal{F}_{a,j}, j\in\mathbb{N}$, To elaborate, the minimum ℓ-shot achievable distortion is defined as:

$$d^*_\ell \triangleq \min_{\substack{(f^{(a)}_{i,j})_{i\in[n_q],j\in[\ell]}\in\mathcal{F}_{a,j}\\ t_j^{n_q\times\kappa}\in\mathbb{R}^{n_q\times\kappa}}} \frac{1}{\ell}\sum_{j=1}^{\ell}\mathbb{E}_{S^{n\times\ell},X^{m\times\ell}}(d(S^n(j),\widehat{S}^n(j))). \qquad (2)$$

The collection of functions $(f^{(a)}_{i,j})_{i\in[n_q],j\in[\ell]}$ and thresholds $t_j^{n_q\times\kappa}$, $j\in[\ell]$ minimizing (2) are called the ℓ-shot optimal functions and thresholds, respectively. Our objective is to characterize d^*_ℓ and the corresponding processing functions and thresholds.

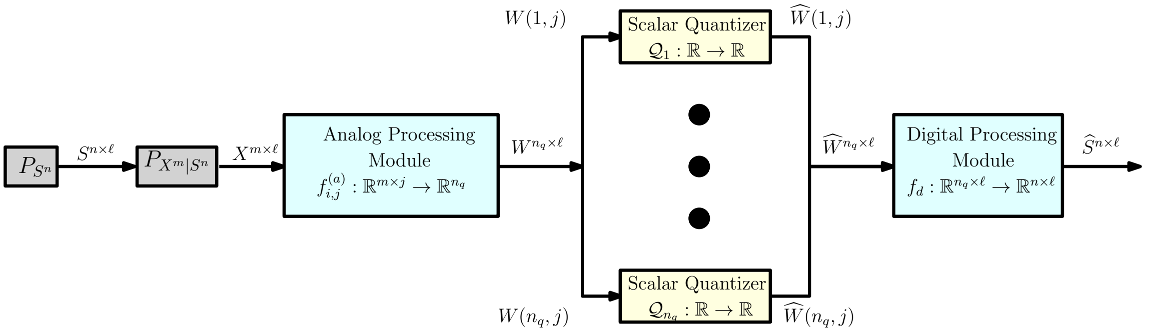
P_{S^n}
$S^{n\times\ell}$
$P_{X^m|S^n}$
$X^{m\times\ell}$
Analog Processing Module
$f_{i,j}^{(a)}:\mathbb{R}^{m\times j}\to\mathbb{R}^{n_q}$
$W^{n_q\times\ell}$
$W(1,j)$
Scalar Quantizer
$\mathcal{Q}_1:\mathbb{R}\to\mathbb{R}$
$\widehat{W}(1,j)$
$W(n_q,j)$
Scalar Quantizer
$\mathcal{Q}_{n_q}:\mathbb{R}\to\mathbb{R}$
$\widehat{W}(n_q,j)$
$\widehat{W}^{n_q\times\ell}$
Digital Processing Module
$f_d:\mathbb{R}^{n_q\times\ell}\to\mathbb{R}^{n\times\ell}$
$\widehat{S}^{n\times\ell}$

Fig. 1. The task-based quantization setup.

III. An Illustrative Gaussian Example

In order to motivate the use of non-linear processing prior to quantization, in this section, we focus on a simple Gaussian example, and provide an intuitive justification of the performance gains due to using non-linear processing over linear processing. In the subsequent sections, we build upon the intuition provided by this example, and study the fundamental performance limits of the general task-based quantization problem using various classes of non-linear analog processing functions. Section V numerically evaluates the achievable distortion in each of the scenarios considered in this section.

Let us take $n = m = 1$, and let the task be characterized by a zero-mean, unit variance, Gaussian random variable, i.e., $S \sim \mathcal{N}(0,1)$. Additionally, let us assume that the measurement vector is produced by passing the task through a Gaussian additive channel, i.e., $X = S+N, N \sim \mathcal{N}(0,\sigma_N^2)$, where $\sigma_N \in \mathbb{R}$ and S and N are independent of each other. Furthermore, let the quantization system be equipped by two one-bit ADCs, i.e., $n_q = \kappa = 2$. Finally, we take $d(s,s') = (s-s')^2$ as the square distortion. We consider four scenarios, and find the minimum achievable distortion in each case.

1) Scenario 1. Linear Analog Processing: In this scenario, we restrict $f_a : \mathbb{R} \to \mathbb{R}$ to affine transformations, i.e., $\mathcal{F}_{a,j} = \{f_a | f_a(x) = bx+c, b, c \in \mathbb{R}\}, j \in [\ell]$. It is straightforward to see, using the orthogonality principle, that the minimum one-shot achievable distortion is given by:

$$d^*_{1,lin} = \min_{\tau_1,\tau_2:\tau_1<\tau_2} \mathbb{E}_{S,N}((S-\widehat{S})^2), \tag{3}$$

where $\widehat{S} \triangleq \mathbb{E}(S|\widehat{X}_{\tau_1,\tau_2})$ and $\widehat{X}_{\tau_1,\tau_2}$ is the quantization output of a three-level ADC with thresholds $(-\infty,\tau_1,\tau_2,\infty)$ for input X (see Equation (1)).

2) Scenario 2. Quadratic Analog Operators: In this scenario, we choose $f_a : \mathbb{R} \to \mathbb{R}$ from the set of all quadratic functions, i.e., $\mathcal{F}_{a,j} = \{f : \mathbb{R} \to \mathbb{R} | f(x) = ax^2+bx+c, a, b, c \in \mathbb{R}\}, j \in [\ell]$. Let $\tau_1 \leq \tau_2 \leq \tau_3$ be arbitrarily chosen real numbers. Define $f_1(x) = (x-\tau_1)(x-\tau_3)$ and $f_2(x) = (x-\tau_2)$. In this case, $W_1 = (X-\tau_1)(X-\tau_3)$ and $W_2 = X-\tau_2$ are the ADC inputs. We set the ADC thresholds to zero, so that $\widehat{W}_1 = \mathbb{1}(X \in [\tau_1,\tau_3])$ and $\widehat{W}_2 = \mathbb{1}(X \in [\tau_2,\infty])$. Thus, receiving $\widehat{W}_1$ and $\widehat{W}_2$ is equivalent to receiving the quantization output for quantizing X with a four-level ADC with thresholds τ_1,τ_2,τ_3. Consequently,

$$d^*_{1,quad} = \min_{\tau_1,\tau_2,\tau_3:\tau_1<\tau_2<\tau_3} \mathbb{E}_{S,N}((S-\widehat{S})^2), \tag{4}$$

where $\widehat{S} = \mathbb{E}(S|\widehat{X}_{\tau_1,\tau_2,\tau_3})$ and $\widehat{X}_{\tau_1,\tau_2,\tau_3}$ is the quantization output of a four-level ADC with thresholds $(-\infty,\tau_1,\tau_2,\tau_3,\infty)$ for input X. Note that this is an improvement over the achievable distortion of Scenario 1. In fact, to achieve $d^*_{1,quad}$ using linear analog processing, one needs to use three one-bit ADCs instead of two one-bit ADCs, thus requiring a fifty percent increase in ADC power consumption.

3) Scenario 3. Envelope Detectors: In this scenario, we assume the quantization system is equipped with envelope detectors, which can perform absolute value operations on the analog signal. Let $\tau_1 \leq \tau_2 \leq \tau_3$ be arbitrarily chosen real numbers. Define $f_1(x) = |x - \frac{\tau_1+\tau_3}{2}|$ and $f_2(x) = x$. Furthermore, let the ADC thresholds be $t_1 = \frac{\tau_3-\tau_1}{2}$ and $t_2 = \tau_2$. Then,

$$\widehat{W}_1 = \mathbb{1}(|X - \frac{\tau_1+\tau_3}{2}| < \frac{\tau_3-\tau_1}{2}) = \mathbb{1}(X \in [\tau_1,\tau_3]),$$
$$\widehat{W}_2 = \mathbb{1}(X > \tau_2).$$

Consequently, the achievable distortion is equal to that of Scenario 2, and improves the distortion in Scenario 1. In general the use of polynomial operators (Scenario 2) leads to lower achievable distortion compared to envelope detectors (Scenario 3), however the circuit design of envelope detectors is more straightforward than that of polynomial operators [24], hence there is a trade-off between design complexity and achievable distortion between these two scenarios.

It can be noted that in Scenarios 1-3, since f_a is memoryless, and its output at time j only depends on the input at time j, the minimum ℓ-shot achievable distortion is equal to the minimum one-shot achievable distortion for all $\ell \in \mathbb{N}$.

4) Scenario 4. Analog Delay Elements: In this scenario, we consider the use of analog delay elements, which allows for causal memory in the analog processing functions. That is, we consider a processing function at time $j \in \mathbb{N}$ which is an affine function of the form $f_{a,j} : \mathbb{R}^j \to \mathbb{R}$ and $f_{a,j}$ takes $X^{m\times j}$ as input. The two-shot minimum achievable distortion is given by:

$$d^*_{2,delay} = \min_{\tau_1,\tau_2,\tau_3,\tau_4,a_1,a_2} \frac{1}{2}\sum_{j=1}^{2} \mathbb{E}((S_j - \widehat{S}_j)^2), \tag{5}$$

where $\widehat{S}_j = \mathbb{E}_{S_j|X_1,X_2}(S_j|\widehat{X_1},\widehat{X_2})$ and $\widehat{X_1}$ is the quantization output of a three-level ADC with thresholds $(-\infty, \tau_1, \tau_2, \infty)$ for input X_1 and $\widehat{X_2}$ is the quantization output of a three-level ADC with thresholds $(-\infty, \tau_3, \tau_4, \infty)$ for input $a_1X_1 + a_2X_2$. It should be noted that the optimization in this scenario is over a larger search space compared to that of Scenario 1, as it allows for two-dimensional quantization, in the (X_1, X_2) space rather than only the X_2 space, in the second time-slot. The optimization reduces to that of Scenario 1 by restricting to $a_1 = 0, a_2 = 1$. This achievable distortion is numerically evaluated in Section V. We show that in this simple scenario, the gains due to the additional delay element are negligible compared to Scenario 1. However, if the use of delay elements is further augmented by analog polynomial operators, then we achieve significant gains over the previous three scenarios.

IV. Fundamental Performance Limits in Task-Based Quantization

A. Finite-degree Polynomial Operators and Delay Elements

We consider a setup equipped with finite-degree polynomial operators with delay elements. That is, we consider the following set of implementable functions:

$$\mathcal{F}^t_{a,j} = \{f(\cdot)|f(x^{m\times j}) = \sum_{\substack{(k_{u,v}, u\in[m], v\in[j]):\\ \sum_{u,v} k_{u,v}\leq t}} b_{k^{m\times j}} \prod_{v\in[m], u\in[j]} x^{k_{u,v}}, b_{k^{m\times j}} \in \mathbb{R}\},$$

$$\mathcal{F}_{a,j} = \cup_{t\in\mathbb{N}} \mathcal{F}^t_{a,j}, \quad j \in \mathbb{N}.$$

Theorem 1. *Consider a task-based quantization setup parametrized by $(n, m, P_{S^n}, P_{X^m|S^n}, (\mathcal{F}_{a,j})_{j\in\mathbb{N}}, n_q, \kappa, d(\cdot,\cdot))$ as described in the prequel. Assume that there exists $\mathbf{s} \in \mathbb{R}^m$ such that $\mathbb{E}(d(S^m, \mathbf{s})) \leq \infty$. The minimum achievable ℓ-shot distortion for asymptotically large ℓ is given by:*

$$\lim_{\ell\to\infty} d^*_\ell = \min_{P_{\widehat{S}^m|X^n}: I(X^n;\widehat{S}^m)\leq n_q} \mathbb{E}_{S,X}(d(S^n, \widehat{S}^n)), \qquad (6)$$

where $P_{S^m,X^n,\widehat{S}^m} \triangleq P_{S^m,X^n}P_{\widehat{S}^m|X^n}$, i.e., the Markov chain $S^n \leftrightarrow X^m \leftrightarrow \widehat{S}^n$ holds.

The distortion is then equal to the indirect distortion-rate function (iDRF) evaluated at compression rate n_q bits per input symbol. The proof follows by noting that using the multivariate Taylor expansion, any quantizer used for indirect source coding can be well-approximated, with arbitrary precision, using a finite-degree polynomial. Consequently, the optimal quantization scheme achieving the iDRF can be implemented using the analog processing functions, and its output (bits) can be passed through the ADCs without any further modification on the digital side. That is, the analog processing function is chosen such that its output is equal to that of the optimal compression function in the equivalent indirect source coding problem. Note that the output of the optimal compression function is binary, hence by setting the ADC thresholds equal to $\frac{1}{2}$, the binary analog processing outputs are recovered without further distortion on the digital side. The complete proof is given in [27].

B. Memoryless Finite-degree Polynomial Operators

Implementing large analog delay elements may not be practically possible due to synchronization and chip space limitation issues. In this section, we consider a task-based quantization setup equipped with finite-degree polynomial operators without delay elements:

$$\mathcal{F}^t_{a,j} = \{f(\cdot)|f(x^m) = \sum_{\substack{(k_u, u\in[m]):\\ \sum_u k_u\leq t}} b_{k^m} \prod_{v\in[m]} x^{k_u}, b_{k^m} \in \mathbb{R}\}, j \in \mathbb{N}.$$

Note that this can be considered as the one-shot version of the scenario considered in Section IV-A.

Theorem 2. *Consider a task-based quantization setup parameterized by $(n, m, P_{S^n}, P_{X^m|S^n}, (\mathcal{F}_{a,j})_{j\in\mathbb{N}}, n_q, \kappa, d(\cdot,\cdot))$. The minimum achievable distortion is given by:*

$$d^*_\ell = \min_{\substack{f:\mathbb{R}^m\to[\kappa^{n_q}]\\ g:[\kappa^{n_q}]\to\mathbb{R}^n}} \mathbb{E}_{S,X}(d(S^n, \widehat{S}^n)), \qquad (7)$$

for all $\ell \in \mathbb{N}$, where $\widehat{S} \triangleq g(f(X)))$.

The proof follows by similar arguments as that of Theorem 1. We provide an outline in the following. We first note that since the system is not equipped with delay elements, the reconstruction at time j only depends on the input at time j. Consequently, the ℓ-shot minimum achievable distortion is the same for all values of $\ell \in \mathbb{N}$. Hence, it suffices to consider the one-shot distortion. Furthermore, the ADCs can produce at most κ^{n_q} Voronoi regions, which implies that the right-hand-side term in (7) is a lower-bound for the achievable distortion. On the other hand, similar to the proof of Theorem 1, using the multi-variate version of Taylor's approximation, any quantizer with κ^{n_q} Voronoi regions can be constructed using finite-degree polynomials and n_q ADCs each with κ quantization levels. This implies that the right-hand-side term in (7) is an upper-bound for the achievable distortion.

C. Low-Degree Polynomials without Delay Elements

It is shown in [10], [24], that although the power consumption of low-degree polynomial operators such as quadratic operators may be significantly smaller than that of ADC components, the power consumption grows with polynomial degree, and becomes significant for high-degree polynomials. As a result, in this section we focus on the use of low degree polynomial operators with no delay elements. To derive computable, closed-form expressions for the achievable distortion, we focus on the scalar measurements and one-bit ADCs, i.e., $m = 1, \kappa = 2$. We consider the set of implementable analog functions $\mathcal{F}^\delta_{a,j} = \{f(\cdot)|f(x) = \sum_{i=0}^{\delta} a_i x^i, a_i \in \mathbb{R}\}$, where $\delta \in \mathbb{N}$ is the maximum polynomial degree. The following characterizes the minimum achievable distortion in this scenario.

Theorem 3. *Consider a task-based quantization setup parameterized by $(n, 1, P_{S^n}, P_{X^m|S^n}, (\mathcal{F}^\delta_{a,j})_{j\in\mathbb{N}}, n_q, 2, d(\cdot,\cdot))$ as described in the prequel. Then,*

$$d^*_\ell = \min_{\substack{(\tau_i)_{i\in[\Gamma]}\\ g:[\Gamma+1]\to\mathbb{R}^n}} \mathbb{E}_{S,X}(d(S^n, g(\widehat{X}))),$$

where $\widehat{X}$ is the quantization output of a $(\Gamma+1)$-level ADC with thresholds $(-\infty, \tau_1, \tau_2, \cdots, \tau_\Gamma, \infty)$ and input X, and

$$\Gamma \triangleq \min(2^{n_q}, \Gamma') \qquad \Gamma' \triangleq \begin{cases} n_q\delta + 1 & \text{if } \delta \text{ is odd,} \\ n_q\delta & \text{otherwise} \end{cases}.$$

Note that since the polynomial operators may have a constant non-zero bias, we may assume without loss of generality that the ADCs have zero thresholds, and incorporate the thresholds into the polynomial bias. Then, the proof of the theorem follows by noting that the output of the ADC changes at the roots of the polynomial operator. Each polynomial operator of degree δ has at most δ distinct roots, and since there are n_q operators, they may have at most δn_q different roots. On the other hand, for even-degree polynomials, the value for asymptotically large negative and positive inputs are the same, hence the ADC output is equal for both. Consequently, there are at most Γ' different quantization Voronoi regions as a result of the ADC operation. The complete proof is provided in [27]. It should be noted that the result may be generalized to $\kappa > 2$ by using Proposition 4 in [24] to characterize the Voronoi regions.

D. Envelope Detectors without Delay elements

As shown in the circuit design and simulations of [24], implementing envelope detectors to produce absolute value functions is less costly in terms of circuit design complexity and power consumption, compared to polynomial operators. Consequently, in this section, we consider

$$\mathcal{F}^{\delta}_{a,j} = \{f(y) = A_s(x, b^s), x \in \mathbb{R} | s \in [\delta], a^s \in \mathbb{R}^s\}, \quad \delta \in \mathbb{N},$$

where $A_1(x, b) \triangleq |x - b|, x, b \in \mathbb{R}$ and $A_s(x, b^s) \triangleq A_1(A_{s-1}(x, b^{s-1}), b_s) = |A_{s-1}(x, b^{s-1}) - b_s|, s \in \mathbb{N}$. That is, $\mathcal{F}^{\delta}_{a,j}$ consists of all functions which can be generated using sequences of $s \leq \delta$ concatenated envelope detectors with thresholds $b_1, b_2, \cdots, b_s$, respectively.

Definition 1 (**Fully-Symmetric Vector**). *A vector* $\mathbf{b} = (b_1, b_2, \cdots, b_{2^n})$ *is called symmetric if* $b_i + b_{2^n-i} = b_j + b_{2^n-j}, i, j \in [2^n - 1]$. *The vector* $\mathbf{b}$ *is called fully-symmetric if it is symmetric and the vectors* $(b_1, b_2, \cdots, b_{2^{n-1}})$ *and* $(b_{2^{n-1}+1}, b_{2^{n-1}+2}, \cdots, b_{2^n})$ *are both fully-symmetric for* $n > 2$ *and symmetric for* $n = 2$.

Theorem 4. *Consider a task-based quantization setup parameterized by* $(n, 1, P_{S^n}, P_{X^m|S^n}, (\mathcal{F}^{\delta}_{a,j})_{j\in\mathbb{N}}, n_q, 2, d(\cdot,\cdot))$ *as described in the prequel. Then,*

$$d^*_\ell = \min_{\substack{(\tau_i)_{i\in[\Gamma]}\in\mathcal{S} \\ g:[\Gamma+1]\to\mathbb{R}^n}} \mathbb{E}_{S,X}(d(S^n, g(\widehat{X}))),$$

where $\Gamma \triangleq \min(2^{n_q}, n_q 2^\delta)$, *and* $\mathcal{S}$ *consists of the set of all vectors of length* $n_q 2^\delta$, *which can be partitioned into* n_q *fully-symmetric subvectors, each of length* 2^δ.

The proof follows by similar arguments as that of Theorem 3 and [24, Proposition 5].

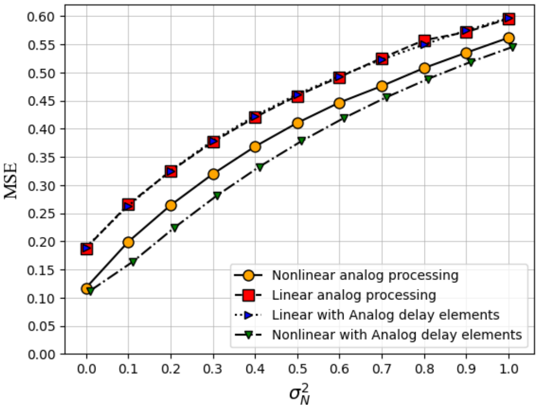

Fig. 2. Comparison of MSE distortion for linear and nonlinear analog processing with and without analog delay elements for a jointly Gaussian scalar task S and measurement X, with $n_q = \kappa = 2$.

V. Simulation Results

Let us consider the task-based quantization setup considered in Section III. Figure 2 provides a numerical evaluation of the achievable distortion in this setup under each of the scenarios considered in Section IV. The linear analog processing plot (red square markers) shows the achievable distortion when only linear analog processing is used without delay element, and it serves as a baseline for the other schemes. It is derived by evaluating Equation (3) and sweeping over all possible values of τ_1, τ_2 with step-size 0.01. The linear processing with delay elements plot (blue triangle markers) is derived by evaluating Equation (5) by sweeping over values of $\tau_1, \tau_2, \tau_3, \tau_4, a_1, a_2$. It can be observed that in this simple scenario, the use of a single delay element while restricting to linear processing does not lead to a tangible performance improvement. The non-linear analog processing plot (orange circle markers) shows the achievable performance when quadratic polynomial operators are used without delay elements. It is derived by optimizing Equation (4). It can be observed that the use of quadratic operators improves the achievable distortion over the baseline. Lastly, the non-linear analog processing with delay elements plot (green triangle markers) shows the performance when polynomial operators with arbitrary degree and arbitrary number of delay elements can be used. It is derived by optimizing Equation (6). This serves as an outer-bound for the achievable distortion in the previously mentioned scenarios as it considers the most general subset of implementable analog functions.

VI. Conclusions

The use of non-linear analog processing prior to quantization using low resolution ADCs in the task based quantization problem was studied. Several classes of non-linear analog processors were considered including analog delay elements, polynomial operators, and envelope detectors. In each scenario, the minimum achievable distortion was characterized and it was shown that the use of non-linear processing improves the achievable distortion. Simulations of a Gaussian task-based quantization setup were provided to illustrate these gains.

References

[1] B. Razavi, *Principles of data conversion system design*. IEEE press New York, 1995, vol. 126.

[2] B. Murmann, "The race for the extra decibel: a brief review of current ADC performance trajectories," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 3, pp. 58–66, 2015.

[3] Y. Chi and H. Fu, "Subspace learning from bits," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4429–4442, 2017.

[4] R. M. Corey and A. C. Singer, "Wideband source localization using one-bit quantized arrays," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.

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

[6] 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)*, 2019, pp. 390–394.

[7] 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.

[8] 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.

[9] N. I. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Capacity bounds for one-bit mimo gaussian channels with analog combining," *IEEE Transactions on Communications*, vol. 70, no. 11, pp. 7224–7239, 2022.

[10] F. Shirani and H. Aghasi, "Quantifying the capacity gains in coarsely quantized siso systems with nonlinear analog operators," in *2022 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2022, pp. 522–527.

[11] ——, "Mimo systems with one-bit adcs: Capacity gains using nonlinear analog operations," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 2511–2516.

[12] Y. C. Eldar, *Sampling theory: Beyond bandlimited systems*. Cambridge University Press, 2015.

[13] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on selected areas in communications*, vol. 17, no. 4, pp. 539–550, 1999.

[14] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, "On low-resolution adcs in practical 5g millimeter-wave massive mimo systems," *IEEE Communications Magazine*, vol. 56, no. 7, pp. 205–211, 2018.

[15] N. I. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Design and analysis of hardware-limited non-uniform task-based quantizers," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1551–1562, 2023.

[16] 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)*, 2022, pp. 5338–5342.

[17] 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)*, 2017, pp. 599–603.

[18] 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.

[19] 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)*, 2021, pp. 5514–5518.

[20] N. Shlezinger and Y. C. Eldar, "Deep task-based quantization," *Entropy*, vol. 23, no. 1, p. 104, 2021.

[21] D. Malak, R. Yazicigil, M. Médard, X. Zhang, and Y. C. Eldar, "Hardware-limited task-based quantization in systems," in *Women in Telecommunications*. Springer, 2023, pp. 105–163.

[22] M. Shohat, G. Tsintsadze, N. Shlezinger, and Y. C. Eldar, "Deep quantization for mimo channel estimation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3912–3916.

[23] N. I. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Hardware-limited non-uniform task-based quantizers," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[24] M. T. Alonso, X. Liu, F. Shirani, and H. Aghasi, "Capacity gains in mimo systems with few-bit adcs using nonlinear analog operators," 2022.

[25] 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*, 2021.

[26] A. Khalili, S. Rini, L. Barletta, E. Erkip, and Y. C. Eldar, "On MIMO channel capacity with output quantization constraints," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1355–1359.

[27] M. T. Alonso, F. Shirani, N. I. Bernardo, and Y. C. Eldar, "Nonlinear analog processing gains in task-based quantization," *arXiv preprint arXiv:2402.01525*, 2024.