Information Theoretic Threshold Tuning in Parallel Stochastic Quantizer Architectures ⋆

Hassan Tavakoli
School of EECS
Oregon State University
Oregon, OR 97331, USA
tavakolh@oregonstate.edu

Thinh Nguyen, *Senior Member, IEEE*
School of EECS
Oregon State University
Oregon, OR 97331, USA
thinhq@eecs.oregonstate.edu

Bella Bose, *Life Fellow, IEEE*
School of EECS
Oregon State University
Oregon, OR 97331, USA
Bella.Bose@oregonstate.edu

***Abstract*—Quantization plays a central role in digital communication by mapping continuous-valued signals to a finite set of levels with minimal distortion. Beyond mean-square error, mutual information between the channel input and the quantizer output provides a powerful metric for signal recovery. However, finding the quantizer that maximizes the mutual information is NP-complete for non-binary inputs. To that end, while not optimal, thresholding schemes, whether single-threshold or multi-threshold, are widely adopted. In this work, we study the parallel stochastic single-threshold quantizer architecture, provide some information-theoretic insights, and introduce a momentum-accelerated gradient ascent algorithm that efficiently tunes a single decision threshold to maximize the mutual information. We demonstrate convergence improvements over exhaustive search and quantify mutual information gains across binary and non-binary input distributions. We also validate our theoretical framework with simulations on the MNIST dataset, demonstrating that increasing the number of parallel quantization branches, i.e., mutual information, significantly improves classification accuracy, especially when quantization thresholds are learned and training data is limited.**

Quantization, Mutual information, Threshold optimization, Momentum gradient ascent, Parallel noisy-adder architecture

I. Introduction

A quantizer maps continuous signals to a finite set of levels and is therefore fundamental in digital communications. Traditional metrics use average distortion such as mean-square error (MSE). From an information-theoretic viewpoint, mutual information (MI) between the channel input and quantizer output is a more general metric: it does not assume continuity of the signals and directly measures the statistical dependence between original and quantized signals, so higher MI implies better recoverability. [1]–[7]

Prior work has attacked quantizer design from several angles. Kurkoski *et al.* cast quantization as impurity minimization using statistical learning ideas [2]. Nguyen and colleagues derived optimal binary quantizers with multiple thresholds and characterized capacity-achieving quantizers and input distributions for binary channels [8], [9], extending the robust quantization framework of [10]. Dulek used convex and decision-theoretic tools to show the existence and geometric structure of optimal quantizers in the continuous-input case [6]. Interestingly, additive noise can sometimes improve information transfer in nonlinear quantizers [11].

Finding the MI-maximizing quantizer is computationally hard for non-binary inputs (NP-complete). [12] Thus practical methods usually restrict the search to threshold quantizers—single thresholds for one-bit designs [9], [13] or multi-threshold/sequential schemes for larger alphabets [14].

Building on [10], this paper provides information-theoretic insight and proposes a momentum-accelerated gradient-ascent heuristic to optimize a single-threshold quantizer for maximum mutual information. We empirically show that learning thresholds on MNIST improves classifier performance versus non-optimal thresholds, and we study how classification accuracy depends on the number of quantization bits (which correlates with MI).

The paper is organized as follows. Section II gives motivation and the system model. Section III presents theoretical results on the optimal threshold and MI for a parallel stochastic quantizer architecture. Because exact optimization is hard, Section IV develops our momentum-accelerated gradient-ascent algorithm. Section V evaluates the algorithm across input distributions and shows gains from adding parallel quantization branches and from learned thresholds, especially with limited training data. Section VI concludes and outlines future work.

II. System Model and Motivation

A. Motivation

The system, Fig.1, uses m parallel, identical stochastic comparators. X is the input (discrete or continuous), Z_i are i.i.d. continuous additive noises, $Q_\theta(\cdot)$ are identical single-threshold comparators, and $Y_i \in \{0, 1\}$ are the outputs. The sum $Y = \sum_i Y_i$ is a sufficient statistic for X and serves as the quantized value.

A drawback of this architecture is the larger comparator count compared with conventional ADCs (e.g., flash, pipelined, sigma–delta). [15], [16] Sigma–delta designs use fewer comparators but need high oversampling; pipelined ADCs use moderate comparator counts but are slower because of sequential approximation. By contrast, the parallel stochastic design uses uniform reference voltages (so it avoids large resistor loads), which simplifies the circuit and enables faster

⋆ This work was supported by the National Science Foundation Grant CCF-2417898.

operation, and it is inherently robust in harsh environments (e.g., radiation) where comparator/op-amp failures are more likely. This architecture is well-suited to low-power sensing and edge computing. Multiple low-resolution (1-bit) comparators operating in parallel consume little power yet can capture enough information for downstream tasks. We show that a neural-network classifier trained on these quantized outputs can reach high accuracy using only a small number of parallel quantizers.

B. System Model

Let X be a random variable denoting the input signals. X can be a discrete or continuous random variable. However, to simplify the discussion, let X be a discrete random variable, taking values $\{x_1, x_2, \ldots, x_N\}$ with probability of $\Pr(X = x_i) = q_i$. Let Z_k, $k = 0, 1, \ldots m$ be independent identical random noise generated by the system. In each branch, the input X is corrupted by an additive noise Z_k. The m resulting signals $X + Z_k$ is then passed through a comparator $Q_\theta(\cdot)$, parameterized by a tunable threshold θ, to produce $Y_k \in \{0, 1\}$. Specifically,

$$Y_k = \begin{cases} 1, & \text{if } X + Z_k \geq \theta \\ 0, & \text{if } X + Z_k < \theta \end{cases} \tag{1}$$

Thus, the probability of $Y_k = 1$ is when $Z_k \geq \theta - x_i$, :

$$q(x_i) = \Pr(Y_k = 1 | X = x_i) = 1 - F_Z(\theta - x_i), \tag{2}$$

where $F_Z(.)$ denotes the cumulative distribution of the generated noise.

These outputs are then aggregated through a final summation block to yield the overall system output $Y = \sum_{j=1}^{m} Y_j$. Thus, the conditional probability of $Y = j$ given $X = x_i$ is:

$$\Pr(Y = j \mid X = x_i) = \binom{m}{j} (q(x_i))^j (1 - q(x_i))^{m-j}. \tag{3}$$

The mutual information between the input signals X and the quantized signals Y is given as:

$$\begin{aligned} I(X;Y) = &\sum_{i=1}^{m} \sum_{j=1}^{n} \Pr(X = x_i, Y = j) \\ &\times \log \left(\frac{\Pr(X = x_i, Y = j)}{\Pr(X = x_i) \Pr(Y = j)} \right) \end{aligned} \tag{4}$$

where $\Pr(Y = j) = \sum_{i=1}^{n} \Pr(X = x_i, Y = j)$ and log denotes $\log_2$.

III. Mutual Information and Optimal Threshold

A parallel stochastic quantizer is a discrete memoryless channel with input X and output Y; each threshold θ induces a specific channel $p(y \mid x; \theta)$. The design task is to choose θ to maximize the mutual information $I(X;Y)$. Recall that for fixed $p(x)$, $I(X;Y)$ is convex in the channel transition matrix $p(y \mid x)$, while for fixed $p(y \mid x)$ it is concave in $p(x)$. But since θ only indirectly determines $p(y \mid x)$, $I(X;Y)$ need not be concave (or otherwise well-behaved) as a function of θ.

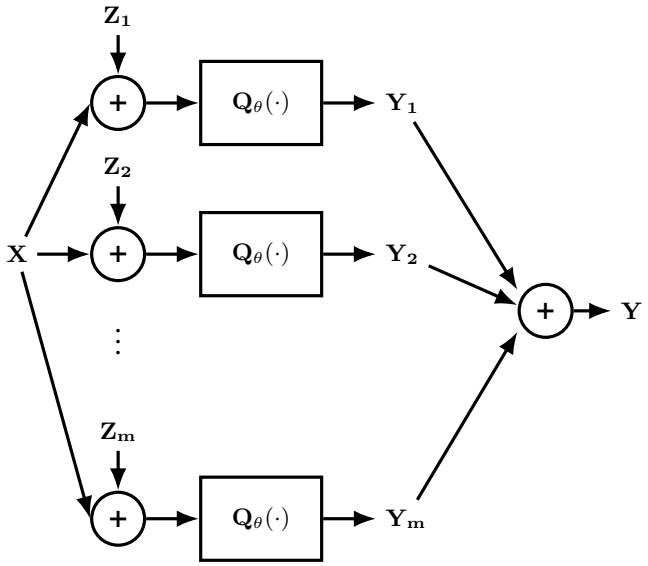

Fig. 1. Parallel stochastic quantizer architecture

This non-concavity makes direct optimization over θ difficult, as formalized in the following proposition.

Proposition 1 (Mutual Information is not concave in θ:): For any fixed branch count m, the mutual information $I(X;Y)$ is not a concave function of the threshold parameter θ.

Proof: See Appendix.

On the other hand, in certain special cases, we can find the optimal threshold θ^* analytically as stated in the proposition below.

Proposition 2 (Symmetric Threshold Optimality): If the input alphabet $\mathcal{X}$ is symmetric about zero and Z_k are i.i.d. zero-mean symmetric noise, then the unique maximizer of $I(X;Y)$ is $\theta^* = 0$.

Proof: See Appendix.

Theorem 1: Let $Y = \sum_{k=1}^{m} Y_k$, where each $Y_k \mid X = x_i \sim \text{Bernoulli}(q(x_i))$. As $m \to \infty$, the distribution of Y conditioned on $X = x_i$ converges in distribution to a normal distribution:

$$Y \xrightarrow{d} \mathcal{N}\big(mq(x_i),\, mq(x_i)(1 - q(x_i))\big).$$

In the special case where $X \sim \mathcal{U}[-\alpha, \alpha]$ and each $Z_i \sim \mathcal{U}[-\alpha, \alpha]$ is drawn independently, and with $\theta = 0$, consider the estimator

$$\hat{X}_m = \frac{1}{m} \sum_{k=1}^{m} Y_k$$

used to estimate X. Then the following results hold:

$$\hat{X}_m \xrightarrow{m.s.} X \text{ on the order } O\left(\frac{1}{m}\right)$$

$$I(X; \hat{X}_m) \xrightarrow{p} H(X)$$

Proof: See Appendix.

Theorem 1 shows that, in the special case described above, the mean-square error estimation of X by using $Y_1, Y_2, \ldots, Y_m$ is reduced to zero as m approaches infinity on the order of $\frac{1}{m}$. As a result, the mutual information between Y and X achieves the maximum, i.e., equals to $H(X)$.

IV. Proposed Algorithm

Because the exact problem is generally intractable, we use a momentum-accelerated gradient-ascent heuristic to approximate the optimal threshold. This method does not guarantee the global optimum, but empirically, it often finds near-optimal solutions. We compare two strategies: Gradient Ascent without momentum (GA) and Gradient Ascent with Momentum (GAM). Momentum is motivated by the nonconvex/nonconcave shape of $I_\theta(X;Y)$ in θ; empirically it helps escape poor local maxima and improves convergence.

Algorithm 1 Threshold Optimizer Algorithm: TOA(η)

Require: Discrete support $X = (x_i)_{i=1}^n$, probabilities $P_X = \big(P_X(x_i)\big)_{i=1}^n$ with $\sum_i P_X(x_i) = 1$; Gaussian noise $Z \sim \mathcal{N}(\mathbb{E}(Z), \sigma^2)$; Initial threshold $\theta^0 = \mathbb{E}(X)$, binomial trials m; Learning rate μ, momentum η; Maximum iterations $K_{\max}$, tolerance ε

Initialize $\gamma_0 \leftarrow 0$, $\theta \leftarrow \theta^0$
for $k = 0$ to $K_{\max} - 1$ **do**
  // Step 1: Compute q_i for current θ
  for $i = 1$ to n **do**
    $q_i \leftarrow 1 - \Phi\left(\dfrac{\theta - x_i - \mathbb{E}(Z)}{\sigma}\right)$
  end for
  // Step 2: Compute $P(Y \mid X)$ and marginalize to P_Y
  for $i = 1$ to n **do**
    for $j = 0$ to m **do**
      $P(Y = j \mid X = x_i) \leftarrow \binom{m}{j} q_i^j (1 - q_i)^{m-j}$
    end for
  end for
  for $j = 0$ to m **do**
    $P_Y(j) \leftarrow \sum_{i=1}^n P_X(x_i) \cdot P(Y = j \mid X = x_i)$
  end for
  // Step 3: Compute mutual information $I(\theta)$
  $I(\theta) \leftarrow \sum_{i=1}^n \sum_{j=0}^m P_X(x_i) \cdot P(Y = j \mid X = x_i) \cdot \log\left(\dfrac{P(Y = j \mid X = x_i)}{P_Y(j)}\right)$
  // Step 4: Using gradient
  Compute $\dfrac{\partial I_\theta(X;Y)}{\partial \theta}$ using Eq. (6)
  // Step 5: Momentum update
  $\gamma_k \leftarrow \eta \cdot \gamma_{k-1} + \mu \cdot \dfrac{\partial I_\theta(X;Y)}{\partial \theta}$
  $\theta \leftarrow \theta + \gamma_k$
  if $|\gamma_k| < \varepsilon$ **then**
    break
  end if
end for
return $\theta^* = \theta$, $I^* = I(\theta^*)$

We package the approach as the Threshold Optimizer Algorithm with momentum η, denoted TOA(η). When $\eta = 0$, TOA reduces to GA; when $\eta \neq 0$ it becomes GAM, so a single parameter unifies both methods. The TOA seeks θ that maximizes the mutual information between a discrete input X (with mass function $P_X(i)$) and a binomial output Y. For each candidate θ we compute q via Eq. (2) (using $Z \sim \mathcal{N}(\mathrm{E}(Z), \sigma_Z^2)$), then the transition probabilities from Eq. (3), the marginals $\Pr(Y = j)$, and the mutual information in Eq.(4). If the stopping condition is not met, we estimate the derivative $\partial I/\partial\theta$ using Eq.(6) and update θ with the momentum update rule.

$$\gamma_k = \eta\,\gamma_{k-1} + \mu\,\frac{\partial I_\theta(X;Y)}{\partial \theta}, \qquad \theta^{(k+1)} = \theta^{(k)} + \gamma_k. \tag{5}$$

Iteration continues until $|\gamma_k| < \varepsilon$ or a maximum number of iterations is reached, yielding the optimized θ^* and its mutual information $I(\theta^*)$.

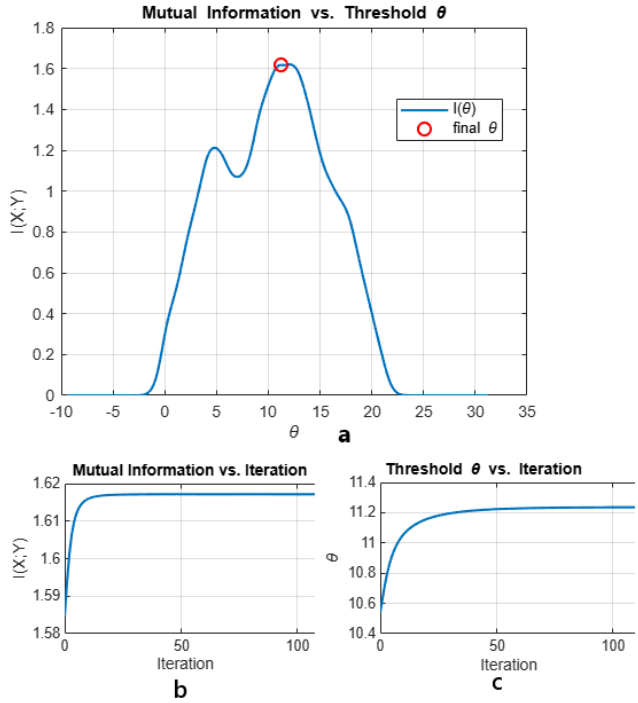

Fig. 2. a) Mutual information vs. threshold θ b) Mutual information vs. iteration, and c) θ vs. iterations for the second setting.

V. Simulation Results

A. Performance Evaluation of TOA (η)

In the simulation results section, we present two different setups based on the proposed Algorithm 1, TOA (η). In both cases, the parameters are set as follows: $m = 10$, $\mu = 0.5$, momentum $\eta = 0.5$, maximum number of iterations $K_{\max} = 50$, and tolerance $\varepsilon = 10^{-3}$.

In the first setting, we set x_i to be $n = 50$ values in the interval $[-10, 10]$ uniformly. Noise $Z \sim \mathcal{N}(1, 1)$. In the

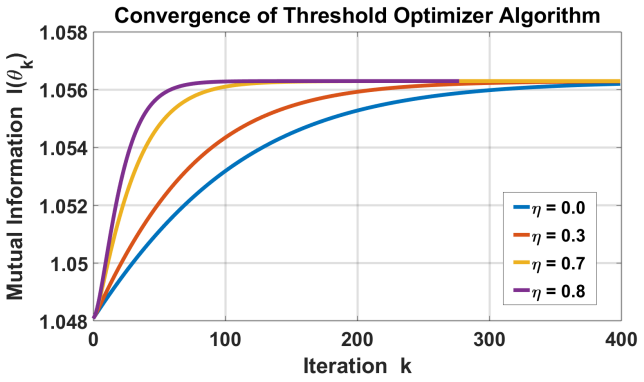

Fig. 3. Convergence comparing of TOA(η) for $\eta \in \{0,\ 0.3,\ 0.7,\ 0.8\}$.

second setting, x_i are now shifted by 10, i.e., $x_i \in [0, 20]$, and the noise Z is shifted by one, i.e., $Z \sim \mathcal{N}(0, 1)$. Similarly to setting one, Fig. 2(a), (b), and (c) show that the proposed algorithm converges quickly to the optimal θ^*.

As shown in Fig. 3, by using first setting parameters, increasing the momentum coefficient η in TOA(η) significantly accelerates convergence. We evaluated four settings: $\eta \in \{0,\ 0.3,\ 0.7,\ 0.8\}$. The case $\eta = 0$ (i.e., standard Gradient Ascent) exhibits the slowest convergence, while the fastest convergence is obtained for $\eta = 0.8$. These results confirm that adding momentum substantially improves the optimizer's speed and robustness.

B. Classifier vs. Mutual Information

In this section, we evaluate the proposed parallel quantization model using the MNIST dataset of handwritten digits. In Fig. 4, each image, originally 28×28 pixels, is first flattened into a 784-dimensional vector. The image is then processed through a bank of m parallel quantization branches. In each branch, i.i.d. uniform noise $Z_i \sim \mathcal{U}(-128, 128)$ is added independently, and a quantizer Q_θ is applied. The outputs from all quantizers, $Y_1, \ldots, Y_m$, are concatenated and used as input to a convolutional neural network (CNN) for digit classification. The architecture and other setups of the neural network, which consists of two hidden layers with 80 and 60 neurons respectively, are such that it achieves an accuracy of approximately 97 $\%$ when trained on the 60,000-sample MNIST training set and evaluated on the 10,000-sample test set.

To examine the effects of both redundancy (m) and training data size (TSS), we vary the number of parallel quantization branches m from 1 to 10 and the training sample size between 1,000 and 60,000. Figure 5 shows the configuration with learned thresholds via our TOA ($\eta = 0.7$), while Figure 6 uses fixed thresholds ($\theta = 128$) for comparison. In both cases, classification accuracy improves as m increases, confirming Theorem 1's prediction that more branches raise the mutual-information bound and thus enrich the representation available to the classifier. Moreover, when thresholds are learned, increasing the training set size further amplifies this benefit: not only does a larger m yield bigger gains, but each doubling of data produces a noticeable lift in accuracy. By contrast, the fixed-threshold system appears to plateau quickly and even degrade slightly with more data, suggesting overfitting to suboptimal thresholds. These results underscore that optimizing thresholds based on mutual information both magnifies the impact of redundancy and unlocks the full value of additional training samples.

VI. Conclusion

We have provided some information-theoretic insights and proposed a fast algorithm for maximizing mutual information in a parallel stochastic quantizer architecture. The algorithm optimizes a single threshold using a momentum-accelerated gradient ascent method. Simulation results confirm that the proposed algorithm achieves faster convergence and higher $I(X;Y)$ compared to brute-force grid search. Future research will expand this methodology to multi-threshold arrays and incorporate adaptive learning of input distributions, with the ultimate aim of bridging the gap to capacity-achieving quantizers [6], [7], [9]. Our simulation results confirm that increasing the number of quantization branches and learning the quantization thresholds both contribute to better performance, in line with our information-theoretic analysis, particularly under limited data conditions.

References

[1] Robert M. Gray and David L. Neuhoff. Quantization. *IEEE transactions on information theory*, 44(6):2325–2383, 1998.

[2] Brian M. Kurkoski and Hideki Yagi. Quantization of binary-input discrete memoryless channels. *IEEE Transactions on Information Theory*, 60(8):4544–4552, Aug 2014.

[3] Ken-ichi Iwata and Shin-ya Ozawa. Quantizer design for outputs of binary-input discrete memoryless channels using smawk algorithm. In *2014 IEEE International Symposium on Information Theory*, pages 191–195. IEEE, 2014.

[4] Alankrita Bhatt, Bobak Nazer, Or Ordentlich, and Yury Polyanskiy. Information-distilling quantizers. *IEEE Transactions on Information Theory*, 67(4):2472–2487, 2021.

[5] T. D. Nguyen and T. Nguyen. On binary quantizer for maximizing mutual information. *IEEE Transactions on Communications*, 68(9):5435–5445, Sep 2020.

[6] Bartłomiej Dulek. On mutual information-based optimal quantizer design. *IEEE Communications Letters*, 26(5):1008–1011, 2022.

[7] Qiqing Zhai and Youguo Wang. A mutual information-maximizing quantizer based on the noise-injected threshold array. *Digital Signal Processing*, 146:104394, 2024.

[8] T. Nguyen and T. Nguyen. Capacity achieving quantizer design for binary channels. *IEEE Communications Letters*, 25(3):759–763, 2021.

[9] T. Nguyen and T. Nguyen. Optimal thresholding quantizer maximizing mutual information of discrete multiple-input continuous one-bit output quantization. In *IEEE International Symposium on Information Theory*, pages 646–651, 2021.

[10] Thinh Nguyen. Robust data-optimized stochastic analog-to-digital converters. *IEEE transactions on signal processing*, 55(6):2735–2740, 2007.

[11] Hao Chen, Lav R Varshney, and Pramod K Varshney. Noise-enhanced information systems. *Proceedings of the IEEE*, 102(10):1607–1621, 2014.

[12] Brett Mumey and Tom Gedeon. Optimal mutual information quantization is np-complete. *Neural Information Coding Workshop*, pages 1–6, 2003.

[13] Gholamreza Alirezaei and Rudolf Mathar. Optimum one-bit quantization. In *2015 IEEE Information Theory Workshop-Fall (ITW)*, pages 357–361. IEEE, 2015.

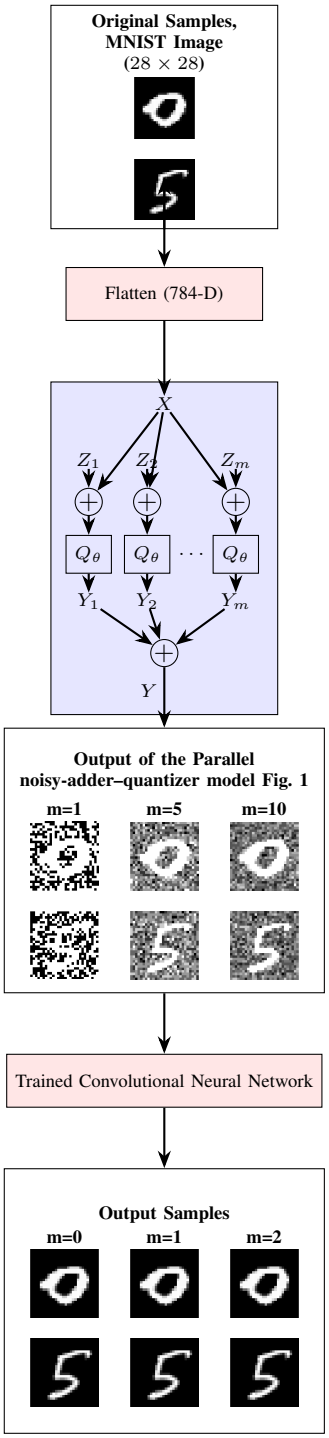

Fig. 4. System model: original MNIST samples are flattened, passed through a bank of noisy quantizers, merged, then classified by a neural network. Each branch adds Gaussian noise Z_i and applies a quantizer $Q_\theta(\cdot)$, yielding output Y_i. These outputs are concatenated and fed into a convolutional neural network for digit classification. Sample of 0 and 5 are shown

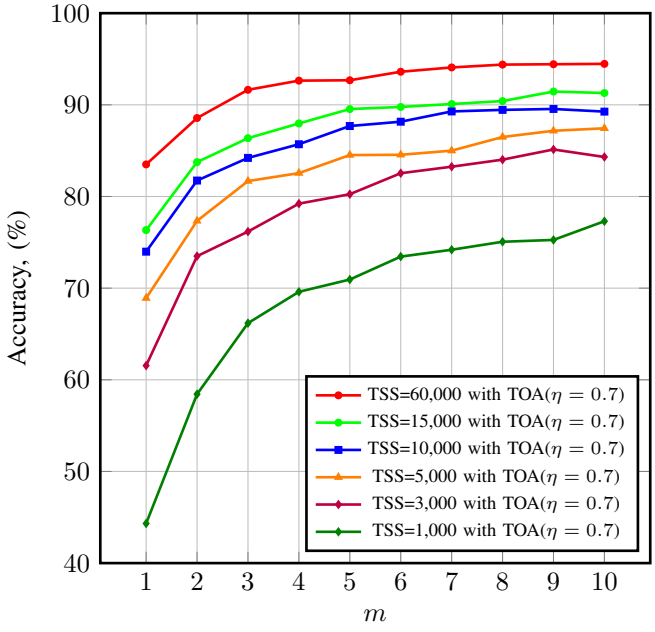

Fig. 5. Accuracy as a function of m for different Training Sample Sizes (TSS). All of Z_i Considered has Uniform distribution in [-128,128].

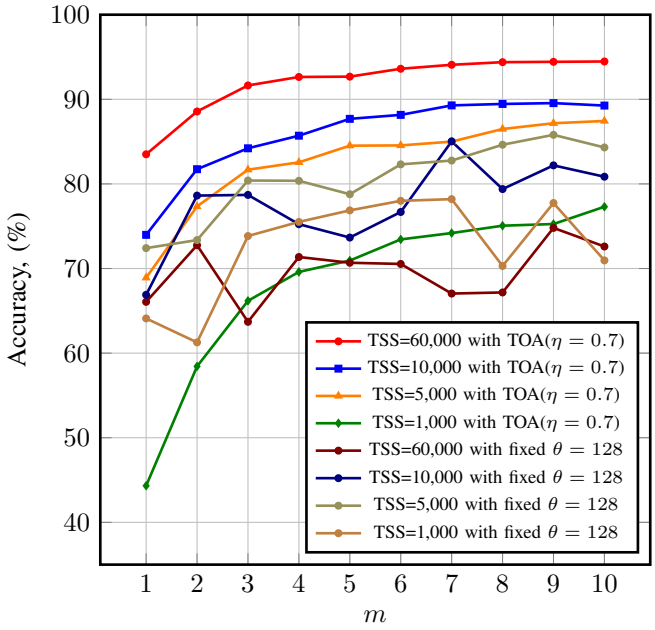

Fig. 6. Accuracy as a function of the number of parallel branches m for various training sample sizes (TSS) under two thresholding schemes: learned thresholds via TOA ($\eta = 0.7$) and fixed threshold ($\theta = 128$). All Z_i are drawn from a uniform distribution over $[-128, 128]$.

[14] Xuan He, Kui Cai, Wentu Song, and Zhen Mei. Dynamic programming for sequential deterministic quantization of discrete memoryless channels. *IEEE Transactions on Communications*, 69(6):3638–3651, 2021.

[15] Behzad Razavi. The flash adc [a circuit for all seasons]. *IEEE Solid-State Circuits Magazine*, 9(3):9–13, 2017.

[16] R.H. Walden. Analog-to-digital converter survey and analysis. *IEEE Journal on Selected Areas in Communications*, 17(4):539–550, 1999.

VII. Appendix

A. Appendix A: Proof of Proposition 1

Proof: Let us get start with Eq. 4, and write it as: $I(X;Y) = \sum_{j=1}^{n} \mathrm{E_X}(P(Y = j \mid X = x_i) \times \log P(Y = j \mid X =$

$x_i)) - \sum_{j=1}^{n} P(Y=j)\log\left(P(Y=j)\right)$. Now, consider that for random variables of $a(\theta)$ and $b(\theta)$ we have $I_k(\theta) = \mathrm{E_X}(a(\theta)\log a(\theta)) - b(\theta)\log b(\theta)$, where $\mathrm{E_X}(a(\theta)) = b(\theta)$, which gives us:

$$\frac{dI_k(\theta)}{d\theta} = \mathrm{E_X}(a'(\theta)\log a(\theta)) - b'(\theta)\log b(\theta), \tag{6}$$

$$\frac{\mathrm{d}^2 I_k(\theta)}{\mathrm{d}\theta^2} = \mathbb{E}_X\left[a''(\theta)\log\left(\frac{a(\theta)}{b(\theta)}\right)\right] - \frac{1}{\ln(2)}\,\mathbb{E}_X\left[\frac{\left(a'(\theta)\right)^2}{a(\theta)}\right] - \frac{1}{\ln(2)}\,\frac{\left[\mathbb{E}_X\big(a'(\theta)\big)\right]^2}{b(\theta)}. \tag{7}$$

So, we have:

$$\frac{d^2 I_\theta(X;Y)}{d\theta^2} = \sum_{j=0}^{m}\left\{\mathbb{E}_X\left[\frac{d^2 P_\theta(Y=j\mid X)}{d\theta^2} \times \log\left(\frac{P_\theta(Y=j\mid X)}{P_\theta(Y=j)}\right)\right] - \frac{1}{\ln 2}\,\mathbb{E}_X\left[\frac{\left(\frac{dP_\theta(Y=j|X)}{d\theta}\right)^2}{P_\theta(Y=j\mid X)}\right] - \frac{1}{\ln 2}\,\frac{\left[\mathbb{E}_X\left(\frac{dP_\theta(Y=j|X)}{d\theta}\right)\right]^2}{P_\theta(Y=j)}\right\}. \tag{8}$$

Moreover, Eq. (3) shows that $P_\theta(Y=j\mid X)$ follows a binomial distribution parameterized by q, where q is a decreasing function of θ. Consequently, the second derivative $\frac{d^2 I_\theta(X;Y)}{d\theta^2}$ can take positive or negative values depending on the value of θ. ■

B. Appendix B: Proof of Proposition 2

Proof: Because the input distribution is symmetric about zero, $P_X(x) = P_X(-x)$ for all x. Likewise, zero-mean symmetric noise implies $f_Z(z) = f_Z(-z)$, so for any threshold θ, we have: $P\big(Y=j\mid X=x,\theta\big) = P\big(Q_\theta(x+Z)=j\big) = P\big(Q_{-\theta}(-x+Z)=j\big) = P\big(Y=j\mid X=-x,-\theta\big)$. Hence the joint distribution satisfies $P_{X,Y}(x,j;\theta) = P_X(x)\,P(Y=j\mid X=x,\theta) = P_X(-x)\,P(Y=j\mid X=-x,-\theta) = P_{X,Y}(-x,j;-\theta)$,

so it is invariant under $\theta\mapsto-\theta$. Consequently, the mutual information obeys $I(\theta) = I(-\theta)$, i.e. it is an even function of θ. In particular, its derivative vanishes at $\theta = 0$. ■

C. Appendix C: Proof of Theorem 1

Proof: For each input symbol x_i, the conditional output $\big(Y=\sum_{k=1}^{m} Y_k \;\big|\; X=x_i\big) \sim \mathrm{Binomial}\big(m,\ q(x_i)\big)$, where $q(x_i) = P(Y_k=1\mid X=x_i)$. By the Central Limit Theorem (CLT),

$$\frac{Y - m\,q(x_i)}{\sqrt{m\,q(x_i)\big(1-q(x_i)\big)}} \to \mathcal{N}(0,1) \quad (m\to\infty), \tag{9}$$

and hence

$$Y \to \mathcal{N}\big(m\,q(x_i),\ m\,q(x_i)(1-q(x_i))\big) \quad (m\to\infty). \tag{10}$$

Convergence in distribution of the channel law $P_{Y|X=x_i}$ to the Gaussian density, together with the Continuous Mapping Theorem, ensures that any continuous functional of the law, such as the entropy terms in the mutual information, also converges appropriately. It is known that mutual information is a continuous functional of the conditional law $P(Y\mid X)$ under standard DMC topologies, and in particular under the one induced by convergence in distribution of each $P_{Y|X=x_i}$.

Based on Proposition 2, since the input alphabet $\mathcal{X}$ is symmetric about zero and the noise samples Z_k are i.i.d. with zero-mean and symmetric distribution, the mutual information $I(X;Y)$ is uniquely maximized when the quantization threshold is set to $\theta^* = 0$. The rest of the proof follows the approach in [10]. Without loss of generality, we assume that the value 0 is mapped to -1, while 1 remains mapped to $+1$. Under this optimal threshold, the quantization estimator is given by $\hat{X} = \frac{\alpha}{m}\sum_{k=1}^{m}\mathrm{sign}(X+Z_k)$, where 'sign' is -1 for negative values and 1 for positive values. The quantization error is: $e = X - \hat{X} = X - \frac{\alpha}{m}\sum_{k=1}^{m}\mathrm{sign}(X+Z_k)$. Let $s = \sum_{k=1}^{m}\mathrm{sign}(X+Z_k)$ be the quantized sum. Then the quantization error power is $E = \int_{-\alpha}^{\alpha}\sum_{s=-m}^{m}\left(x - \frac{\alpha s}{m}\right)^2 p(x,s)\,dx$, where $p(x,s)$ is the joint distribution of x and s. Note that s only takes even values: $-m, -m+2, \ldots, m$. Using Bayes' rule and expanding the square:

$$E = \int_{-\alpha}^{\alpha}\sum_{s=-m}^{m}\left(x^2 - \frac{2\alpha x s}{m} + \frac{\alpha^2 s^2}{m^2}\right)p(s|x)p(x)dx. \tag{11}$$

Let i be the number of $+1$ outputs; then the number of -1 outputs is $j = m-i$. Since $s = i-j = 2i-m$, we can write the conditional distribution of i as: $p(i|x) = \binom{m}{i}p(1|x)^i(1-p(1|x))^{m-i}$, where $p(1|x) = \mathbb{P}[\mathrm{sign}(x+n)=1] = \frac{x+\alpha}{2\alpha}$, due to uniform noise over $[-\alpha,\alpha]$. Substituting $s = 2i-m$ and using the above probability and for term $\sum_{i=0}^{m}p(i|x) = 1$, $\int_{-\alpha}^{\alpha}x^2p(x)\,dx = \frac{\alpha^2}{3}$. For the second term, using $\mathbb{E}[i|x] = mp(1|x)$, we obtain: $\frac{2\alpha}{m}\int_{-\alpha}^{\alpha}x(2\mathbb{E}[i|x]-m)p(x)dx = 2\alpha\int_{-\alpha}^{\alpha}x\,(2p(1|x)-1)\,p(x)dx$. Using $p(1|x) = \frac{x+\alpha}{2\alpha}$ and $p(x) = \frac{1}{2\alpha}$, we evaluate: $2\alpha\int_{-\alpha}^{\alpha}x\cdot\frac{x}{\alpha}\cdot\frac{1}{2\alpha}\,dx = \frac{2\alpha^3}{3\alpha^2} = \frac{2\alpha}{3}$. For the last term, we use:

$$\mathbb{E}[s^2\mid x] = \mathrm{Var}[s\mid x] + (\mathbb{E}[s\mid x])^2 = 4m\,p(1\mid x)\big(1-p(1\mid x)\big) + \big(2m\,p(1\mid x)-m\big)^2.$$

and integrate to obtain: $\frac{(m-1)\alpha^4}{3m\alpha^2} + \frac{\alpha^2}{m}$. Putting all together, the total quantization error power is:

$$E = \frac{\alpha^2}{3} - \frac{2\alpha^2}{3} + \frac{(m-1)\alpha^2}{3m} + \frac{\alpha^2}{m}. \tag{12}$$

Finally, observe that as $m\to\infty$, the mean-squared error tends to zero, i.e., $E\to 0$. This implies that the reconstruction $\hat{X}(Y_1,\ldots,Y_m)$ becomes asymptotically accurate, and hence the quantized outputs $(Y_1,\ldots,Y_m)$ determine X with vanishing uncertainty. Therefore, the conditional entropy converges to zero: $\lim_{m\to\infty}H(X\mid Y_1,\ldots,Y_m) = 0$, (via Fano' inequality) which completes the proof. ■