On Binary Quantizer For Maximizing Mutual Information

Thuan Nguyen and Thinh Nguyen, *Senior Member, IEEE*

***Abstract*—We consider a channel with a binary input X being corrupted by a continuous-valued noise that results in a continuous-valued output Y. An optimal binary quantizer is used to quantize the continuous-valued output Y to the final binary output Z to maximize the mutual information $I(X;Z)$. We show that when the ratio of the channel conditional density $r(y)=\frac{P(Y=y|X=0)}{P(Y=y|X=1)}$ is a strictly increasing or decreasing function of y, then a quantizer having a single threshold can maximize mutual information. Furthermore, we show that an optimal quantizer (possibly with multiple thresholds) is the one with the thresholding vector whose elements are all the solutions of $r(y)=r^*$ for some constant $r^*>0$. In addition, we also characterize necessary conditions using fixed point theorem for the optimality and uniqueness of a quantizer. Based on these conditions, we propose an efficient procedure for determining all locally optimal quantizers, and thus, a globally optimal quantizer can be found. Our results also confirm some previous results using alternative elementary proofs.**

***Index Terms*—Channel quantization, mutual information, threshold, optimization.**

I. Introduction

Quantization techniques play a vital role in signal processing, communication, and information theory. A classical quantization technique maps a given real number to an element in a given finite discrete set that minimizes/maximizes a certain objective. In compression, quantization is often used to minimize the distortion (e.g. mean square error (MSE)) between the original data and its quantized version [1], [2]. In graphics, color quantization is used to reduce the number of colors in the images for displays with various capabilities [3]. In communication, quantization is often used to minimize the decoding errors. Broadly, any conversion of a high-resolution signal to a low-resolution signal requires quantization. In this paper, we consider the quantization in the context of a communication channel where the transmitted binary signal is corrupted by a continuous noise, resulting in a continuous-valued signal at the receiver. To recover the transmitted signal, the receiver performs a quantization algorithm that maps the received continuous-valued signal to the quantized signal such that the objective function between the input and the quantized output is maximized/minimized. There is a rich literature on quantizer design that minimizes various objectives. One popular objective is to minimize the average decoding error. Another fundamental objective is to maximize the mutual information between the discrete transmitted inputs and the quantized outputs. Equivalently, this objective minimizes the information loss between the inputs and the outputs, and is related to the capacity of the channel. Specifically, for a given discrete memoryless channel (DMC) specified by a channel matrix M, its capacity is found by maximizing the mutual information between the input and the output with respect to the input distribution p [4], [5]. On the other hand, our work is focused on maximizing the mutual information with respect to the quantization parameters, i.e, it is equivalent to designing a channel matrix M for a fixed distribution p that maximizes the capacity. This situation often arises in real-world scenarios where the distribution of input is already given. In addition, many recent works have proposed to use quantization strategies that maximize the mutual information in the designs of low density parity check codes (LDPC) [6], [7] and polar codes [8].

We consider a channel with binary input X that is corrupted by a given continuous noise to produce continuous-valued output Y. An optimal binary quantizer is then used to quantize the continuous-valued output Y to the final binary output Z to maximize the mutual information $I(X;Z)$. We show that when the ratio of the channel conditional density $r(y)=\frac{P(Y=y|X=0)}{P(Y=y|X=1)}$ is a strictly increasing or decreasing function of y, then a quantizer having a single threshold can maximize mutual information. Furthermore, we show that an optimal quantizer (possibly with multiple thresholds) is the one with the thresholding vector whose elements are all the solutions of $r(y)=r^*$ for some constant $r^*>0$. In addition, we characterize necessary conditions for optimality and uniqueness of a quantizer via a fixed point theorem. Based on this, we propose an efficient algorithm that is able to determine all of the locally optimal quantizers that finally results in the globally optimal quantizer. Our results also confirm some previous results using alternative elementary proofs.

The outline of the paper is as follows. First, we discuss a few related works in Section II. In Section III, we formulate the problem of designing the optimal quantizer that maximizes the mutual information. In Section IV, we describe the structure of optimal quantizers. In Section V, we describe the sufficient conditions via the fixed point theorem for the optimality and uniqueness of a quantizer, together with an efficient procedure for finding the globally optimal quantizer.

Thuan Nguyen is with the School of Electrical Engineering and Computer Science, Oregon State University, Oregon, OR, 97331 USA, e-mail: (nguyeth9@oregonstate.edu).

Thinh Nguyen is with the School of Electrical Engineering and Computer Science, Oregon State University, Oregon, OR, 97331 USA, e-mail: (thinhq@eecs.oregonstate.edu).

II. Related Work

Research on quantization techniques has a long history, including many earliest works in 1960s [9] that aim to minimize the distortion between the original signal and the quantized signal. From a communication perspective, designing the quantizers that maximize the information capacity for Gaussian channels have also been proposed in 1970s [10]. Recently, in constructing efficient codes such as LDPC and polar codes, a number of works have made use of quantizers that maximize the mutual information [6], [7], [8]. Many advanced quantization algorithms have also been proposed to maximize the mutual information between the input and the quantized output over the past decade [11], [12], [13], [14], [15], [16]. In [11], the channel is assumed to have discrete input and discrete output, and the optimal quantizers can be found efficiently using dynamic programming that has polynomial time complexity [17]. On the other hand, we study the channels with discrete binary inputs and continuous-valued outputs which are then quantized to binary outputs. The continuous-valued output is a direct result of the conditional channel density. We note that it is possible to first discretize the continuous-valued output, then use the existing quantization algorithms for the discrete input-discrete output channels [11]. However, in many scenarios, this may result in loss of efficiencies. In particular, many analytical and computational techniques for dealing with continuous-valued functions are more efficient than their discrete counterparts.

Our work is also related to the classification problem in learning theory. Burshtein et al. gave the condition on the existence of an optimal quantizer which minimizes the impurity of partitions [18]. Because of the similarity between maximizing mutual information and minimizing conditional entropy function [11], [19], the result in [18] can be applied for finding the optimal quantizer. A similar result also can be found in [20]. In [21], Zhang et al. show that finding an optimal quantizer is equivalent to finding an optimal clustering. Therefore, a locally optimal solution can be found using k-means algorithm with the Kullback-Leibler (KL) divergence as the distance metric. Recently, there have also been many works on approximating the optimal clustering that minimize the impurity function for high dimensional data [22], [23], [24].

There are also works on finding channel capacity by maximizing the mutual information over both input probability mass function (pmf) and thresholds variables. This problem remains a hard problem [12], [25], [26], [27], [28]. Although the mutual information is a convex function in the input pmf, it is not a convex function in the quantization parameters. As such, many successful convex optimization techniques for finding the optimal solution are not applicable. In [25], a heuristic near optimal quantization algorithm is proposed. However, the algorithm only works well when the SNR ratio is high. In [12], R. Mathar et al. investigated an optimal quantization strategy for binary input-multiple output channels using two support points. These results are only applicable to approximate the optimal point between two supporting points. In [19], Kurkoski et al. constructed a sufficient condition such that a single threshold quantizer is optimal for arbitrary binary-input, continuous-output channels based on Burshtein et al.'s theorem on optimal classification [18]. On the other hand, our work describes the generalized conditions for the existence of a single threshold optimal quantizer together with a simple procedure that is able to find the globally optimal quantizer efficiently.

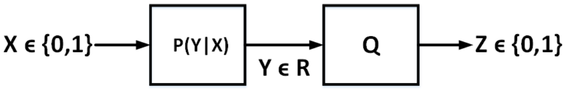

Figure 1. Channel model: binary input X is corrupted by continuous noise to result in continuous-valued Y at the receiver. The receiver attempts to recover X by quantizing Y into binary signal Z.

III. Problem description

We consider the channel shown in Fig. 1 where the binary signals $x \in X = \{0,1\}$ are transmitted and corrupted by a continuous noise source to produce a continuous-valued output $y \in \mathbb{R}$ at the receiver. Specifically, y is specified by the a channel conditional density $p(y|x)$. $p(y|x)$ models the distortion caused by noise. The receiver recovers the original binary signal x by decoding the received continuous-valued signal y to $z \in Z = \{0,1\}$ using a quantizer Q. Since $y \in \mathbb{R}$, the quantization parameters can be specified by a thresholding vector

$$\mathbf{h} = (h_1, h_2, \dots, h_n) \in \mathbb{R}^n,$$

with $h_1 < h_2 < \cdots < h_{n-1} < h_n$, where n is assumed a finite number. Theoretically, it might be perceivably possible to construct the conditional densities $p(y|x_0)$ and $p(y|x_1)$ such that the optimal quantizer might consist an infinite number of thresholds. On the other hand, for a practical implementation, especially when the quantizer is implemented using a lookup table, then a finite number of thresholds must be used. To that end, the optimal quantizer in this paper refers to the best quantizer in the class of all quantizers with a finite number of thresholds.

In particular, $\mathbf{h}$ induces $n+1$ disjoint partitions:

$$H_1 = (-\infty, h_1), H_2 = [h_1, h_2), \dots, H_n = [h_{n-1}, h_n), H_{n+1} = [h_n, \infty).$$

Let $\mathbb{H} = \bigcup_{i \in odd} H_i$ and $\bar{\mathbb{H}} = \bigcup_{i \in even} H_i$, then $\mathbb{H} \cap \bar{\mathbb{H}} = \emptyset$ and $\mathbb{H} \cup \bar{\mathbb{H}} = \mathbb{R}$.

The receiver uses a quantizer $Q : Y \rightarrow Z$ to quantize Y to Z as:

$$Z = \begin{cases} 0 & \text{if } Y \in \mathbb{H}, \\ 1 & \text{if } Y \in \bar{\mathbb{H}}. \end{cases} \tag{1}$$

Note that we can also switch the rule such that Q quantizes Y to $Z = 1$ if $y \in \mathbb{H}$ and quantizes Y to $Z = 0$ if $y \in \bar{\mathbb{H}}$. The main point is that $\mathbf{h}$ divides $\mathbb{R}$ into $n+1$ contiguous disjoint segments, each maps to either 0 or 1 alternatively. Our goal is to design an optimal quantizer Q^*, specifically $\mathbf{h}^*$ that maximizes the mutual information $I(X;Z)$ between the input X and the quantized output Z:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} I(X;Z). \tag{2}$$

We note that both the values of thresholds h_i's and the number of thresholds n are the optimization variables. The maximization in (2) assumes that the input probability mass function $p(x)$ and the channel conditional density $p(y|x)$ are given.

IV. Optimal Quantizer Structure

For convenience, we use the following notations:

1) $\mathbf{p} = (p_0, p_1)$ denotes the probability mass function for the input X, with $p_0 = P(X = 0)$ and $p_1 = P(X = 1)$.
2) $\mathbf{q} = (q_0, q_1)$ denotes probability mass function for the output Z, with $q_0 = P(Z = 0)$ and $q_1 = P(Z = 1)$.
3) $\phi_0(y) = p(y|x = 0)$ and $\phi_1(y) = p(y|x = 1)$ denote conditional density functions of the received signal Y given the input signal $X = 0$ and $X = 1$, respectively.

Furthermore, we make two following assumptions:

Assumptions:

1) $r(y) = \dfrac{\phi_0(y)}{\phi_1(y)}$ will play a central role this paper. All the results in this paper assume that $r(y)$ is a continuous function, and has a finite number of stationary points. Equivalently, $r(y) = r'$ has a finite number of solutions for any constant $r' > 0$. Note that this assumption will hold for most $\phi_0(y)$ and $\phi_1(y)$.
2) Both $\phi_0(y)$ and $\phi_1(y)$ are differentiable everywhere.

Using the notations and the assumptions above, a 2×2 channel matrix A associated with a discrete memoryless channel (DMC) with input X and output Z is:

$$A = \begin{bmatrix} A_{11} & 1 - A_{11} \\ 1 - A_{22} & A_{22} \end{bmatrix},$$

where

$$A_{11} = \int_{y \in \mathbb{H}} \phi_0(y) dy, \tag{3}$$

$$A_{22} = \int_{y \in \bar{\mathbb{H}}} \phi_1(y) dy. \tag{4}$$

The simplest quantizer (decoding scheme) uses only a single threshold to quantize a continuous received signal into binary outputs. Specifically,

$$Z = \begin{cases} 0 & \text{if } Y < h_1, \\ 1 & \text{otherwise.} \end{cases}$$

In general, this quantizer is not optimal, i.e., does not maximize the mutual information $I(X; Z)$. Using the results of Burshtein et al. [18], Kurkoski et al. [19] showed a sufficient condition on $p(y|x)$ for which the single threshold quantizer is indeed an optimal quantizer. Our first contribution is to show that the optimal binary quantizer with multiple thresholds, specified by a thresholding vector $\mathbf{h}^* = (h_1^*, h_2^*, \dots, h_n^*)$ with $h_i^* < h_{i+1}^*$, must satisfy the conditions stated in the Theorem 1.

Theorem 1. Let $\mathbf{h}^* = (h_1^*, \dots, h_n^*)$ be a thresholding vector of an optimal quantizer Q^*, then:

$$\frac{\phi_0(h_i^*)}{\phi_1(h_i^*)} = \frac{\phi_0(h_j^*)}{\phi_1(h_j^*)} = r^*, \tag{5}$$

for $\forall\ i, j \in \{1, 2, \dots, n\}$ and some optimal constant $r^* > 0$.

Proof. We note that using the optimal thresholding vector $\mathbf{h}^*$, the quantization mapping follows (1). $\mathbf{h}^*$ divides $\mathbb{R}$ into $n + 1$ contiguous disjoint segments, each maps to either 0 or 1 alternatively. The overall DMC in Fig. 1 has the channel matrix

$$A^* = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

and the mutual information can be written as a function of $\mathbf{h}$ as:

$$I(\mathbf{h}) = H(Z) - H(Z|X) = H(q_0) - [p_0 H(A_{11}) + p_1 H(A_{22})], \tag{6}$$

where for any $w \in [0, 1]$, $H(w) = -[w \log(w) + (1 - w) \log(1 - w)]$ and $q_0 = P(Z = 0) = p_0 A_{11} + p_1 A_{21}$.

This is an optimization problem that maximizes $I(\mathbf{h})$. The theory of optimization requires that an optimal point must satisfy the KKT conditions [29]. In particular, define the Lagrangian function as:

$$L(\mathbf{h}, \lambda) = I(\mathbf{h}) + \sum_{i=1}^{n-1} \lambda_i (h_i - h_{i+1}), \tag{7}$$

then the KKT conditions [29] states that, an optimal point $\mathbf{h}^*$ and $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_{n-1}^*)$ must satisfy:

$$\begin{cases} \frac{\partial L(\mathbf{h}, \lambda)}{\partial h_i} |_{\mathbf{h} = \mathbf{h}^*, \lambda = \lambda^*} = 0, i = 1, 2, \dots, n - 1, \\ \lambda_i^* (h_i - h_{i+1}) = 0, i = 1, 2, \dots, n - 1, \\ \lambda_i^* \geq 0, i = 1, 2, \dots, n - 1. \end{cases} \tag{8}$$

Since the structure of the quantizer requires that $h_i < h_{i+1}$, the second and the third conditions in (8) together imply that $\lambda_i^* = 0, i = 1, 2, \dots, n - 1$. Consequently, from (7) and the first condition in (8), we have:

$$\frac{\partial L(\mathbf{h}, \lambda)}{\partial h_i} |_{\mathbf{h} = \mathbf{h}^*, \lambda = \lambda^*} = \frac{\partial I(\mathbf{h})}{\partial h_i} |_{\mathbf{h} = \mathbf{h}^*} = 0.$$

The stationary points can be found by setting the partial derivatives with respect to each h_i to zero:

$$\begin{aligned} \frac{\partial I(\mathbf{h})}{\partial h_i} &= (\log \frac{1 - q_0}{q_0}) \frac{\partial q_0}{\partial h_i} - p_0 (\log \frac{1 - A_{11}}{A_{11}}) \frac{\partial A_{11}}{\partial h_i} \\ &- p_1 (\log \frac{1 - A_{22}}{A_{22}}) \frac{\partial A_{22}}{\partial h_i} \\ &= (\log \frac{1 - q_0}{q_0}) (p_0 \frac{\partial A_{11}}{\partial h_i} - p_1 \frac{\partial A_{22}}{\partial h_i}) \\ &- p_0 (\log \frac{1 - A_{11}}{A_{11}}) \frac{\partial A_{11}}{\partial h_i} - p_1 (\log \frac{1 - A_{22}}{A_{22}}) \frac{\partial A_{22}}{\partial h_i} \quad (9) \\ &= p_0 \frac{\partial A_{11}}{\partial h_i} (\log \frac{1 - q_0}{q_0} - \log \frac{1 - A_{11}}{A_{11}}) \\ &- p_1 \frac{\partial A_{22}}{\partial h_i} (\log \frac{1 - q_0}{q_0} + \log \frac{1 - A_{22}}{A_{22}}) = 0, \quad (10) \end{aligned}$$

with (9) due to $q_0 = p_0 A_{11} + p_1 A_{21} = p_0 A_{11} + p_1 (1 - A_{22})$.

Since $\frac{\partial A_{11}}{\partial h_i} = \phi_0(h_i)$ and $\frac{\partial A_{22}}{\partial h_i} = -\phi_1(h_i)$, from (10), we have:

$$\frac{\phi_0(h_i^*)}{\phi_1(h_i^*)} = -\frac{p_1}{p_0} \frac{\log \frac{1 - q_0}{q_0} + \log \frac{1 - A_{22}}{A_{22}}}{\log \frac{1 - q_0}{q_0} - \log \frac{1 - A_{11}}{A_{11}}} = r^*. \tag{11}$$

Since $r^* > 0$ (please see Appendix E) and (11) holds for $\forall\ i$, the RHS of (11) equals to some constant $r^* > 0$ for a quantizer Q^*. Theorem 1 follows. □

Remark: The importance of Theorem 1 is as follows. Suppose the optimal value r^* is given and the equation $r(y) = r^*$ has m solutions: $y_1 < y_2 < \cdots < y_m$. Then, Theorem 1 says that the optimal quantizer must either have its thresholding vector be $(y_1, y_2, \ldots, y_m)$ or one of its ordered subsets, e.g., $(h_1^*, h_2^*) = (y_1, y_3)$, or both. In Theorem 2 below, we will show that the quantizer whose thresholding vector is all the solutions of $r(y) = r^*$, will be at least as good as any quantizer whose thresholding vector is an ordered subset of the set of all solutions. Moreover, we will show that under some sufficient conditions via Banach's fixed point theorem, r^* is unique, and describe an efficient procedure for finding r^* in Section V.

Theorem 2. Let $y_1^* < y_2^* < \cdots < y_n^*$ be the solutions of $r(y) = r^*$ for the optimal constant $r^* > 0$. Let $Q_{r^*}^n$ be the quantizer whose thresholding vector is all the solutions, i.e., $h_i^* = y_i^*, i = 1, 2, \ldots, n$, then for $k < n$, $Q_{r^*}^n$ is at least as good as any quantizer $Q_{r^*}^k$ whose thresholding vector is an ordered subset of k elements of the set of $(h_1^*, h_2^*, \ldots, h_n^*)$.

Proof. Let $(h_1^*, h_2^*, \ldots, h_m^*)$ be an optimal thresholding vector for all the quantizers having m thresholds ($m \leq n$). Let $(z_1^*, z_2^*, \ldots, z_{m-1}^*)$ be an optimal thresholding vector for all quantizers having $m-1$ thresholds. The mutual information can be written as a function of these quantizers as: $I(h_1^*, h_2^*, \ldots, h_m^*)$ and $I(z_1^*, z_2^*, \ldots, z_{m-1}^*)$. We will first show that $I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(z_1^*, z_2^*, \ldots, z_{m-1}^*)$, for any $m > 0$. This will be proved using contradiction.

Assume that $I(h_1^*, h_2^*, \ldots, h_m^*) < I(z_1^*, z_2^*, \ldots, z_{m-1}^*)$, then

$$I(z_1^*, z_2^*, \ldots, z_{m-1}^*) = I(h_1^*, h_2^*, \ldots, h_m^*) + \delta, \tag{12}$$

where δ is a positive constant.

Since $(h_1^*, h_2^*, \ldots, h_m^*)$ is optimal,

$$I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(h_1, h_2, \ldots, h_{m-1}, h_m), \tag{13}$$

for any $h_i < h_{i+1}$, $i = 1, 2, \ldots, m-1$.

Now replacing $h_i = z_i^*$, for $i = 1, 2, \ldots, m-1$ into (13), we have:

$$I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(z_1^*, z_2^*, \ldots, z_{m-1}^*, h_m). \tag{14}$$

Since $\int_{-\infty}^{\infty} \phi_i(y)dy = 1$, $\forall\ i = 1, 2$,

$$\lim_{y \to \infty} \phi_i(y) = 0, i = 1, 2.$$

Consequently,

$$\lim_{h_m \to \infty} I(z_1^*, z_2^*, \ldots, z_{m-1}^*, h_m) = I(z_1^*, z_2^*, \ldots, z_{m-1}^*).$$

Equivalently, there exists an $h_m > N_\epsilon$ such that

$$|I(z_1^*, z_2^*, \ldots, z_{m-1}^*, h_m) - I(z_1^*, z_2^*, \ldots, z_{m-1}^*)| \leq \epsilon, \tag{15}$$

for any $\epsilon > 0$. Next, we pick a N_ϵ such that $\epsilon < \delta$. Then,

$$\begin{aligned} & I(h_1^*, h_2^*, \ldots, h_m^*) & (16) \\ = \ & I(z_1^*, \ldots, z_{m-1}^*) + I(h_1^*, h_2^*, \ldots, h_m^*) - I(z_1^*, z_2^*, \ldots, z_{m-1}^*) \\ \geq \ & I(z_1^*, \ldots, z_{m-1}^*) - |I(h_1^*, h_2^*, \ldots, h_m^*) - I(z_1^*, z_2^*, \ldots, z_{m-1}^*)| \\ \geq \ & I(h_1^*, h_2^*, \ldots, h_m^*) + \delta - \epsilon, & (17) \end{aligned}$$

where (17) is due to (12) and (15). Since $\delta - \epsilon > 0$ by assumption, (17) indicates that $I(h_1^*, h_2^*, \ldots, h_m^*)$ is strictly greater than itself which is a contradiction. Thus, $I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(z_1^*, z_2^*, \ldots, z_{m-1}^*)$.

Next, since $(z_1^*, z_2^*, \ldots, z_{m-1}^*)$ is an optimal thresholding vector for all quantizers having $m-1$ thresholds, $I(z_1^*, z_2^*, \ldots, z_{m-1}^*) \geq I(\bar{h}_1^*, \bar{h}_2^*, \ldots, \bar{h}_{m-1}^*)$ where $(\bar{h}_1^*, \bar{h}_2^*, \ldots, \bar{h}_{m-1}^*)$ is an arbitrary subset of $(h_1^*, h_2^*, \ldots, h_m^*)$. Thus, $I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(z_1^*, z_2^*, \ldots, z_{m-1}^*) \geq I(\bar{h}_1^*, \bar{h}_2^*, \ldots, \bar{h}_{m-1}^*)$. Consequently, the optimal quantizer having n thresholds is at least as good as the optimal quantizer having $n-1$ thresholds. Similarly, the optimal quantizer having $n-1$ thresholds is at least as good as the optimal quantizer having $n-2$ thresholds and so on. Thus, by induction, $Q_{r^*}^n$ is at least as good as any quantizer $Q_{r^*}^k$, $\forall$ $k < n$.

□

Corollary 1. If

$$r(y) = \frac{\phi_0(y)}{\phi_1(y)} \tag{18}$$

is a strictly increasing or decreasing function, then the optimal quantizer consists of only a single threshold h_1^*.

Proof. Noting that since $r(y)$ is a strictly increasing or decreasing function. Therefore, $r(y_1) \neq r(y_2)$ for $y_1 \neq y_2$. Thus, (5) will not hold for $h_1^* \neq h_2^*$. Consequently, the optimal quantizer has only a single threshold. □

We note that in a previous result [19], an optimality condition for a single threshold quantizer is that:

$$s(y) = \log \frac{\phi_0(y)}{\phi_1(y)} \tag{19}$$

is a monotonic function. If $\dfrac{\phi_0(y)}{\phi_1(y)}$ is a strictly monotonic function, then previous result is a consequence of Corollary 1 since $\log(.)$ is a strictly monotonic function, any strictly monotonic function $\dfrac{\phi_0(y)}{\phi_1(y)}$ results in a strictly monotonic function $s(y)$.

Corollary 2. If

$$\phi_0(y - \mu) = \phi_1(y) \text{ for some constant } \mu, \tag{20}$$

and $\phi_0(y)$ is a strictly log-concave or log-convex function, then using a single threshold quantizer is optimal.

Proof. Taking derivative of $r(y)$, we have:

$$\frac{dr(y)}{dy} = \frac{\phi_0'(y)\phi_1(y) - \phi_0(y)\phi_1'(y)}{\phi_1(y)^2} > 0, \tag{21}$$

which is equivalent with:

$$\frac{\phi_0'(y)}{\phi_0(y)} > \frac{\phi_1'(y)}{\phi_1(y)}. \tag{22}$$

Using (20), we have:

$$\frac{\phi_0'(y)}{\phi_0(y)} > \frac{\phi_0'(y-\mu)}{\phi_0(y-\mu)}. \tag{23}$$

Now, a function $f(x)$ is strictly log-convex if and only if $\frac{f'(x)}{f(x)}$ is a strictly increasing function [29]. Thus, if $\phi_0(y)$ is strictly log-convex, then

$$\frac{\phi_0'(y)}{\phi_0(y)} > \frac{\phi_0'(y-\mu)}{\phi_0(y-\mu)}. \tag{24}$$

Thus, $r'(y) > 0$ or $r(y)$ is a strictly increasing function which satisfies the condition for having an optimal single threshold quantizer in Corollary 1. A similar proof can be established for log-concave functions. □

V. Necessary Conditions For Optimality and Uniqueness of a Quantizer Via Fixed Point Theorem and Fixed Point Algorithm

In this section, we characterize necessary conditions for optimality and uniqueness of a quantizer via a fixed point theorem. Using this new conditions, we describe an efficient procedure based on fixed point algorithm for finding all the possible r^* that results in a globally optimal quantizer Q^*.

A. Necessary Conditions for Optimality via Fixed Point Theorem

For ease of analysis, we define a new variable a as:

$$a = \frac{p_1\phi_1(y)}{p_0\phi_0(y) + p_1\phi_1(y)} = \frac{1}{1 + \frac{p_0\phi_0(y)}{p_1\phi_1(y)}} = \frac{1}{1 + (\frac{p_0}{p_1})r}, \tag{25}$$

where

$$r = \frac{\phi_0(y)}{\phi_1(y)}.$$

We note that $a \in (0,1)$. In addition, the mapping from r to a is a one-to-one mapping. Furthermore, each value of a corresponds to a different value of r which in turn, corresponds to a quantizer in a set of possible quantizers that contains an optimal quantizer. As an example, Fig. 2 shows two conditional densities $\phi_0(y)$ and $\phi_1(y)$, and the corresponding $r(y)$ and $u(y)$ are shown in Fig. 3 and Fig. 4, respectively. Now, the mutual information $I(X;Z)$ can be re-written as a function of a, and is denoted as $I(X;Z)_a$. Thus, finding the optimal r^* is equivalent to finding the optimal a^* that maximizes $I(X;Z)_a$. Furthermore, the optimal thresholds $\mathbf{h}^* = (h_1^*, \dots, h_n^*)$ can be directly determined as the solutions of

$$\frac{p_1\phi_1(h)}{p_0\phi_0(h) + p_1\phi_1(h)} = a^*. \tag{26}$$

First, let

$$u(y) = \frac{p_1\phi_1(y)}{p_0\phi_0(y) + p_1\phi_1(y)}. \tag{27}$$

For given a, define $\mathbb{H}_a = \{y : u(y) < a\}$ and $\bar{\mathbb{H}}_a = \{y : u(y) \geq a\}$. The sets $\mathbb{H}_a$ and $\bar{\mathbb{H}}_a$ together specify a binary quantizer that maps y to $z \in \{0,1\}$, depending on whether y belongs to $\mathbb{H}_a$ or $\bar{\mathbb{H}}_a$ as shown in Fig. 4.

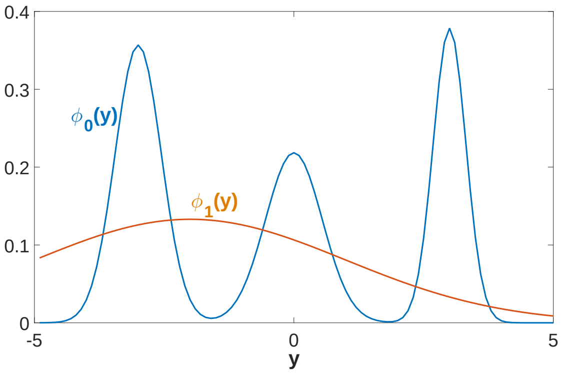

Figure 2. Conditional densities $\phi_0(y) = 0.3N(0,\sqrt{0.3}) + 0.4N(-3,\sqrt{0.2}) + 0.3N(3,\sqrt{0.1})$ and $\phi_1(y) = N(-2,3)$. They are used in Fig. 3 and Fig. 4.

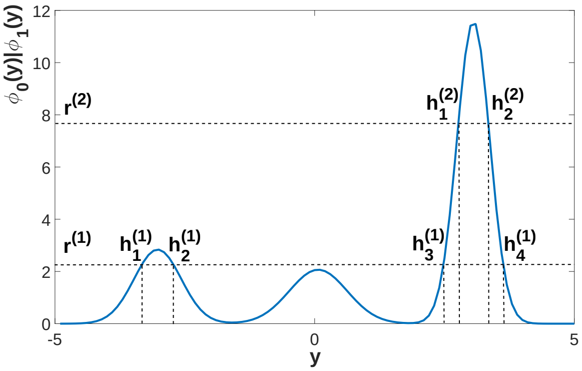

Figure 3. Two thresholding vectors: $\mathbf{h}^{(1)} = (h_1^{(1)}, h_2^{(1)}, h_3^{(1)}, h_4^{(1)})$ and $\mathbf{h}^{(2)} = (h_1^{(2)}, h_2^{(2)})$ correspond to two different values of r are shown. $\phi_0(y) = 0.3N(0,\sqrt{0.3}) + 0.4N(-3,\sqrt{0.2}) + 0.3N(3,\sqrt{0.1})$, $\phi_1(y) = N(-2,3)$.

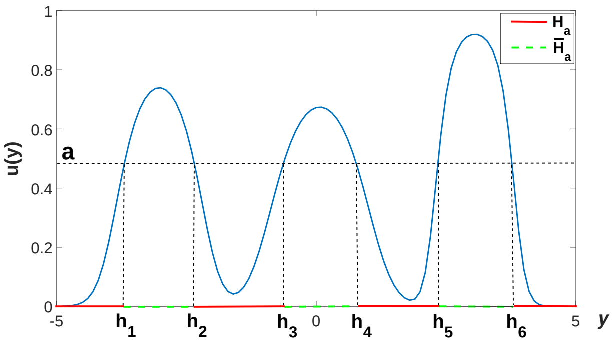

Figure 4. Illustration of the sets $\mathbb{H}_a$ and $\bar{\mathbb{H}}_a$. $\mathbb{H}_a$ consists of solid red segments while $\bar{\mathbb{H}}_a$ consists of green dotted segments. In this example, there exists a quantizer with 6 thresholds $h_1, h_2, \dots, h_6$ that correspond to a specific value of $a = 0.5$. $p_0 = p_1 = 0.5$, $\phi_0(y) = 0.3N(0,\sqrt{0.3}) + 0.4N(-3,\sqrt{0.2}) + 0.3N(3,\sqrt{0.1})$, $\phi_1(y) = N(-2,3)$.

Without the loss of generality, suppose we use the following quantizer:

$$z = \begin{cases} 0 & y \in \mathbb{H}_a, \\ 1 & y \in \bar{\mathbb{H}}_a, \end{cases} \tag{28}$$

then the channel matrix of the overall DMC is:

$$A = \begin{bmatrix} f(a) & 1-f(a) \\ 1-g(a) & g(a) \end{bmatrix},$$

where $f(a) \triangleq p(z=0|x=0)$ and $g(a) \triangleq p(z=1|x=1)$. $f(a)$ and $g(a)$ can be written in terms of $\phi_0(y)$ and $\phi_1(y)$ as:

$$f(a) = \int_{y \in \mathbb{H}_a} \phi_0(y) dy, \tag{29}$$

$$g(a) = \int_{y \in \bar{\mathbb{H}}_a} \phi_1(y) dy. \tag{30}$$

Now, let us consider the special cases where $a=1$ or $a=0$. In these cases, $I(X;Z)=0$ due to $f(a)=1$ and $g(a)=0$ or vice-versa. Therefore, $a=1$ and $a=0$ cannot be the optimal points. Thus, we can assume that $a \in (0,1)$ and $0 < f(a), g(a) < 1$. Lemmas 1 and 2 below provide the properties of $f(a)$ and $g(a)$ and the relationship with each other.

Lemma 1. Derivatives of $f(a)$ and $g(a)$ are related through the following equation:

$$\frac{dg(a)}{da} = -\frac{ap_0}{(1-a)p_1}\frac{df(a)}{da}. \tag{31}$$

Proof. Please see the proof in Appendix A. □

Lemma 2. For $\forall\ a \in (0,1)$,
(1) $g'(a) < 0$ and $f'(a) > 0$.
(2) $f(a) + g(a) > 1$.

Proof. Please see the proof in Appendix B. □

Define

$$\begin{aligned} \mathbf{l}_a &= [\frac{p_0 f(a)}{p_0 f(a) + p_1(1-g(a))}, \frac{p_1(1-g(a))}{p_0 f(a) + p_1(1-g(a))}], \\ \mathbf{r}_a &= [\frac{p_0(1-f(a))}{p_0(1-f(a)) + p_1 g(a)}, \frac{p_1 g(a)}{p_0(1-f(a)) + p_1 g(a)}], \\ \mathbf{a} &= [1-a, a]. \end{aligned}$$

Let $D_{KL}(\mathbf{x}, \mathbf{y})$ denote the Kullback-Leibler (KL) divergence between two vectors $\mathbf{x} = [1-x, x]$ and $\mathbf{y} = [1-y, y]$ for $x, y \in (0,1)$,

$$D_{KL}(\mathbf{x}||\mathbf{y}) = x\log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y}). \tag{32}$$

Lemma 3. Each optimal quantizer Q^* (local or global) corresponds to an optimal a^* such that

$$D_{KL}(\mathbf{a}^*||\mathbf{l}_{a^*}) = D_{KL}(\mathbf{a}^*||\mathbf{r}_{a^*}).$$

Proof. Using Lemma 1, setting derivative of $I(X;Z)_a$ to zero, we have:

$$\begin{aligned} \frac{dI(X;Z)_a}{da} &= p_1 g'(a)[\frac{a-1}{a}\big(\log(\frac{f(a)}{1-f(a)}) && (33) \\ &- \log(\frac{p_0 f(a) + p_1(1-g(a))}{p_0(1-f(a)) + p_1 g(a)})\big) \\ &+ \log(\frac{g(a)}{1-g(a)}) + \log(\frac{p_0 f(a) + p_1(1-g(a))}{p_0(1-f(a)) + p_1 g(a)})] \\ &= p_1 g'(a) F(a) = 0, && (34) \end{aligned}$$

where

$$\begin{aligned} F(a) &= \frac{a-1}{a}\log(\frac{f(a)}{1-f(a)}) + \log(\frac{g(a)}{1-g(a)}) \\ &+ \frac{1}{a}\log(\frac{p_0 f(a) + p_1(1-g(a))}{p_0(1-f(a)) + p_1 g(a)}). \end{aligned} \tag{35}$$

From Lemma 2 $g'(a) < 0$ and $p_1 > 0$, thus, the stationary points of $I(X;Z)_a$ must occur at $F(a) = 0$. Applying definitions of $\mathbf{a}, \mathbf{l}_a, \mathbf{r}_a$, and KL divergence, it can be shown that

$$F(a) = \frac{1}{a}\big[D_{KL}(\mathbf{a}||\mathbf{l}_a) - D_{KL}(\mathbf{a}||\mathbf{r}_a)\big].$$

Please see the proof in Appendix C. Thus,

$$F(a) = 0 \leftrightarrow \big[D_{KL}(\mathbf{a}||\mathbf{l}_a) - D_{KL}(\mathbf{a}||\mathbf{r}_a)\big] = 0.$$

In other words, each optimal quantizer Q^* (local or global) corresponds to an optimal a^* such that

$$D_{KL}(\mathbf{a}^*||\mathbf{l}_{a^*}) = D_{KL}(\mathbf{a}^*||\mathbf{r}_{a^*}).$$

□

Lemma 4. Let $\mathbf{c}_a = [1-c(a), c(a)]$ then

$$c(a) = \frac{\log(\dfrac{1-f(a)}{f(a)}\dfrac{p_0 f(a) + p_1(1-g(a))}{p_0(1-f(a)) + p_1 g(a)})}{\log(\dfrac{1-f(a)}{f(a)}\dfrac{1-g(a)}{g(a)})} \tag{36}$$

if and only if

$$D_{KL}(\mathbf{c}_a||\mathbf{l}_a) = D_{KL}(\mathbf{c}_a||\mathbf{r}_a).$$

Proof. By using the definitions of $\mathbf{c}_a, \mathbf{l}_a, \mathbf{r}_a$, and KL divergence, (37) follows. Now, $(1-f(a))(1-g(a)) = 1 - f(a) - g(a) + f(a)g(a) < f(a)g(a)$ due to $f(a) + g(a) > 1$. Thus, $\log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)})) \neq 0$. Therefore, $D_{KL}(\mathbf{c}_a||\mathbf{l}_a) - D_{KL}(\mathbf{c}_a||\mathbf{r}_a) = 0$ if and only if $c(a)$ satisfies (36). □

We now characterize the optimality condition for a quantizer via the fixed point theorem.

Theorem 3. Let a quantizer Q^* be an optimal quantizer with an optimal a^*, then $c(a^*) = a^*$ where $c(a)$ is defined in (36).

Proof. From Lemma 3, the optimal quantizer Q^* corresponds to an optimal vector $\mathbf{a}^* = [1-a^*, a^*]$ must have $D_{KL}(\mathbf{a}^*||\mathbf{l}_{a^*}) = D_{KL}(\mathbf{a}^*||\mathbf{r}_{a^*})$. Now, from Lemma 4 for given $\mathbf{l}_{a^*}$ and $\mathbf{r}_{a^*}$, there exists a unique vector $\mathbf{c}_{a^*} = [1-c(a^*), c(a^*)]$ such that $D_{KL}(\mathbf{c}_{a^*}||\mathbf{l}_{a^*}) = D_{KL}(\mathbf{c}_{a^*}||\mathbf{r}_{a^*})$ where $c(a)$ is defined in (36). Combining Lemma 3 and 4, we have $c(a^*) = a^*$. □

We will use Theorem 3 in our algorithm for finding optimal quantizers. To do that, we will show some interesting properties of $c(a)$ in Theorem 4 and Theorem 5 below.

Theorem 4. $c(a) \in (0,1)$ and is a smooth (derivative exists), non-decreasing function of a.

Proof. Please see Appendix D for the proof. □

Lemma 5. The sequence $a^{i+1} = c(a^i)$ must converge to a fixed point a^* for any initial point $a^0 \in (0,1)$.

$$\begin{aligned}
D_{KL}(\mathbf{c}_a||\mathbf{l}_a) - D_{KL}(\mathbf{c}_a||\mathbf{r}_a) &= \big(c(a)\log(\frac{c(a)}{\dfrac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}}) + (1-c(a))\log(\frac{1-c(a)}{\dfrac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}})\big) \\
&- \big(c(a)\log(\frac{c(a)}{\dfrac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}) + (1-c(a))\log(\frac{1-c(a)}{\dfrac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}})\big) \\
&= c(a)\log(\frac{\dfrac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}{\dfrac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}}) + (1-c(a))\log(\frac{\dfrac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}}{\dfrac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}}) \\
&= \log(\frac{\dfrac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}}{\dfrac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}}) - c(a)\big(\log(\frac{\dfrac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}}{\dfrac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}}) - \log(\frac{\dfrac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}{\dfrac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}})\big) \\
&= \log((\frac{1-f(a)}{f(a)})(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)})) - c(a)\log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)})) \\
&= \log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)}))\big(\frac{\log((\dfrac{1-f(a)}{f(a)})(\dfrac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)}))}{\log((\dfrac{1-f(a)}{f(a)})(\dfrac{1-g(a)}{g(a)}))} - c(a)\big). \qquad (37)
\end{aligned}$$

Proof. From Theorem 4, $c(a)$ is a non-decreasing function and $c(a) \in (0,1)$. Thus, the sequence generated by $a^{i+1} = c(a^i)$, starting from any a^0 is monotone, i.e., $a^{i+1} \geq a^i \; \forall \; i$ or $a^{i+1} \leq a^i \; \forall \; i$. Specifically, if $a^1 \leq a^0$, then $a^2 = c(a^1) \leq c(a^0) = a^1$, therefore, $a^2 \leq a^1$. By induction method, if $a^1 \leq a^0$ then $a^{i+1} \leq a^i \; \forall \; i$. Similarly, if $a^1 \geq a^0$ then $a^{i+1} \geq a^i \; \forall \; i$. Thus, the sequence a^i is monotone. From Theorem 4, $c(a^i) \in (0,1)$ or the sequence a^i is bounded in $(0,1)$. Thus, sequence a^i has a limit a^* such that $a^* = c(a^*)$. □

Theorem 5. For any initial point $a^0 \in (0,1)$, if $\lim_{i \to +\infty} a^i = a^*$ where $a^{i+1} = c(a^i)$, then there is no other solution a' such that $a' = c(a')$ between a_0 and a^*.

Proof. We will prove by contradiction. For the case where $a^0 \leq a^*$, assume that there is a a' such that $a' = c(a')$ and $a^0 < a' < a^*$. Since the sequence a^i is monotone, there exists an i such that $a^i < a' < a^{i+1}$. Since $c(a)$ is non-decreasing, we have $a^{i+1} = c(a^i) \leq c(a') = a'$ which contradicts the assumption that $a' < a^{i+1}$. Similarly, we can show that there is no other solution a' in the interval (a^*, a^0) for the case $a^0 > a^*$.

Fig. 5 illustrates the convergence of sequence a^i to a^* from the initial point a^0. □

B. Outline of Algorithm for Finding All Solutions to $a^* = c(a^*)$

A straightforward way of computing the optimal a^* is the iteration method by starting with a^0. However, depending on the starting point a^0, the iterations may lead to a local optimal solution. In other words, when the equation $a = c(a)$ has more than one solution, we need a procedure capable of finding all the solutions of $a = c(a)$. Using Theorem 5, we outline an efficient procedure that can find all the solutions to $a = c(a)$. A global solution then can be chosen among these solutions that maximize the mutual information.

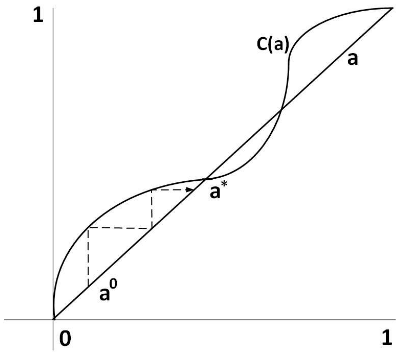

Figure 5. Illustration of the convergence of sequence a^i to a^* from the initial point a^0.

Our procedure initiates two iteration loops using two starting points $a_l^0 = \epsilon$ and $a_r^0 = 1 - \epsilon$ where ϵ is a small number. Suppose that the first iteration loop converges to a_l^*, and the second iteration loop converges to a_r^*. If $a_l^* = a_r^*$, then the procedure terminates with $a^* = a_r^*$ being the optimal point. This is due to Theorem 5 which states that there is no solution of $a = c(a)$ in either (ϵ, a^*) or $(a^*, 1-\epsilon)$. We assume that the optimal solution is not in $(0, \epsilon)$ or $(1-\epsilon, 1)$ since we can make ϵ arbitrarily small. Otherwise, if $a_l^* < a_r^*$, we need to check whether or not there exists some other solutions in the interval (a_l^*, a_r^*). In order to find them, the procedure initiates another iteration loop using a starting point $a^0 = (a_l^* + a_r^*)/2$. After this iteration loop converges to a_c^*, one needs to run the iterations over two intervals $(a_l^*, \min(a^0, a_c^*))$ and $(\max(a^0, a_c^*), a_r^*)$. If any of these intervals is nonempty, then the procedure recursively repeats the previous steps until the whole interval $(0,1)$ has been completely searched. When all a^*'s are found, we pick the one that maximizes the mutual information. Note that this fixed point method is much faster than an exhaustive search through all the values of a. Finally, we note that our procedure is based on the algorithm in [30].

Next, we state a sufficient condition for which a^* is unique.

Corollary 3. Let $d(x, y)$ is an arbitrary distance metric between x and y. If there exists a $q \in [0, 1)$ such that for all $x, y \in (0, 1)$

$$d(c(x), c(y)) \leq q d(x, y), \tag{38}$$

then there exists a unique a^* such that $c(a^*) = a^*$.

Proof. From Theorem 4, obviously that $a \in (0, 1)$ and $c(a) \in (0, 1)$. Thus, $c(a)$ maps to itself. If existing q and $d(,)$ such that $d(c(x), c(y)) \leq q d(x, y)$ for all $x, y \in (0, 1)$ then $c(.)$ is a contraction mapping. From Banach's fixed point theorem [31], there exists a unique a^* such that $c(a^*) = a^*$. □

Note that if we use $d(x, y) = |x - y|$, then it is straight forward to show that if $0 < c'(a) < 1$, then a^* is unique.

VI. Conclusion

In this paper, we show that if the ratio of the channel conditional densities of the inputs $r(y) = \frac{P(Y=y|X=0)}{P(Y=y|X=1)}$ is a strictly increasing or decreasing function, then the quantizers having a single threshold are optimal. Furthermore, we show that an optimal quantizer (possibly with multiple thresholds) is the one with the thresholding vector whose elements are all the solutions of $r(y) = r^*$ for some constant $r^* > 0$. We also describe a necessary condition for optimality, a sufficient condition for uniqueness via a fixed point theorem, together with an algorithm for finding the globally optimal quantizer.

Appendix

A. Proof for Lemma 1

From (25), we have:

$$\phi_1(h_i) = \frac{a p_0}{(1-a) p_1} \phi_0(h_i), \forall i \in \{1, 2, \ldots, n\}. \tag{39}$$

Now, suppose that $u(y) = a$ having n solutions $\{h_1, h_2, \ldots, h_n\}$. Without loss of generality, suppose that $\mathbb{H}_a = \{(-\infty, h_1) \cup [h_2, h_3) \cup \cdots \cup [h_n, +\infty)\}$ and $\bar{\mathbb{H}}_a = \mathbb{R} \setminus \mathbb{H}_a = \{[h_1, h_2) \cup [h_3, h_4) \cup \cdots \cup [h_{n-1}, h_n)\}$. From (29) and (30)

$$\frac{df(a)}{da} = \frac{\partial f(a)}{\partial h} \frac{\partial h}{\partial a} = + \phi_0(h_1) \frac{\partial h_1}{\partial a} - \phi_0(h_2) \frac{\partial h_2}{\partial a} + \ldots - \phi_0(h_n) \frac{\partial h_n}{\partial a}, \tag{40}$$

$$\frac{dg(a)}{da} = \frac{\partial g(a)}{\partial h} \frac{\partial h}{\partial a} = - \phi_1(h_1) \frac{\partial h_1}{\partial a} + \phi_1(h_2) \frac{\partial h_2}{\partial a} - \ldots + \phi_1(h_n) \frac{\partial h_n}{\partial a}. \tag{41}$$

Combining Eqs. (39), (40) and (41), we have the desired proof. We note that $f'(a)$ and $g'(a)$ have the opposite sign. As a result, if $f(a)$ increases, then $g(a)$ decreases and vice-versa.

■

B. Proof for Lemma 2

(1) From (26), $f(a)$ represents the quantized bit "0" which is the area of $u(y)$ (defined in (27)) where $u(y) < a$. Therefore, if a is increasing, $f(a)$ is obviously increasing. Thus, $f'(a) > 0$. A similar proof can be established for $g(a)$ which corresponds to the area of $u(y)$ where $u(y) \geq a$.

(2) We note that $f(a)$ and $g(a)$ represent the quantized bits "0" and "1" which correspond to the areas of $u(y) < a$ and $u(y) \geq a$, respectively. Let $\mathbb{H}_a = \{y | u(y) < a\}$ and $\bar{\mathbb{H}}_a = \{y | u(y) \geq a\}$. From (26)

$$a p_0 \phi_0(y) > (1-a) p_1 \phi_1(y), \forall y \in \mathbb{H}_a, \tag{42}$$

$$a p_0 \phi_0(y) \leq (1-a) p_1 \phi_1(y), \forall y \in \bar{\mathbb{H}}_a. \tag{43}$$

We consider two possible cases: $a > p_1$ and $a \leq p_1$. In both cases, we will show that $f(a) + g(a) > 1$.

- If $a < p_1$ then $1 - a > 1 - p_1 = p_0$. Thus, from (42), $\phi_0(y) > \phi_1(y)$ for $\forall\ y \in \mathbb{H}_a$. Therefore,

$$f(a) + g(a) = \int_{y \in \mathbb{H}_a} \phi_0(y) dy + \int_{y \in \bar{\mathbb{H}}_a} \phi_1(y) dy \tag{44}$$

$$> \int_{y \in \mathbb{H}_a} \phi_1(y) dy + \int_{y \in \bar{\mathbb{H}}_a} \phi_1(y) dy \tag{45}$$

$$= 1. \tag{46}$$

- If $a \geq p_1$ then $1 - a \leq 1 - p_1 = p_0$. Thus, from (43), $\phi_0(y) \leq \phi_1(y)$ for $\forall\ y \in \bar{\mathbb{H}}_a$. Therefore,

$$f(a) + g(a) = \int_{y \in \mathbb{H}_a} \phi_0(y) dy + \int_{y \in \bar{\mathbb{H}}_a} \phi_1(y) dy \tag{47}$$

$$\geq \int_{y \in \mathbb{H}_a} \phi_0(y) dy + \int_{y \in \bar{\mathbb{H}}_a} \phi_0(y) dy \tag{48}$$

$$= 1. \tag{49}$$

■

Remark: The necessary condition for inequality (49) becomes equality is $\phi_0(y) = \phi_1(y)$ for $\forall\ y \in \bar{\mathbb{H}}_a$ that contradicts to the assumption that $r(y)$ has a finite number of stationary points. Thus, $f(a) + g(a) > 1$.

C. Proof of Lemma 3

By using the definitions of $\mathbf{a}, \mathbf{l}_a, \mathbf{r}_a$ and KL divergence, it can be shown that (54) holds with (50) due to $a \log a + (1-a) \log(1-a)$ is cancelled after summing up, (51), (52) and (53) due to a bit of algebra, (54) due to the definition of $F(a)$ in (35).

D. Proof Theorem 4

We will use the following lemmas and the order notion of 2-dimensional vector below to prove Theorem 4.

Vector Order. Consider two binary probability vectors $\mathbf{x} = [1-x, x]$ and $\mathbf{y} = [1-y, y]$, $x, y \in (0, 1)$, we define the vector order $\mathbf{y} \geq \mathbf{x}$ if and only if $y \geq x$.

Lemma 6. For any three binary probabiity vectors $\mathbf{a} = [1-a, a]$, $\mathbf{b} = [1-b, b]$ and $\mathbf{c} = [1-c, c]$ such that $\mathbf{a} \leq \mathbf{b} \leq \mathbf{c}$ (or $a \leq b \leq c$), then

- (a) $D_{KL}(\mathbf{a}||\mathbf{b}) \leq D_{KL}(\mathbf{a}||\mathbf{c})$

$$
\begin{aligned}
& \frac{1}{a}\big[D_{KL}(\mathbf{a}||\mathbf{l}_a) - D_{KL}(\mathbf{a}||\mathbf{r}_a)\big] \\
= \; & \frac{1}{a}\big[\big(a\log(\frac{a}{\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}}) + (1-a)\log(\frac{1-a}{\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}})\big) \\
- \; & \big(a\log(\frac{a}{\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}) + (1-a)\log(\frac{1-a}{\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}})\big)\big] \\
= \; & \frac{1}{a}\big[-\big(a\log(\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}) + (1-a)\log(\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))})\big) \\
+ \; & \big(a\log(\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}) + (1-a)\log(\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)})\big)\big] \qquad (50) \\
= \; & \frac{1}{a}\big[(1-a)\log(\frac{p_0(1-f(a))}{p_0f(a)}) + a\log(\frac{p_1g(a)}{p_1(1-g(a))}) + (1-a+a)\log(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)})\big] \qquad (51) \\
= \; & \frac{1}{a}\big[(a-1)\log(\frac{f(a)}{1-f(a)}) + a\log(\frac{g(a)}{1-g(a)}) + \log(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)})\big] \qquad (52) \\
= \; & \frac{a-1}{a}\log(\frac{f(a)}{1-f(a)}) + \log(\frac{g(a)}{1-g(a)}) + \frac{1}{a}\log(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)}) \qquad (53) \\
= \; & F(a), \qquad (54)
\end{aligned}
$$

- (b) $D_{KL}(\mathbf{c}||\mathbf{b}) \leq D_{KL}(\mathbf{c}||\mathbf{a})$
- (c) $D_{KL}(\mathbf{b}||\mathbf{a}) \leq D_{KL}(\mathbf{c}||\mathbf{a})$
- (d) $D_{KL}(\mathbf{b}||\mathbf{c}) \leq D_{KL}(\mathbf{a}||\mathbf{c})$

Proof. Proof of (a). For a given $\mathbf{a}$, we show that $D_{KL}(\mathbf{a}||\mathbf{b})$ is a non-decreasing function of b. Let $D(b) = D_{KL}(\mathbf{a}||\mathbf{b}) = a\log(\frac{a}{b}) + (1-a)\log(\frac{1-a}{1-b})$

$$D'(b) = \frac{1-a}{1-b} - \frac{a}{b}. \qquad (55)$$

Since $a \leq b$ then $1-a \geq 1-b$, thus $\frac{1-a}{1-b} \geq 1 \geq \frac{a}{b}$ and $D'(b) \geq 0 \; \forall \; b \geq a$. Since $b \leq c$, $D(b) \leq D(c)$ or $D_{KL}(\mathbf{a}||\mathbf{b}) \leq D_{KL}(\mathbf{a}||\mathbf{c})$. The equality happens if and only if $b = c$.

We omit the proofs of (b), (c), and (d) since they are similar to the proof of (a). □

Lemma 7. If $D_{KL}(\mathbf{c}_a||\mathbf{l}_a) = D_{KL}(\mathbf{c}_a||\mathbf{r}_a)$, then $\mathbf{l}_a \leq \mathbf{c}_a \leq \mathbf{r}_a$

Proof. First, we show that $\mathbf{l}_a < \mathbf{r}_a$, $\forall$ a. Indeed, consider

$$
\begin{aligned}
& \frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)} - \frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))} \\
= \; & \frac{p_0p_1(g(a)f(a) - (1-g(a))(1-f(a)))}{(p_0(1-f(a))+p_1g(a))(p_0f(a)+p_1(1-g(a)))} \\
= \; & \frac{p_0p_1(f(a)+g(a)-1)}{(p_0(1-f(a))+p_1g(a))(p_0f(a)+p_1(1-g(a)))} \\
> \; & 0,
\end{aligned}
$$

where the last inequality is due to $f(a)+g(a) > 1$ (Lemma 2), and all other terms in the last equation are positive. Thus, the second entry of $\mathbf{r}_a$ is strictly greater than the second entry of $\mathbf{l}_a$ or $\mathbf{r}_a > \mathbf{l}_a$.

Now, suppose that $\mathbf{c}_a < \mathbf{l}_a < \mathbf{r}_a$, by Lemma 6 part (a), $D_{KL}(\mathbf{c}_a||\mathbf{l}_a) < D_{KL}(\mathbf{c}_a||\mathbf{r}_a)$ that contradicts to $D_{KL}(\mathbf{c}_a||\mathbf{l}_a) = D_{KL}(\mathbf{c}_a||\mathbf{r}_a)$. Thus, $\mathbf{l}_a \leq \mathbf{c}_a$. A similar proof can be constructed to show that $\mathbf{c}_a \leq \mathbf{r}_a$. Thus, $\mathbf{l}_a \leq \mathbf{c}_a \leq \mathbf{r}_a$. □

Lemma 8. Consider a_1 and a_2 such that $0 < a_1 \leq a_2 < 1$, then $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2}$ and $\mathbf{r}_{a_1} \leq \mathbf{r}_{a_2}$.

Proof. First, we show that $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2}$. Indeed, consider the function $s(a)$ as the ratio of the second entry over the first entry of $\mathbf{l}_a$, i.e., $s(a) = \frac{p_1(1-g(a))}{p_0f(a)}$. We have

$$
\begin{aligned}
s'(a) \; & = \; \frac{-p_1g'(a)p_0f(a) - p_1(1-g(a))p_0f'(a)}{(p_0f(a))^2} \\
& = \; p_0p_1f'(a)\big(\frac{ap_0}{(1-a)p_1}f(a) - (1-g(a))\big), \qquad (56)
\end{aligned}
$$

with (56) due to Lemma 2. Also from (42),

$$\phi_1(y) < \frac{ap_0}{(1-a)p_1}\phi_0(y), \forall y \in \mathbb{H}_a.$$

Moreover, from the definitions of $f(a)$ and $g(a)$ in (29) and (30), $f(a)$ and $1-g(a)$ are the integrals of $\phi_0(y)$ and $\phi_1(y)$, respectively over $\mathbb{H}_a$, respectively. Thus, $\frac{ap_0}{(1-a)p_1}f(a) - (1-g(a)) > 0$. From Lemma 2 $f'(a) > 0$, thus $s'(a) > 0$. That said, the ratio of the second entry over the first entry of $\mathbf{l}_a$ is an increasing function of a. Furthermore, $\mathbf{l}_a$ is a probability vector, i.e., the summation of the first entry and the second entry equals one. Therefore, the second entry of $\mathbf{l}_a$ is an increasing function of a or $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2}$.

A similar proof can be constructed to show that $\mathbf{r}_{a_1} \leq \mathbf{r}_{a_2}$. □

Lemma 9. Consider 4 vectors $\mathbf{a} = [1-a, a]$, $\mathbf{b} = [1-b, b]$, $\mathbf{c} = [1-c, c]$ and $\mathbf{d} = [1-d, d]$ such that $\mathbf{a} \leq \mathbf{b} \leq \mathbf{c} \leq \mathbf{d}$ (or $a \leq b \leq c \leq d$), then

- (a) $D_{KL}(\mathbf{d}||\mathbf{a}) \geq D_{KL}(\mathbf{c}||\mathbf{b})$.
- (b) $D_{KL}(\mathbf{a}||\mathbf{d}) \geq D_{KL}(\mathbf{b}||\mathbf{c})$.

Proof. Proof of (a). We have $D_{KL}(\mathbf{c}||\mathbf{b}) \le D_{KL}(\mathbf{c}||\mathbf{a})$ and $D_{KL}(\mathbf{c}||\mathbf{a}) \le D_{KL}(\mathbf{d}||\mathbf{a})$ due to Lemma 6 part (b) and (c), respectively. Thus, $D_{KL}(\mathbf{d}||\mathbf{a}) \ge D_{KL}(\mathbf{c}||\mathbf{b})$. The equality happens if and only if $a = b$ and $c = d$.

Proof of (b). Similar to proof of part (a), $D_{KL}(\mathbf{a}||\mathbf{d}) \ge D_{KL}(\mathbf{b}||\mathbf{d})$ and $D_{KL}(\mathbf{b}||\mathbf{d}) \ge D_{KL}(\mathbf{b}||\mathbf{c})$ due to Lemma 6 part (a) and (d), respectively. Thus, $D_{KL}(\mathbf{a}||\mathbf{d}) \ge D_{KL}(\mathbf{b}||\mathbf{c})$. The equality happens if and only if $a = b$ and $c = d$. □

Now, we are ready to prove Theorem 4.

Proof of $c(a) \in (0, 1)$.

From Lemma 7, we have $\mathbf{l}_a \le \mathbf{c}_a \le \mathbf{r}_a$. Equivalently,

$$0 < \frac{p_1(1-g(a))}{p_0 f(a) + p_1(1-g(a))} \le c(a) \le \frac{p_1 g(a)}{p_0(1-f(a)) + p_1 g(a)} < 1. \tag{57}$$

Proof for the smoothness of $c(a)$. Since $0 < f(a), g(a) < 1$, $p_0(1-f(a))+p_1 g(a) > 0$ and $f(a)g(a) > 0$, thus all of the denominators of (36) is positive. In addition, one can verify that

$$(1-f(a))(1-g(a)) = 1-f(a)-g(a)+f(a)g(a) < f(a)g(a).$$

Thus, $\log(\frac{(1-f(a))(1-g(a))}{f(a)g(a)})$ is non-zero. In addition, if $f'(a)$ and $g'(a)$ exist, it is straight forward to show that $c'(a)$ also exists. Therefore, $c(a)$ is a well-defined and smooth function of a.

Proof for the non-decreasing of $c(a)$.

Suppose that there exists $a_1 \le a_2$ such that $D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) = D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1})$ and $D_{KL}(\mathbf{c}_{a_2}||\mathbf{l}_{a_2}) = D_{KL}(\mathbf{c}_{a_2}||\mathbf{r}_{a_2})$ but $c(a_1) > c(a_2)$. From Lemma 8, $\mathbf{l}_{a_1} \le \mathbf{l}_{a_2}$, $\mathbf{r}_{a_1} \le \mathbf{r}_{a_2}$. From Lemma 7, $\mathbf{l}_{a_1} \le \mathbf{c}_{a_1} \le \mathbf{r}_{a_1}$ and $\mathbf{l}_{a_2} \le \mathbf{c}_{a_2} \le \mathbf{r}_{a_2}$. From the assumption that $c_{a_1} > c_{a_2}$, $\mathbf{c}_{a_1} > \mathbf{c}_{a_2}$. Therefore,

$$\mathbf{l}_{a_1} \le \mathbf{l}_{a_2} \le \mathbf{c}_{a_2} < \mathbf{c}_{a_1} \le \mathbf{r}_{a_1} \le \mathbf{r}_{a_2}.$$

Now, using Lemma 9 part (a) for $\mathbf{l}_{a_1} \le \mathbf{l}_{a_2} \le \mathbf{c}_{a_2} < \mathbf{c}_{a_1}$,

$$D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) > D_{KL}(\mathbf{c}_{a_2}||\mathbf{l}_{a_2}). \tag{58}$$

Similarly, using Lemma 9 part (b) for $\mathbf{c}_{a_2} < \mathbf{c}_{a_1} \le \mathbf{r}_{a_1} \le \mathbf{r}_{a_2}$,

$$D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1}) < D_{KL}(\mathbf{c}_{a_2}||\mathbf{r}_{a_2}). \tag{59}$$

From (58) and (59),

$$D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) > D_{KL}(\mathbf{c}_{a_2}||\mathbf{l}_{a_2}) = D_{KL}(\mathbf{c}_{a_2}||\mathbf{r}_{a_2}) > D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1})$$

that contradicts to our assumption that $D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) = D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1})$. By contradiction method, $c(a_1) \le c(a_2)$ if $a_1 \le a_2$. Thus, $c(a)$ is a non-decreasing function of a. Combining with (57), we have the proof for Theorem 4.

E. Proof of $r^* > 0$

From (11) and $p_0 > 0$, $p_1 > 0$, $r^* > 0$ is equivalent to

$$-\frac{\log\dfrac{1-q_0}{q_0} + \log\dfrac{1-A_{22}}{A_{22}}}{\log\dfrac{1-q_0}{q_0} - \log\dfrac{1-A_{11}}{A_{11}}} > 0.$$

Thus, we need to show that

$$\log(\frac{1-q_0}{q_0}\frac{1-A_{22}}{A_{22}})\log(\frac{q_0}{1-q_0}\frac{1-A_{11}}{A_{11}}) > 0.$$

Since $\log(x) > 0$ if and only if $x > 1$, we can show that

$$(\frac{1-q_0}{q_0}\frac{1-A_{22}}{A_{22}} - 1)(\frac{q_0}{1-q_0}\frac{1-A_{11}}{A_{11}} - 1) > 0.$$

Using a bit of algebra, (60) is equivalent to

$$(A_{11} - q_0)(A_{22} - q_1) > 0. \tag{60}$$

However, $A_{11} = f(a)$, $A_{22} = g(a)$, thus $A_{11} + A_{22} > 1$ by Lemma 2. From $A_{21} + A_{22} = 1 < A_{11} + A_{22}$, $A_{21} < A_{11}$. Similarly, $A_{12} < A_{22}$. Therefore,

$$q_0 = p_0 A_{11} + p_1 A_{21} < p_0 A_{11} + p_1 A_{11} = A_{11}, \tag{61}$$

$$q_1 = p_0 A_{12} + p_1 A_{22} < p_0 A_{22} + p_1 A_{22} = A_{22}. \tag{62}$$

Combining (61) and (62), (60) follows. The proof is complete.

References

[1] Morris Goldberg, P Boucher, and Seymour Shlien. Image compression using adaptive vector quantization. *IEEE Transactions on Communications*, 34(2):180–187, 1986.

[2] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*, 2014.

[3] Lale Akarun, Y Yardunci, and A Enis Cetin. Adaptive methods for dithering color images. *IEEE transactions on image processing*, 6(7):950–955, 1997.

[4] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[5] Thuan Nguyen and Thinh Nguyen. On closed form capacities of discrete memoryless channels. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2018.

[6] Francisco Javier Cuadros Romero and Brian M Kurkoski. Decoding ldpc codes with mutual information-maximizing lookup tables. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 426–430. IEEE, 2015.

[7] Jiadong Wang, Thomas Courtade, Hari Shankar, and Richard D Wesel. Soft information for ldpc decoding in flash: Mutual-information optimized quantization. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–6. IEEE, 2011.

[8] Ido Tal and Alexander Vardy. How to construct polar codes. *arXiv preprint arXiv:1105.6164*, 2011.

[9] Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, 1960.

[10] Joel G Smith. The information capacity of amplitude-and variance-constrained sclar gaussian channels. *Information and Control*, 18(3):203–219, 1971.

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

[12] Rudolf Mathar and Meik Dörpinghaus. Threshold optimization for capacity-achieving discrete input one-bit output quantization. In *2013 IEEE International Symposium on Information Theory*, pages 1999–2003. IEEE, 2013.

[13] Yuta Sakai and Ken-ichi Iwata. Suboptimal quantizer design for outputs of discrete memoryless channels with a finite-input alphabet. In *2014 International Symposium on Information Theory and its Applications*, pages 120–124. IEEE, 2014.

[14] 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.

[15] Andreas Winkelbauer, Gerald Matz, and Andreas Burg. Channel-optimized vector quantization with mutual information as fidelity criterion. In *Signals, Systems and Computers, 2013 Asilomar Conference on*, pages 851–855. IEEE, 2013.

[16] Tobias Koch and Amos Lapidoth. At low snr, asymmetric quantizers are better. *IEEE Trans. Information Theory*, 59(9):5421–5445, 2013.

[17] Xuan He, Kui Cai, Wentu Song, and Zhen Mei. Dynamic programming for discrete memoryless channel quantization. *CoRR*, abs/1901.01659, 2019.

[18] David Burshtein, Vincent Della Pietra, Dimitri Kanevsky, Arthur Nadas, et al. Minimum impurity partitions. *The Annals of Statistics*, 20(3):1637–1646, 1992.

[19] Brian M Kurkoski and Hideki Yagi. Single-bit quantization of binary-input, continuous-output channels. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2088–2092. IEEE, 2017.

[20] Don Coppersmith, Se June Hong, and Jonathan RM Hosking. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197–217, 1999.

[21] Jiuyang Alan Zhang and Brian M Kurkoski. Low-complexity quantization of discrete memoryless channels. In *2016 International Symposium on Information Theory and Its Applications (ISITA)*, pages 448–452. IEEE, 2016.

[22] Bobak Nazer, Or Ordentlich, and Yury Polyanskiy. Information-distilling quantizers. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 96–100. IEEE, 2017.

[23] Eduardo S Laber, Marco Molinaro, and Felipe A Mello Pereira. Binary partitions with approximate minimum impurity. In *International Conference on Machine Learning*, pages 2860–2868, 2018.

[24] Ferdinando Cicalese, Eduardo Laber, and Lucas Murtinho. New results on information theoretic clustering. In *International Conference on Machine Learning*, pages 1242–1251, 2019.

[25] Thuan Nguyen, Yu-Jung Chu, and Thinh Nguyen. On the capacities of discrete memoryless thresholding channels. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2018.

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

[27] Jaspreet Singh, Onkar Dabeer, and Upamanyu Madhow. On the limits of communication with low-precision analog-to-digital conversion at the receiver. *IEEE Trans. Communications*, 57(12):3629–3639, 2009.

[28] Brendan Mumey and Tomáš Gedeon. Optimal mutual information quantization is np-complete. In *Proc. Neural Inf. Coding (NIC) Workshop*, 2003.

[29] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[30] Xiaolin Wu. Optimal quantization by matrix searching. *Journal of algorithms*, 12(4):663–673, 1991.

[31] William A Kirk and Brailey Sims. *Handbook of metric fixed point theory*. Springer Science & Business Media, 2013.

Thuan Nguyen received a B.S. degree in Electrical Engineering (honors program) from Post and Telecommunication Institute of Technology, Vietnam, in 2013. He is currently a Ph.D. student at Oregon State University, Corvallis, Oregon, USA. His research interests include information theory, signal processing and machine learning.

Thinh Nguyen (M04) received the B.S. degree from the University of Washington, Seattle, WA, USA, in 1995 and the Ph.D. degree from the University of California, Berkeley, CA, USA, in 2003, both in electrical engineering. He is currently a Professor with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA. He is interested in all things stochastic, with applications to signal processing, distributed systems, wireless networks, network coding, and quantum walks.

Dr. Nguyen has served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the IEEE TRANSACTIONS ON MULTIMEDIA