Optimal Quantizer Structure For Maximizing Mutual Information Under Constraints

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

***Abstract*—Consider a channel whose the input alphabet set $\mathbb{X} = \{x_1, x_2, \ldots, x_K\}$ contains K discrete symbols modeled as a discrete random variable X having a probability mass function $\mathbf{p}(\mathbf{x}) = [p(x_1), p(x_2), \ldots, p(x_K)]$ and the received signal Y being a continuous random variable. Y is a distorted version of X caused by a channel distortion, characterized by the conditional densities $p(y|x_i) = \phi_i(y)$, $i = 1, 2, \ldots, K$. To recover X, a quantizer Q is used to quantize Y back to a discrete output $\mathbb{Z} = \{z_1, z_2, \ldots, z_N\}$ corresponding to a random variable Z with a probability mass function $\mathbf{p}(\mathbf{z}) = [p(z_1), p(z_2), \ldots, p(z_N)]$ such that the mutual information $I(X;Z)$ is maximized subject to an arbitrary constraint on $\mathbf{p}(\mathbf{z})$. Formally, we are interested in designing an optimal quantizer Q^* that maximizes $\beta I(X;Z) - C(Z)$ where β is a positive number that controls the trade-off between maximizing $I(X;Z)$ and minimizing an arbitrary cost function $C(Z)$. Let $\mathbf{p}(\mathbf{x}|y) = [p(x_1|y), p(x_2|y), \ldots, p(x_K|y)]$ be the posterior distribution of X for a given value of y, we show that for any arbitrary cost function $C(.)$, the optimal quantizer Q^* separates the vectors $\mathbf{p}(\mathbf{x}|y)$ into convex regions. Using this result, a method is proposed to determine an upper bound on the number of thresholds (decision variables on y) which is used to speed up the algorithm for finding an optimal quantizer. Numerical results are presented to validate the findings.**

***Index Terms*—Channel quantization, mutual information, channel capacity, threshold, constraints.**

I. Introduction

Motivated by the development of polar codes [1] and LDPC codes [2], finding optimal quantizers that maximize the mutual information between the input and output has been a topic of interest in recent years. Many practical algorithms and theoretical results for such optimal quantizers have been proposed over the past decade [3]–[9]. Finding an optimal quantizer that maximizes the mutual information in a general setting is an NP-hard problem [10]. Consequently, using an exhaustive search is intractable even for the modest size of the input and output sets. Therefore, existing algorithms typically find an approximate solution [4], [5], [7]. On the other hand, under certain restrictions e.g., binary input channel, there exist polynomial-time algorithms [3], [6], [11] for finding the exact solution.

While there exist many exact and approximate algorithms for finding an optimal quantizer that maximizes the mutual information between the input and output under different settings, the problem of finding an optimal quantizer that maximizes the mutual information subject to some constraints on the output, receives less attention. In this paper, we are interested in studying the optimal quantizers in the following communication setting. We consider a sender transmits K discrete symbols $\mathbb{X} = \{x_1, x_2, \ldots, x_K\}$ modeled as a discrete random variable X having a probability mass function $\mathbf{p}(\mathbf{x}) = [p(x_1), p(x_2), \ldots, p(x_K)]$ over an arbitrary continuous channel. As such, the received signal Y is a distorted version of X caused by the channel distortion that is characterized by the conditional densities $p(y|x_i) = \phi_i(y)$, $i = 1, 2, \ldots, K$. To recover X, a quantizer Q is used to quantize Y back to a discrete output $\mathbb{Z} = \{z_1, z_2, \ldots, z_N\}$ corresponding to a random variable Z with a probability mass function $\mathbf{p}(\mathbf{z}) = [p(z_1), p(z_2), \ldots, p(z_N)]$ such that the mutual information $I(X;Z)$ is maximized subject to an arbitrary constraint on $\mathbf{p}(\mathbf{z})$. Formally, we are interested in designing an optimal quantizer Q^* that maximizes $\beta I(X;Z) - C(Z)$ where β is a positive number that controls the trade-off between maximizing $I(X;Z)$ and minimizing an arbitrary cost function $C(Z)$.

This problem is a generalized version of the Deterministic Information Bottleneck [12], and has many applications. Specifically, using the entropy constraint on Z, our problem is exactly the DIB. Imposing entropy constraint on Z is useful in many applications that use low-bandwidth channels or limited storage systems. For example, suppose one wants to quantize a continuous data source before applying entropy coding, e.g., Huffman code, to gain compression. Ideally, one wants to minimize the distortion between the original continuous data and the quantized data. However, minimizing the distortion may result in a high entropy of the quantized data which may exceed a given storage capacity after compression. Thus, one needs to impose a constraint on the entropy of the quantized data to guarantee that the size of the resulted compressed data is below the storage capacity while retaining much information in the original source. Similarly, if the quantized data must be transmitted over a limited bandwidth channel, it is important to reduce the entropy of the data source below a certain threshold in order to reduce the bit rate to match the limited channel bandwidth.

To that end, the contributions of this paper are as follows. We showed that there exists a convex quantizer that is optimal. Specifically, let $\mathbf{p}(\mathbf{x}|y) = [p(x_1|y), p(x_2|y), \ldots, p(x_K|y)]$ be the posterior distribution of X for a given value of y, we show that for any arbitrary cost function $C(.)$, the optimal quantizer Q^* separates the vectors $\mathbf{p}(\mathbf{x}|y)$ into convex cells. Although using a different approach, our result is similar to the result previously established for the quantization problems without the constraint [3], [13]. In particular, we show that for any given quantizer $Q(y)$, there exists a convex quantizer $\tilde{Q}(y)$ such that: (1) $\tilde{Q}(y)$ produces the same $\mathbf{p}(\mathbf{z})$ as that of $Q(y)$, therefore, the same cost function $C(Z)$, and (2) $I(X;Z)$

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

produced by $\tilde{Q}(y)$ is at least as large as that produced by $Q(y)$. Therefore, a class of convex quantizers should contain at least one optimal quantizer. In addition, using this result, we describe a method for determining an upper bound on the number of thresholds used in a convex quantizer, which narrows down the search space for finding an optimal quantizer. Numerical results are presented to validate the findings.

The outline of our paper is as follows. Section II is the related work. We describe the problem formulation in Section III. All the notations, definitions, preliminary results are introduced in Section IV. Section V investigates the optimal quantizer's structure. An upper bound on the number of optimal thresholds is constructed in Section VI. The numerical results can be found in Section VII. Finally, we provide a few concluding remarks in Section VIII.

II. Related work

When the input is binary, it can be shown that an optimal quantizer (without output constraints) has the structure of convex cells in the space of posterior distribution [3], [6], [11]. Based on this optimality structure, an optimal quantizer can be found efficiently in polynomial time via dynamic programming technique [3]. In particular, the structure of optimal binary-input quantizers in [3] and [6] is constructed based on the well-known result in [13] for the K-ary inputs. The results in [13] and [14] showed that for K-ary input, an optimal quantizer separates space of the posterior probability distribution into convex cells via a number of hyper-plane cuts. The number of hyper-plane cuts can be shown to be polynomial in the data size. Thus, there exists a polynomial time algorithm to find an optimal quantizer by exhaustively searching over all the possible hyper-plane cuts in the posterior distribution space [13].

There also exist a few results on finding a quantizer that maximizes the mutual information subject to some constraints on the output. Finding an optimal quantizer for maximizing/minimizing an objective function other than the mutual information subject to certain output constraints, has a long history. For example, the problem of entropy-constrained scalar quantization [15], [16] and entropy-constrained vector quantization [17], [18] have been well established. The objectives in these problems are minimizing a specific distortion function, typically the mean square error (MSE) between the input and the output while keeping the output entropy less than a certain threshold. The imposed entropy constraint is crucial in applications that use limited communication channels and limited storage systems. Notably, the Deterministic Information Bottleneck (DIB) method of Strouse et al. [12] is most related to our work. Strouse et al. proposed a linear time iterative algorithm to find a locally optimal quantizer that maximizes the mutual information under the entropy constraint of the output. On the other hand, our work is focused on the structure of the optimal quantizer, and can find the exact solution albeit with higher complexity. Our results also generalize the result in [13] for the problem of minimizing impurity without constraints. Specifically, the result in [13] states that the optimal partitions are separated by hyper-plane cuts in the space of the posterior distribution. We show that this structure is also valid for the problem of maximizing mutual information subject to any output constraints. Our proposed approach also relates closely to the work of Gyorgy and Linder [15], [16] which constructed the optimal structure of entropy-constrained scalar quantization. Finally, we note that a part of our work was presented in [19].

III. Problem Formulation

We consider a discrete input source $\mathbb{X} = \{x_1, x_2, \ldots, x_K\}$ modeled as a discrete random variable X consisting K discrete symbols with a given p.m.f $\mathbf{p}(\mathbf{x}) = [p(x_1), p(x_2), \ldots, p(x_K)]$. x_i is transmitted over a given arbitrary continuous channel that distorts/maps x_i to a continuous value $y \in \mathbb{R}$ at the receiver. Let Y be a random variable that models the received signal, then the channel distortion is characterized by K conditional densities $p(y|x_i) = \phi_i(y)$, $i = 1, 2, \ldots, K$. A quantizer Q is used to map the continuous random variable Y to a discrete output $\mathbb{Z} = \{z_1, z_2, \ldots, z_N\}$ corresponding to a discrete random variable Z consisting of N discrete outcomes $z_1, z_2, \ldots, z_N$ with the p.m.f $\mathbf{p}(\mathbf{z}) = [p(z_1), p(z_2), \ldots, p(z_N)]$. We note that $\mathbf{p}(\mathbf{z})$ depends on Q. Let $C(Z)$ be an arbitrary cost function of $\mathbf{p}(\mathbf{z})$. Our goal is to find an optimal quantizer Q^* that maximizes the trade-off between the mutual information $I(X;Z)$ and the cost function $C(Z)$. Formally, we want to solve the following optimization problem:

$$Q^* = \max_{Q} \beta I(X;Z) - C(Z), \tag{1}$$

where β is a pre-specified positive number that controls the trade-off between maximizing $I(X;Z)$ and minimizing $C(Z)$. A well-known constraint $C(Z)$ is the entropy $H(Z) = -\sum_{i=1}^{N} p(z_i) \log p(z_i)$ which we will be used to validate our findings in Section VII.

IV. Preliminaries

A. Notations and definitions

For convenience, we use the following notations and definitions:

1) $\mathbf{p}(\mathbf{x}) = [p(x_1), p(x_2), \ldots, p(x_K)] = [p_1, p_2, \ldots, p_K]$ denotes the p.m.f of the input random variable X.

2) $p(y|x_i) = \phi_i(y)$, $i = 1, 2, \ldots, K$ denotes the conditional density of received-output y for a given transmitted input x_i. Unlike a AWGN channel, $\phi_i(y)$ and $\phi_j(y)$ can be quite different as the channel may distort signals x_i and x_j differently. We assume that $\phi_i(y)$ is a continuous, positive, and differentiable function.

3) $p(y)$ denotes the density function of y. Specifically,

$$p(y) = \sum_{i=1}^{K} p_i \phi_i(y). \tag{2}$$

4) $\mathbf{p}(\mathbf{x}|y) = [p(x_1|y), p(x_2|y), \ldots, p(x_K|y)]$ denotes the conditional probability vector of X given a $y \in Y$ where,

$$p(x_i|y) = \frac{p_i \phi_i(y)}{\sum_{j=1}^{K} p_j \phi_j(y)}. \tag{3}$$

5) The output set $\mathbb{Z}_i$ denotes the set of y's that is mapped to the i^{th} output z_i by $Q(y)$. Formally,

$$\mathbb{Z}_i = \{y : Q(y) = z_i\}. \quad (4)$$

Definition 1. (**Convex quantizer** (***quañtizer***)) Let $\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_N$ be the N sets induced by a quantizer $Q(y)$. $\tilde{Q}(y)$ is a convex quantizer (denoted by *quañtizer*) if for any $\mathbb{Z}_i$ and $\mathbb{Z}_j$, $i \neq j$, there exists a hyper-plane that separates the two conditional probability vectors $\mathbf{p}(\mathbf{x}|y_i)$ and $\mathbf{p}(\mathbf{x}|y_j)$, $\forall y_i \in \mathbb{Z}_i, \forall y_j \in \mathbb{Z}_j$.

We note that a *quañtizer* produces the N convex regions in the K dimensional space of the posterior distribution $\mathbf{p}(\mathbf{x}|y)$, but not the N convex regions in y.

Definition 2. (**Kullback-Leibler (KL) divergence**) The KL divergence of two probability vectors $\mathbf{a} = (a_1, a_2, \ldots, a_K)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_K)$ is defined by:

$$D(\mathbf{a}||\mathbf{b}) = \sum_{i=1}^{K} a_i \log(\frac{a_i}{b_i}). \quad (5)$$

Definition 3. (**Centroid**) The centroid of output set $\mathbb{Z}_i$ is a K-dimensional probability vector $\mathbf{c}_i = [c_i^1, c_i^2, \ldots, c_i^K]$ that minimizes the total KL divergence from $\mathbf{p}(\mathbf{x}|y)$ to $\mathbf{c}_i$, $\forall y \in \mathbb{Z}_i$. Formally,

$$\mathbf{c}_i = \arg\min_{\mathbf{c}} \int_{y \in \mathbb{Z}_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}) p(y) dy. \quad (6)$$

Definition 4. (**Distortion measurement**) The distortion of a quantizer Q that induces N output sets $\{\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_N\}$ is:

$$D(Q) = \sum_{i=1}^{N} D(Q_{\mathbb{Z}_i}) = \sum_{i=1}^{N} \int_{y \in \mathbb{Z}_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i) p(y) dy, \quad (7)$$

where $\mathbf{c}_i$ is the centroid of $\mathbb{Z}_i$ and $D(Q_{\mathbb{Z}_i})$ is the distortion induced for each $\mathbb{Z}_i$,

$$D(Q_{\mathbb{Z}_i}) = \int_{y \in \mathbb{Z}_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i) p(y) dy. \quad (8)$$

B. Optimal quantizer and optimal clustering using Kullback-Leibler divergence

It is well-known that finding an optimal quantizer that minimizes a concave impurity function can be solved using an iterative clustering algorithm with a suitable distance from a data point to its centroid [20]. In a special case where the impurity function is the entropy, minimizing entropy impurity is equivalent to maximizing mutual information [3], [4]. Consequently, Zhang and Kurkoski showed that finding an optimal quantizer Q^* that maximizes the mutual information between the input and the output is equivalent to determining the optimal clustering that minimizes the distortion using KL divergence as the distance [4]. The result in [4] was constructed for discrete domain but it can be extended to continuous domain. For ease of analysis, we will provide a proof sketch. For a given y and a given quantizer Q that maps y to $\mathbb{Z}_i$ with centroid $\mathbf{c}_i$, the KL-divergence between the posterior distribution $\mathbf{p}(\mathbf{x}|y)$ and $\mathbf{c}_i$ is denoted by $D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)$. If the expectation is taken over Y, from Lemma 1 in [4], we have:

$$\mathbb{E}_Y[D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)] = I(X;Y) - I(X;Z).$$

Since $\mathbf{p}(\mathbf{x})$ and $\phi_i(y)$ are given, $I(X;Y)$ is given and independent of the quantizer Q. Thus, maximizing $I(X;Z)$ over Q is equivalent to minimizing $\mathbb{E}_Y[D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)]$ with an optimal quantizer being a solution to:

$$Q^* = \min_Q \mathbb{E}_Y[D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)] \quad (9)$$

$$= \min_Q \sum_{i=1}^{N} \int_{y \in \mathbb{Z}_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i) p(y) dy, \quad (10)$$

where $p(y)$ is the density of $y \in Y$. Now, the problem of finding the optimal quantizer maximizing mutual information can be cast as the problem of finding the optimal clustering that minimizes the KL divergence. Thus, in the rest of this paper, we will focus on finding the optimal clustering minimizing the KL divergence. Also, KL divergence is a special case of Bregman divergence, and for a given quantized-output set $\mathbb{Z}_i$, its centroid $\mathbf{c}_i$ can be computed by a closed-form expression (Proposition 1, [21]).

V. Structure of optimal quantizer

We show that an optimal quantizer can be found within a class of convex quantizers as defined in Definition 1. Our approach is to show that any quantizer can be replaced by an equal or better convex quantizer that maximizes the objective function $\beta I(X;Z) - C(Z)$. Specifically, we show that for any quantizer Q, there exists a convex quantizer $\tilde{Q}$ such that: (1) $\tilde{Q}$ produces the same output distribution as Q and (2) the total distortion induced by $D(\tilde{Q})$ is less than or equal to $D(Q)$, or equivalently $I(X;Z)$ produced by $\tilde{Q}$ is at least as large as that produced by Q. Thus, the optimal quantizer that maximizes $\beta I(X;Z) - C(Z)$ must belong to the class of convex quantizers. Consequently, an algorithm for finding the best quantizer in the set of all convex quantizers will find an optimal quantizer. The main point for doing this is that it is easier from an algorithmic viewpoint to search for an optimal quantizer in a set of convex quantizers than to search through all the possible quantizers. We now consider a simple case of binary quantization.

A. Structure of an optimal quantizer for binary output ($N = 2$)

Theorem 1. Let Q be an arbitrary quantizer that induces two disjoint discrete output sets $\mathbb{Z}_1$ and $\mathbb{Z}_2$ with two corresponding centroids $\mathbf{c}_1 = [c_1^1, c_1^2, \ldots, c_1^K], \mathbf{c}_2 = [c_2^1, c_2^2, \ldots, c_2^K]$. There exists a convex quantizer $\tilde{Q}$ associated with a hyper-plane that separates the space of the posterior distribution $\mathbf{p}(\mathbf{x}|y)$ into two discrete sets $\{\tilde{\mathbb{Z}}_1, \tilde{\mathbb{Z}}_2\}$ having the corresponding centroids $\{\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2\}$ such that (1) $p(\mathbb{Z}_i) \triangleq P(y \in \mathbb{Z}_i) = p(\tilde{\mathbb{Z}}_i) \triangleq P(y \in \tilde{\mathbb{Z}}_i), i = 1, 2$, and (2) $D(\tilde{Q}) \leq D(Q)$.

Proof. Let Q be a given arbitrary quantizer. Q induces $\mathbb{Z}_1$, $\mathbb{Z}_2$, $\mathbf{c}_1$ and $\mathbf{c}_2$. Let $F(\mathbf{p}(\mathbf{x}|y)) = D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)$, then:

$$\begin{aligned} F(\mathbf{p}(\mathbf{x}|y)) &= D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2) \\ &= \sum_{i=1}^{K} p(x_i|y) \log \frac{p(x_i|y)}{c_1^i} - \sum_{i=1}^{K} p(x_i|y) \log \frac{p(x_i|y)}{c_2^i} \\ &= \sum_{i=1}^{K} p(x_i|y) \log \frac{c_2^i}{c_1^i} = \mathbf{a}^T \mathbf{p}(\mathbf{x}|y), \end{aligned} \quad (11)$$

where $\mathbf{a} = [a_1, a_2, \ldots, a_K]$ be a K-dimensional vector where $a_i = \log \frac{c_2^i}{c_1^i}$, $i = 1, 2, \ldots, K$.

Now, let us consider a family of hyper-planes $\mathcal{H}(h)$ in the K-dimensional space parametrized by $h \in \mathbb{R}$ in the following equation:

$$\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) = h. \tag{12}$$

For a given h, the hyper-plane $\mathcal{H}(h)$ separates the K-dimensional posterior distribution $\mathbf{p}(\mathbf{x}|y)$ into two disjoint sets corresponding to $F(\mathbf{p}(\mathbf{x}|y)) \leq h$ and $F(\mathbf{p}(\mathbf{x}|y)) > h$. Based on Definition 1, there is also a family of convex quantizers $\tilde{Q}$ for each h. Our goal is to show that there exists a hyper-plane $\mathcal{H}(\tilde{h})$ associated with a convex quantizer $\tilde{Q}$ that separates the space of posterior distribution into two disjoint sets $\{\tilde{\mathbb{Z}}_1, \tilde{\mathbb{Z}}_2\}$ such that $p(\mathbb{Z}_i) = p(\tilde{\mathbb{Z}}_i)$, and $D(\tilde{Q}) \leq D(Q)$.

Proof of claim (1). Assume that Q produces two output sets $\mathbb{Z}_1$ and $\mathbb{Z}_2$ with the probability $p(\mathbb{Z}_1)$ and $p(\mathbb{Z}_2)$, $p(\mathbb{Z}_1) + p(\mathbb{Z}_2) = 1$. Our first claim is that one can always find a convex quantizer $\tilde{Q}$ corresponding to a hyper-plane $\mathcal{H}(\tilde{h})$ that produces $\tilde{\mathbb{Z}}_1$ and $\tilde{\mathbb{Z}}_2$ such that $p(\tilde{\mathbb{Z}}_1) = p(\mathbb{Z}_1)$ and $p(\tilde{\mathbb{Z}}_2) = p(\mathbb{Z}_2)$.

Consider the following convex quantizer:

$$\tilde{Q}(y) = \begin{cases} \tilde{\mathbb{Z}}_1 & \text{if } F(\mathbf{p}(\mathbf{x}|y)) \leq h, \\ \tilde{\mathbb{Z}}_2 & \text{if } F(\mathbf{p}(\mathbf{x}|y)) > h. \end{cases} \tag{13}$$

By increasing value of h, $h \in (-\infty, +\infty)$, the set $\tilde{\mathbb{Z}}_1$ must enlarge while $\tilde{\mathbb{Z}}_2$ must reduce. Thus, by increasing/decreasing the value of h, one can always choose an appropriate value of $h = \tilde{h}$ such that $p(\tilde{\mathbb{Z}}_1) = p(\mathbb{Z}_1)$ and $p(\tilde{\mathbb{Z}}_2) = p(\mathbb{Z}_2)$. $\tilde{h}$ corresponds to the hyper-plane $\mathcal{H}(\tilde{h})$ of the convex quantizer $\tilde{Q}$.

Proof of claim (2). Our second claim is that $D(\tilde{Q}) \leq D(Q)$. Indeed, using the hyper-plane $\mathcal{H}(\tilde{h})$ in the proof of claim (1) which produces two discrete output sets $\tilde{\mathbb{Z}}_1$ and $\tilde{\mathbb{Z}}_2$. Let $\mathbb{A} = \tilde{\mathbb{Z}}_1 \cap \mathbb{Z}_2$ and $\mathbb{B} = \tilde{\mathbb{Z}}_2 \cap \mathbb{Z}_1$. Note that if $\mathbb{A}$ or $\mathbb{B}$ is empty set then Q can be readily shown to be a convex quantizer. Let $p(\mathbb{A}) = P(y \in \mathbb{A})$ and $p(\mathbb{B}) = P(y \in \mathbb{B})$. We first show that $p(\mathbb{A}) = p(\mathbb{B})$ as follows.

$$\begin{aligned} p(\mathbb{Z}_1) &\overset{\tilde{\mathbb{Z}}_1 \cap \tilde{\mathbb{Z}}_2 = \emptyset}{=} p((\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1) \cup (\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_2)) \\ &= p(\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1) + p(\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_2) = p(\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1) + p(\mathbb{B}). \end{aligned} \tag{14}$$

Similarly,

$$\begin{aligned} p(\tilde{\mathbb{Z}}_1) &\overset{\mathbb{Z}_1 \cap \mathbb{Z}_2 = \emptyset}{=} p((\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1) \cup (\tilde{\mathbb{Z}}_1 \cap \mathbb{Z}_2)) \\ &= p(\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1) + p(\tilde{\mathbb{Z}}_1 \cap \mathbb{Z}_2) = p(\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1) + p(\mathbb{A}). \end{aligned} \tag{15}$$

Since $p(\mathbb{Z}_1) = p(\tilde{\mathbb{Z}}_1)$, from (14) and (15), we have $p(\mathbb{A}) = p(\mathbb{B})$.

Next, let $p(y)$ be the density of $y \in Y$. From $F(\mathbf{p}(\mathbf{x}|y_i)) \leq \tilde{h} < F(\mathbf{p}(\mathbf{x}|y_j))$, $\forall y_i \in \tilde{\mathbb{Z}}_1$ and $\forall y_j \in \tilde{\mathbb{Z}}_2$, together with $\mathbb{A} = \tilde{\mathbb{Z}}_1 \cap \mathbb{Z}_2$ and $\mathbb{B} = \tilde{\mathbb{Z}}_2 \cap \mathbb{Z}_1$, then $F(\mathbf{p}(\mathbf{x}|y_i)) \leq \tilde{h} < F(\mathbf{p}(\mathbf{x}|y_j))$, $\forall y_i \in \mathbb{A}$ and $\forall y_j \in \mathbb{B}$, (16) is established.

By adding $\int_{y \in \{\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) p(y) dy + \int_{y \in \{\mathbb{Z}_2 \cap \tilde{\mathbb{Z}}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2) p(y) dy$ to both sides of (16), we obtain (17).

By moving $-\int_{y \in \mathbb{A}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2) p(y) dy$ to the right hand side and $-\int_{y \in \mathbb{B}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2) p(y) dy$ to the left hand side of (17), we obtain (18).

Now, since $\mathbb{Z}_1 \cap \mathbb{Z}_2 = \emptyset$, $\mathbb{A} \cap \{\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1\} = \{\tilde{\mathbb{Z}}_1 \cap \mathbb{Z}_2\} \cap \{\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1\} = \emptyset$. Thus, the integral over $\mathbb{A}$ and $\{\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1\}$ is equivalent to the integral over $\mathbb{A} \cup \{\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1\} = \tilde{\mathbb{Z}}_1$. Similarly, using $\mathbb{B} \cup \{\mathbb{Z}_2 \cap \tilde{\mathbb{Z}}_2\} = \tilde{\mathbb{Z}}_2$, $\mathbb{B} \cup \{\mathbb{Z}_1 \cap \tilde{\mathbb{Z}}_1\} = \mathbb{Z}_1$ and $\mathbb{A} \cup \{\mathbb{Z}_2 \cap \tilde{\mathbb{Z}}_2\} = \mathbb{Z}_2$, (19) is obtained from (18).

Let $\tilde{\mathbf{c}}_1$ and $\tilde{\mathbf{c}}_2$ be the new centroids of $\tilde{\mathbb{Z}}_1$ and $\tilde{\mathbb{Z}}_2$. From Definition 3, (20) follows.

Finally, from (19) and (20), (21) is established. Combining (21) and Definition 4, $D(\tilde{Q}) \leq D(Q)$. Therefore, for any arbitrary quantizer Q, there exists a convex quantizer $\tilde{Q}$ that produces the same output distribution together with a distortion is equal or smaller than that of Q. □

B. Structure of an optimal quantizer for $N > 2$ quantization levels

Theorem 2. Let Q be an arbitrary quantizer having discrete output sets $\{\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_N\}$ with N centroids $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N$, there exists a convex quantizer $\tilde{Q}$ with N output sets $\{\tilde{\mathbb{Z}}_1, \tilde{\mathbb{Z}}_2, \ldots, \tilde{\mathbb{Z}}_N\}$ such that $\tilde{\mathbb{Z}}_i$ and $\tilde{\mathbb{Z}}_j$ are separated by a hyper-plane $\mathcal{H}(h_{ij})$ in the space of posterior distribution $\forall i, j$, $p(\mathbb{Z}_i) = p(\tilde{\mathbb{Z}}_i)$ $\forall i$, and $D(\tilde{Q}) \leq D(Q)$.

Proof. Let Q be an arbitrary quantizer that produces N output sets $\{\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_N\}$. Consider any two output sets $\mathbb{Z}_i$ and $\mathbb{Z}_j$, $i \neq j$. Now, let $\mathbb{Y}_{ij} = \mathbb{Z}_i \cup \mathbb{Z}_j$. Based on Theorem 1, there is a convex quantizer $\tilde{Q}$ corresponding to a hyper-plane $\mathcal{H}(h_{ij})$ separates the K-dimensional points $\mathbf{p}(\mathbf{x}|y), \forall y \in \mathbb{Y}_{ij}$ into two sets $\tilde{\mathbb{Z}}_i$, and $\tilde{\mathbb{Z}}_j$ with $p(\mathbb{Z}_i) = p(\tilde{\mathbb{Z}}_i)$, $p(\mathbb{Z}_j) = p(\tilde{\mathbb{Z}}_j)$ and $D(\tilde{Q}) \leq D(Q)$. Specifically, we have:

$$\tilde{Q}(y) = \begin{cases} \tilde{\mathbb{Z}}_i & \text{if } y \in \mathbb{Y}_{ij} \text{ and } F(\mathbf{p}(\mathbf{x}|y)) \leq h_{ij}, \\ \tilde{\mathbb{Z}}_j & \text{if } y \in \mathbb{Y}_{ij} \text{ and } F(\mathbf{p}(\mathbf{x}|y)) > h_{ij}, \end{cases} \tag{22}$$

where h_{ij} is a real number corresponding to the hyper-plane $\mathcal{H}(h_{ij})$.

Since the distortion is additive, and the result holds for arbitrary $\mathbb{Z}_i$ and $\mathbb{Z}_j$, by repeating the above process for at most $\frac{N(N-1)}{2}$ pairs of $\mathbb{Z}_i$ and $\mathbb{Z}_j$, one can construct a convex quantizer $\tilde{Q}$ which produces $\{\tilde{\mathbb{Z}}_1, \tilde{\mathbb{Z}}_2, \ldots, \tilde{\mathbb{Z}}_N\}$ such that $p(\mathbb{Z}_i) = p(\tilde{\mathbb{Z}}_i)$ $\forall i$, and $D(\tilde{Q}) \leq D(Q)$. □

Remark 1. (*Optimality*) For a given quantizer Q, there exists a convex quantizer $\tilde{Q}$ having the same output probability $\mathbf{p}(\mathbf{z})$ with a lower distortion. This leads to the same cost function $C(Z)$ for both Q and $\tilde{Q}$. Since the distortion $D(\tilde{Q})$ is smaller or at most equal than that of $D(Q)$, $I(X;Z)$ induced by $\tilde{Q}$ is at least as large as that produced by Q. Thus, we can conclude that an optimal quantizer that maximizes the objective function $\beta I(X;Z) - C(Z)$ must belong to the set of convex quantizers.

Remark 2. (*Complexity*) Since the set of convex quantizers is a subset of all the possible quantizers, searching over the set of convex quantizers is faster than searching over all the possible quantizers. Specifically, if the continuous variable $y \in \mathbb{R}$ is

$$\begin{aligned}\int_{y\in\mathbb{A}} F(\mathbf{p}(\mathbf{x}|y))p(y)dy &= \int_{y\in\mathbb{A}} [D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)]p(y)dy \\ &\leq \int_{y\in\mathbb{B}} [D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)]p(y)dy = \int_{y\in\mathbb{B}} F(\mathbf{p}(\mathbf{x}|y))p(y)dy \end{aligned} \tag{16}$$

$$\begin{aligned}&\int_{y\in\mathbb{A}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy - \int_{y\in\mathbb{A}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy + \int_{y\in\{\mathbb{Z}_1\cap\tilde{\mathbb{Z}}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\{\mathbb{Z}_2\cap\tilde{\mathbb{Z}}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy \\ \leq\ &\int_{y\in\mathbb{B}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy - \int_{y\in\mathbb{B}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy + \int_{y\in\{\mathbb{Z}_1\cap\tilde{\mathbb{Z}}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\{\mathbb{Z}_2\cap\tilde{\mathbb{Z}}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy \end{aligned} \tag{17}$$

$$\begin{aligned}&\Big(\int_{y\in\mathbb{A}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\{\mathbb{Z}_1\cap\tilde{\mathbb{Z}}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy\Big) + \Big(\int_{y\in\{\mathbb{Z}_2\cap\tilde{\mathbb{Z}}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy + \int_{y\in\mathbb{B}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy\Big) \\ \leq\ &\Big(\int_{y\in\mathbb{B}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\{\mathbb{Z}_1\cap\tilde{\mathbb{Z}}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy\Big) + \Big(\int_{y\in\{\mathbb{Z}_2\cap\tilde{\mathbb{Z}}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy + \int_{y\in\mathbb{A}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy\Big) \end{aligned} \tag{18}$$

$$\int_{y\in\tilde{\mathbb{Z}}_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\tilde{\mathbb{Z}}_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy \leq \int_{y\in\mathbb{Z}_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\mathbb{Z}_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy \tag{19}$$

$$\int_{y\in\tilde{\mathbb{Z}}_1} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_1)p(y)dy + \int_{y\in\tilde{\mathbb{Z}}_2} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_2)p(y)dy \leq \int_{y\in\tilde{\mathbb{Z}}_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\tilde{\mathbb{Z}}_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy \tag{20}$$

$$\int_{y\in\tilde{\mathbb{Z}}_1} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_1)p(y)dy + \int_{y\in\tilde{\mathbb{Z}}_2} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_2)p(y)dy \leq \int_{y\in\tilde{\mathbb{Z}}_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)p(y)dy + \int_{y\in\tilde{\mathbb{Z}}_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)p(y)dy \tag{21}$$

discretized into M discrete data points, an exhaustive search over all the possible partitions of M points into N disjoint subsets will have an exponential time complexity of $O(N^M)$. On the other hand, the time complexity of an exhaustive search over all the possible hyper-planes (or over all the possible convex quantizers) is only $O(M^{K-1})$ [13]. Typically, $M >> K$, thus searching over the set of convex quantizers is much faster.

Remark 3. (*Tractable case: binary inputs*) For a special setting of binary input channel $K = 2$, a hyper-plane in the space of the posterior distribution is a scalar and the dynamic programming algorithm is capable to determine an optimal quantizer in $O(M^3)$. We refer the reader to the work in [19] for the details.

Remark 4. (*Locally optimal solution*) While this paper aims to determine a globally optimal quantizer for a general scenario, its time complexity is still high $O(M^{K-1})$. However, it is possible to derive an optimality condition for a locally optimal quantizer which is similar to the result in [17], [12]. Indeed, using a similar approach in [17], [12], it is possible to show that a locally optimal quantizer Q^* must satisfy:

$$Q(y) \to \mathbb{Z}_i \longleftrightarrow d(y, \mathbb{Z}_i) \leq d(y, \mathbb{Z}_j), \forall j \neq i,$$

where the "distance" $d(y, \mathbb{Z}_i)$ from y to $\mathbb{Z}_i$ is defined by:

$$d(y, \mathbb{Z}_i) = \beta D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i) + \frac{\mathrm{d}C(Z)}{\mathrm{d}p(z_i)}.$$

Based on this optimality condition, an iterative algorithm which is similar to k-means algorithm can be used to find a locally optimal solution in linear time complexity [17], [12].

VI. Bounds on the number of thresholds for an optimal quantizer

We note that a convex quantizer Q quantizes a point y based on which convex regions (separated by a set of hyper-planes) the corresponding posterior distribution $\mathbf{p}(\mathbf{x}|y)$ lies in. This requires mapping a point y to its posterior distribution, then successively narrowing down which regions it lies in using the hyper-plane equations. Often times, it is desirable to determine a set of thresholds $t_i \in \mathbb{R}, i = 1, 2, \ldots, S$ that separates y into multiple disjoint regions $\mathbb{R}_i \in \mathbb{R}$ directly. That said, two high-dimensional points $\mathbf{p}(\mathbf{x}|y_1)$ and $\mathbf{p}(\mathbf{x}|y_2)$ that belong to the same convex region in the posterior distribution space may map to multiple disjoint regions $\mathbb{R}_i$'s. Using t_i's, one is able to quantize y directly based on its value. In this section, we determine an upper bound on the number of thresholds t_i that separates the regions $\mathbb{R}_i$'s associated with an optimal quantizer.

As an example, if the output is binary, i.e., $\mathbb{Z} = \{z_1, z_2\}$, then $t_1, t_2, \ldots, t_S$ divide $\mathbb{R}$ into $S + 1$ contiguous disjoint segments $\mathbb{R}_i = (t_{i-1}, t_i)$, with $t_0 = -\infty$ and $t_{S+1} = \infty$. Each y in $\mathbb{R}_i$ is mapped to either z_1 or z_2 alternatively. For a given number of thresholds and the search step size (grid resolution), one can exhaustively search over all the possible $t_1, t_2, \ldots, t_S$ to determine an optimal quantizer. In [6], Kurkoski and Yagi gave a condition for which an optimal quantizer requires only a single threshold to maximize the mutual information between the input and the output of binary-input binary-output channels. Thus, an exhaustive search is practical. In [19], the author extended the single threshold condition in [6] for binary channels under the quantized-output constraint. However, for K-ary input channels, $K > 2$, finding the minimum number of thresholds that is possible to achieve the maximum of mutual

information between the input and the output is still an open problem. In this section, we utilize the results in Theorem 1 and Theorem 2 to construct an upper bound on the required number of thresholds t_i's for an optimal quantizer.

Theorem 1 and Theorem 2 state that the optimal output sets are separated by hyper-planes in the posterior distribution space which correspond to a number of thresholds t_i's in $y \in \mathbb{R}$. In particular, if a hyper-plane is specified by an equation, then the corresponding number of thresholds t_i's associated with the two sets separated by this hyper-plane is at most equal to the number of distinct real solutions of this equation. Thus, an upper bound on the number of thresholds can be obtained by determining an upper bound on the number of solutions of the set of equations specified the hyper-planes of an optimal quantizer. Theorem 3 formally states this result.

Theorem 3. Let $\mathbb{R}_l \cup \mathbb{R}_r = \mathbb{R}$ and $\mathbb{R}_l \cap \mathbb{R}_r = \emptyset$. If $\forall y_l \in \mathbb{R}_l$ and $\forall y_r \in \mathbb{R}_r$,

$$\mathbf{a}^T \mathbf{p}(\mathbf{x}|y_r) \geq h, \ \mathbf{a}^T \mathbf{p}(\mathbf{x}|y_l) < h \tag{23}$$

for given $h > 0$ and $\mathbf{a}$, then $\mathbb{R}_l$ and $\mathbb{R}_r$ are separated by at most S thresholds $t_1, t_2, \ldots, t_S \in \mathbb{R}$ where S is the number of real distinct solutions y to the equation:

$$\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) = h. \tag{24}$$

Proof. Since $\phi_i(y)$ is assumed to be continuous, positive and differentiable everywhere and $h \in \mathbb{R}$, $s(y) = \mathbf{a}^T \mathbf{p}(\mathbf{x}|y) - h$ is a continuous function. Furthermore, if $s(y)$ has S real distinct solutions, then we need exactly S thresholds to separate $\mathbb{R}$ into $S+1$ contiguous disjoint segments, each alternatively maps to either $\mathbb{R}_l$ if $\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) < h$ or $\mathbb{R}_r$ if $\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) \geq h$. □

Theorem 3 provides a concrete approach to determine the number of required thresholds by finding the number of solutions of a hyper-plane equation. Next, using the result in Theorem 3, we construct an upper bound on the number of thresholds for additive white Gaussian noise (AWGN) channels.

Theorem 4. For an additive white Gaussian noise (AWGN) channel, the input symbols satisfy $x_{i+1} - x_i = \delta, i = 1, 2, \ldots N-1$, where δ is a constant, and K quantization levels, the optimal quantizer requires no more than $\dfrac{N(N-1)(K-1)}{2}$ thresholds.

Proof. Using (3), (24) can be rewritten by:

$$a_1 \frac{p_1\phi_1(y)}{\sum_{i=1}^K p_i\phi_i(y)} + a_2 \frac{p_2\phi_2(y)}{\sum_{i=1}^K p_i\phi_i(y)} + \cdots + a_K \frac{p_K\phi_K(y)}{\sum_{i=1}^K p_i\phi_i(y)} = h, \tag{25}$$

or,

$$\sum_{i=1}^K (a_i - h) p_i \phi_i(y) = 0, \tag{26}$$

where

$$\phi_i(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-x_i}{\sigma})^2}. \tag{27}$$

Since $x_{i+1} - x_i = \delta$, $x_i - x_1 = (i-1)\delta$. Substituting (27) into (26) and using $x_i - x_1 = (i-1)\delta$, we have:

$$\begin{aligned}
&\sum_{i=1}^K (a_i - h) p_i \phi_i(y) \\
=& \sum_{i=1}^K (a_i - h) p_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-x_i}{\sigma})^2} \\
=& \sum_{i=1}^K (a_i - h) p_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y^2 - 2yx_i + x_i^2 - 2yx_1 + 2yx_1}{\sigma^2})} \\
=& \sum_{i=1}^K (a_i - h) p_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y^2 - 2yx_1 + x_i^2 - 2y(x_i - x_1)}{\sigma^2})} \\
=& \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2 - 2yx_1}{2\sigma^2}} \Big(\sum_{i=1}^K (a_i - h) p_i e^{-\frac{x_i^2}{2\sigma^2}} e^{\frac{y(x_i - x_1)}{\sigma^2}} \Big) \\
=& \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2 - 2yx_1}{2\sigma^2}} \Big(\sum_{i=1}^K (a_i - h) p_i e^{-\frac{x_i^2}{2\sigma^2}} e^{\frac{y(i-1)\delta}{\sigma^2}} \Big).
\end{aligned} \tag{28}$$

Let $e^{\frac{y}{\sigma^2}} = w$, $\sum_{i=1}^K (a_i - h) p_i e^{-\frac{x_i^2}{2\sigma^2}} = b_i$, and since $\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2 - 2yx_1}{2\sigma^2}} \neq 0$, from (26) and (28), we have:

$$\sum_{i=1}^K b_i (w^\delta)^{i-1} = 0. \tag{29}$$

This follows that w^δ must be roots of a polynomial function having a degree at most $K-1$ which can have at most $K-1$ solutions. Since w^δ and $e^{\frac{y}{\sigma^2}}$ are both monotonic functions, (29) has at most $K-1$ solutions in y which results in at most $K-1$ thresholds in $\mathbb{R}$.

Next, since N partitioned-outputs require at most $N(N-1)/2$ hyper-plane cuts, a quantizer with $N(N-1)(K-1)/2$ thresholds is sufficient to maximize the mutual information. □

Remark 5. AWGN is one of the most common channels in telecommunication, and the assumption of $x_{i+1} - x_i = \delta$ is not too restricted. Indeed, many amplitude modulation techniques such as Amplitude Shift Keying (ASK), On-Off Keying (OOK), and Pulse Amplitude Modulation (PAM) satisfy the condition in Theorem 4.

Remark 6. As a consequence of Theorem 4, if the channel is an AWGN binary-input channel, i.e. $N = 2$, then an optimal quantizer requires at most $K-1$ thresholds. This agrees with the results in [3], [19]. Furthermore, if the channel is AWGN binary-input binary-output ($N = K = 2$), then a single threshold quantizer is optimal.

Remark 7. Based on the proposed upper bound on the number of thresholds, an exhaustive search algorithm can be used for finding the globally optimal quantizer of AWGN channels for small K and N. For example, if $N = 2$, an optimal quantizer

requires at most $K-1$ thresholds which divides $\mathbb{R}$ into K contiguous disjoint segments, each maps to either z_1 or z_2 alternatively. Since $h \in \mathbb{R}$, let ϵ denote the search resolution i.e., the distance between two consecutive points on the search grid. For a given ϵ, an exhaustive search algorithm would have time complexity of $O(M^{K-1})$ where $M = \frac{L}{\epsilon}$, and L is the search range.

Remark 8. Note that our proposed method can be used to determine the number of thresholds for other additive noise channels such as additive exponential distribution, additive uniform distribution, and additive gamma distribution.

VII. Numerical results

First, we want to refer the reader to the numerical results in [19] which can be considered as special cases for illustrating our Theorem 1 and Theorem 2. In this section, we only focus on providing some examples to verify the theoretical results in our proposed Theorem 4.

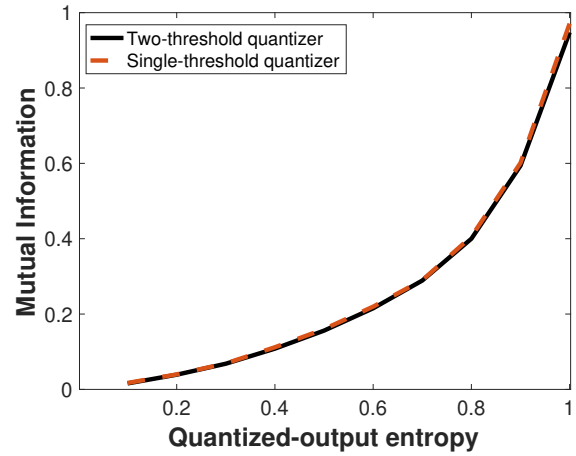
Example 1. We consider a binary-input channel having $\mathbb{X} = \{x_1 = -10, x_2 = 10\}$ and $\mathbf{p}(\mathbf{x}) = [0.6, 0.4]$. X is corrupted by an additive white Gaussian noise having probability density function $G(\mu = 0, \sigma = 2)$ with $\phi_1(y) = G(-10, 2)$ and $\phi_2(y) = G(10, 2)$. Next, we want to design an optimal quantizer Q that quantizes $y \in \mathbb{R}$ to a binary output $\mathbb{Z} = \{z_1, z_2\}$ such that the mutual information $I(X;Z)$ is maximized while $H(Z) \leq \gamma$ for a given γ.

Since $N = K = 2$, Theorem 4 points out that a single-threshold quantizer is optimal. To confirm this theoretical result, we exhaustively search over all the possible single-threshold and two-threshold quantizers in the interval $[-15, 15]$ with the resolution $\epsilon = 0.1$. The maximum values of $I(X;Z)$ using single-threshold quantizers and two-threshold quantizers are denoted by the red-dash curve and the black-curve in Fig. 1, respectively. As seen, the maximum values of mutual information using single-threshold quantizers are slightly larger than the optimal values of mutual information provided by two-threshold quantizers, for $\gamma = 0.1, 0.2, \ldots, 0.9, 1$.

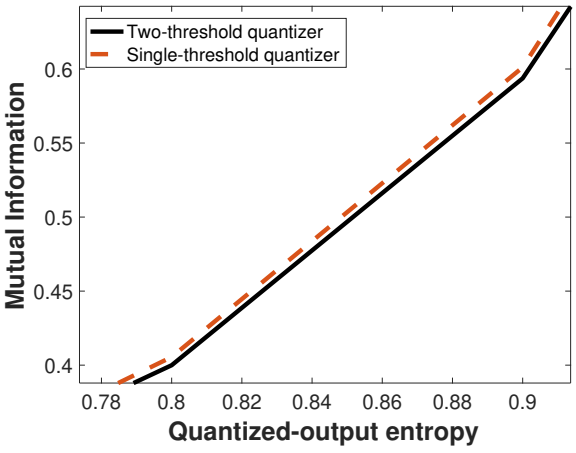
This numerical result indicates that if the channel is AWGN binary-input binary-output ($N = K = 2$), then an optimal quantizer can have a single threshold. Thus, our example confirms the result in Theorem 4.

Example 2. We consider a channel having input $\mathbb{X} = \{x_1 = -10, x_2 = 0, x_3 = 10\}$ and $\mathbf{p}(\mathbf{x}) = [0.3, 0.4, 0.3]$. Similar to Example 1, X is corrupted by an additive white Gaussian noise having probability density function $G(\mu = 0, \sigma = 1)$ with $\phi_1(y) = G(-10, 1)$, $\phi_2(y) = G(0, 1)$, and $\phi_3(y) = G(10, 1)$. We want to design an optimal quantizer Q that quantizes $y \in \mathbb{R}$ to a binary quantized-output $\mathbb{Z} = \{z_1, z_2\}$ to maximize $I(X;Z)$ while $H(Z) \leq \gamma$, for a given γ.

Based on Theorem 4, using $K = 3$ and $N = 2$, the optimal quantizer requires at most 2 thresholds. To verify the upper bound on the number of thresholds, we exhaustively search over all the possible single-threshold, two-threshold and three-threshold quantizers, respectively. Due to a high time-complexity of performing an exhaustive search algorithm with three thresholds, we limit the searching range in $[-15, 15]$ with the resolution $\epsilon = 0.2$. The maximum values of $I(X;Z)$ for single-threshold quantizers, two-threshold quantizers, and three-threshold quantizers are denoted by the black curve, the black-dash curve, and the green curve in Fig. 2, respectively. As seen, $I(X;Z)$ provided by two-threshold quantizers are slightly larger than that of three-threshold quantizers. On the other hand, $I(X;Z)$ provided by single-threshold quantizers are always less than that produced by both two-threshold and three-threshold quantizers. This numerical result implies that two-threshold quantizers are optimal in this example which confirms the result in Theorem 4.



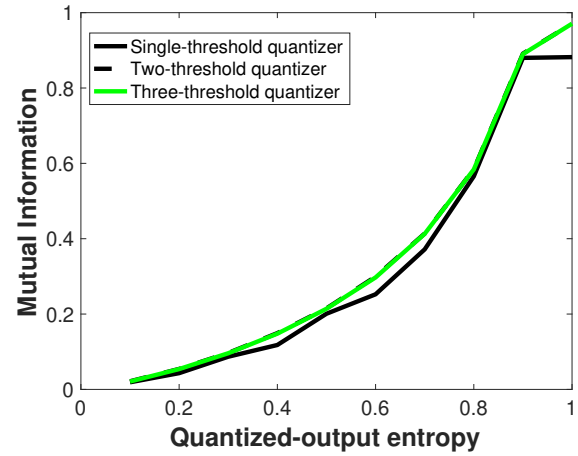
(a)



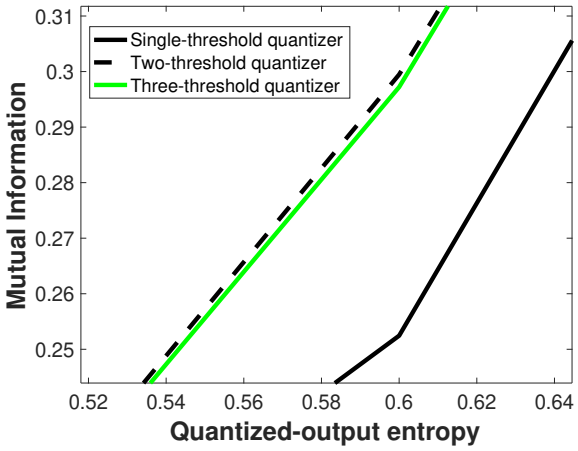
(b)

Figure 1: Example 1: (a) maximum values of mutual information using single-threshold quantizers *vs.* two-threshold quantizers under output constraint $H(Z) \leq \gamma$ for various values of $\gamma = 0.1, 0.2, \ldots, 0.9, 1$; (b) zoom in for $\gamma \in [0.7, 1]$.

We note that that finding an optimal quantization for a large number of inputs is an extremely hard (NP-complete) problem [10]. While our result on bounding the maximum number of thresholds for an optimal quantization can be used to reduce the computations of an exhaustive search on small problems, it is not feasible for arbitrarily large problems. Our future work will focus on alleviating this shortcoming.



(*a*)



(*b*)

Figure 2: Example 2: (a) Maximum values of mutual information using single-threshold quantizers, two-threshold quantizers and three-threshold quantizers under output constraint $H(Z) \leq \gamma$ for various values of $\gamma = 0.1, 0.2, \ldots, 0.9, 1$; (b) zoom in for $\gamma \in [0.5, 0.7]$.

VIII. Conclusion

In this paper, we investigate the structure of optimal quantizers that maximize the mutual information between the input and the output under an arbitrary constraint on the output distribution. Our result shows that the optimal quantizer must belong to a class of convex quantizers. Furthermore, we describe an upper bound on the number of thresholds for an optimal quantizer. Based on this upper bound, an exhaustive search algorithm with polynomial time complexity can be used to determine an optimal solution.

References

[1] I. Tal and A. Vardy. How to construct polar codes. *IEEE Transactions on Information Theory*, 59(10):6562–6582, 2013.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Thuan Nguyen and Thinh Nguyen. On thresholding quantizer design for mutual information maximization: Optimal structures and algorithms. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5. IEEE, 2020.

[9] Thuan Nguyen, Yu-Jung Chu, and Thinh Nguyen. A new fast algorithm for finding capacity of discrete memoryless thresholding channels. In *2020 International Conference on Computing, Networking and Communications (ICNC)*, pages 56–60. IEEE, 2020.

[10] Brendan Mumey and Tomáš Gedeon. Optimal mutual information quantization is np-complete. In *Neural Information Coding (NIC) workshop poster, Snowbird UT*, pages 1932–4553, 2003.

[11] Thuan Duc Nguyen and Thinh Nguyen. On binary quantizer for maximizing mutual information. *IEEE Transactions on Communications*, 68(9):5435–5445, 2020.

[12] DJ Strouse and David J Schwab. The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630, 2017.

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

[14] Thuan Nguyen and Thinh Nguyen. A linear time partitioning algorithm for frequency weighted impurity functions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5375–5379. IEEE, 2020.

[15] A. Gyorgy and Tamás Linder. On the structure of entropy-constrained scalar quantizers. *Proceedings. 2001 IEEE International Symposium on Information Theory (IEEE Cat. No.01CH37252)*, pages 29–, 2001.

[16] Andras Gyorgy and Tamás Linder. Codecell convexity in optimal entropy-constrained vector quantization. *IEEE Transactions on Information Theory*, 49(7):1821–1828, 2003.

[17] Philip A. Chou, Tom D. Lookabaugh, and Robert M. Gray. Entropy-constrained vector quantization. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 37:31–42, 1989.

[18] Allen Gersho and Robert M. Gray. Vector quantization and signal compression. In *The Kluwer international series in engineering and computer science*, 1991.

[19] T. Nguyen and T. Nguyen. Structure of optimal quantizer for binary-input continuous-output channels with output constraints. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1450–1455, 2020.

[20] Philip A. Chou. Optimal partitioning for classification and regression trees. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):340–354, 1991.

[21] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Thuan Nguyen received a B.S. degree in Electrical Engineering (honors program) from Post and Telecommunication Institute of Technology, Vietnam, in 2013 and the M.S. and Ph.D. degrees in electrical computer engineering from Oregon State University in 2018 and 2021, respectively. His research interests include information theory, signal processing and machine learning. He is currently a Postdoc researcher at Tufts University, Boston, MA, USA.

Thinh Nguyen 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.