Structure of Optimal Quantizer for Binary-Input Continuous-Output Channels with Output Constraints

Thuan Nguyen
School of Electrical and
Computer Engineering
Oregon State University
Corvallis, OR, 97331
Email: nguyeth9@oregonstate.edu

Thinh Nguyen
School of Electrical and
Computer Engineering
Oregon State University
Corvallis, 97331
Email: thinhq@eecs.oregonstate.edu

***Abstract*—In this paper, we consider a channel whose the input is a binary random source $X \in \{x_1, x_2\}$ with the probability mass function (pmf) $\mathbf{p}_X = [p_{x_1}, p_{x_2}]$ and the output is a continuous random variable $Y \in \mathbb{R}$ as a result of a continuous noise, characterized by the channel conditional densities $p_{y|x_1} = \phi_1(y)$ and $p_{y|x_2} = \phi_2(y)$. A quantizer Q is used to map Y back to a discrete set $Z \in \{z_1, z_2, \ldots, z_N\}$. To retain most amount of information about X, an optimal Q is one that maximizes $I(X;Z)$. On the other hand, our goal is not only to recover X but also ensure that $\mathbf{p}_Z = [p_{z_1}, p_{z_2}, \ldots, p_{z_N}]$ satisfies a certain constraint. In particular, we are interested in designing a quantizer that maximizes $\beta I(X;Z) - C(\mathbf{p}_Z)$ where β is a trade-off parameter and $C(\mathbf{p}_Z)$ is an arbitrary cost function of $\mathbf{p}_Z$. Let the posterior probability $p_{x_1|y} = r_y = \dfrac{p_{x_1}\phi_1(y)}{p_{x_1}\phi_1(y) + p_{x_2}\phi_2(y)}$, our result shows that the structure of the optimal quantizer separates r_y into convex cells. In other words, the optimal quantizer has the form: $Q^*(r_y) = z_i$, if $a^*_{i-1} \le r_y < a^*_i$, for some optimal thresholds $a^*_0 = 0 < a^*_1 < a^*_2 < \cdots < a^*_{N-1} < a^*_N = 1$. Based on this optimal structure, we describe some fast algorithms for determining the optimal quantizers.**

Keyword: quantization, mutual information, constraints.

I. Introduction

Motivated by the use of quantizers in the decoders for polar codes [1] and LDPC codes [2], designing the quantizers that maximize the mutual information between input and quantized output has received much attention in recent years. Over the past decade, many algorithms and theoretical results on designing such quantizers have been proposed [3]–[13]. Finding an optimal quantizer for an arbitrary number of discrete inputs and outputs remains a difficult problem [14]. Existing practical algorithms typically find a locally optimal solution or an approximate globally optimal solution [4], [6], [7], [9], [12], [15], [16]. Under some certain restrictions, there are efficient algorithms to find the globally optimal quantizer that maximize the mutual information between the input and the quantized output [3], [10], [11]. One such important case is when the channel input is binary, then the optimal quantizer has the structure of convex cells in the space of posterior distribution. In particular, let $X \in \{x_1, x_2\}$ be a binary random input with the probability mass function (pmf) $\mathbf{p}_X = [p_{x_1}, p_{x_2}]$ to a given channel, Y be the continuous output due to a continuous noise source where Y is characterized by the channel conditional densities $p_{y|x_1} = \phi_1(y)$ and $p_{y|x_2} = \phi_2(y)$. A quantizer Q is used to map Y back to a discrete set $Z \in \{z_1, z_2, \ldots, z_N\}$. To retain the most amount of information about X, an optimal Q is one that maximizes $I(X;Z)$. Let the posterior probability $p_{x_1|y} = r_y = \dfrac{p_{x_1}\phi_1(y)}{p_{x_1}\phi_1(y) + p_{x_2}\phi_2(y)}$, then it has been shown that the structure of the optimal quantizer separates r_y into convex cells [3]. In other words, the optimal quantizer is of the form:

$$Q^*(r_y) = z_i, \text{ if } a^*_{i-1} \le r_y < a^*_i, \tag{1}$$

for some optimal thresholds $a^*_0 = 0 < a^*_1 < \cdots < a^*_{N-1} < a^*_N = 1$. These quantizers are called *convex cell quantizer* in r_y. Using this convex cell structure, an optimal quantizer can be found efficiently in a polynomial time via dynamic programming technique. In [5] and [13], the time complexity can be further reduced to a linear time complexity using the famous SMAWK algorithm [17]. We note that it is also well known that if y is used instead of r_y, an optimal quantizer $Q^*(y)$ might not separate y into the N convex regions, i.e., multiple disjoint regions of y might map to the same z_i. Consequently, it is more difficult to find an optimal quantizer from the algorithmic viewpoint.

While there are many results on finding the optimal quantizer that maximizes $I(X;Z)$, the problem of finding an optimal quantizer that maximizes $I(X;Z)$ while the output Z must satisfy a certain constraint, receives less attention. However, we note that finding the optimal quantizer for the objectives other than the mutual information under some constraints on the output, has a long history. For example, the problem of entropy-constrained scalar quantization [18], [19], [20] and entropy-constrained vector quantization [21], [22], [23] have been well investigated. The objectives in these problems are to minimize a specific distortion function, typically the mean square error between the input and the

output while ensuring that the output entropy $H(Z)$ is less than a certain amount. The constrained-entropy quantization is important in compression applications in limited storage systems. For example, suppose one wants to quantize a data source before applying entropy coding to gain compression. Ideally, one wants zero distortion between the original data and the quantized data for viewing. However, this may result in high entropy in the quantized data, which after compression performed on the quantized data might exceed a given storage capacity. Similarly, in a limited communication channel, it is important to reduce the entropy in the source in order to reduce the bit rate to match the channel bandwidth. Constraint on output entropy is only an example. In many scenarios, constraints such as power consumption and time delay can be modeled as constraints on the outputs.

The most related works of this paper are that of Strouse et al. [24] and Gyorgy and Linder [20]. In [24], Strouse et al. proposed an iterative algorithm to find a local optimal quantizer that maximizes the mutual information under the entropy constraint on the output. In [25], the authors generalized the results in [24] to find a local optimal quantizer that minimizes an arbitrary impurity function while output constraint can be an arbitrary concave function. However, to the best of our knowledge, there is no work that can determine a globally optimal quantizer that maximizes the mutual information between input and quantized-output for a given output constraint, even for the binary input channels. On the other hand, our work is similar to the work of Gyorgy and Linder that proves the optimality of convex cell quantizers [20].

To that end, we investigate the problem of designing a quantizer that maximizes $F(X,Z) = \beta I(X;Z) - C(\mathbf{p}_Z)$ where β is a trade-off parameter and $C(\mathbf{p}_Z)$ is an arbitrary cost function of $\mathbf{p}_Z$. Our contribution is to show that, similar to the result in [3], the structure of the optimal quantizer does not change, i.e., $Q^*(y)$ separates r_y into convex cells as defined in Eq. (1). Our approach to obtain this result is to show that for any given quantizer $Q(y)$, there exists a *convex cell quantizer* $\bar{Q}(r_y)$ such that: (1) $\bar{Q}(r_y)$ produces the same $\mathbf{p}_Z$ as that of $Q(y)$ (therefore, the same $C(\mathbf{p}_Z)$) and (2) $I(X;Z)$ produced by $\bar{Q}(r_y)$ is at least as large as that produced by $Q(y)$. Thus, the optimal quantizer must belong to the class of *convex cell quantizers*. It is worth noting that our approach is very similar to the work in [20] that characterizes the optimal structure of entropy-constrained scalar quantizers. In addition, we outline a fast algorithm for finding an optimal quantizer for the case $N = 2$ and discuss a sufficient condition for which a single threshold quantizer is optimal. Furthermore, under a certain mild restriction on $C(\mathbf{p}_Z)$, we describe an efficient algorithm for finding the optimal quantizers for $N > 2$.

II. Problem Formulation

Fig. 1 illustrates the problem setting. The binary random input $X = \{x_1, x_2\}$ with a given pmf $\mathbf{p}_X = [p_{x_1}, p_{x_2}] = [p_1, p_2]$ is transmitted over a noisy channel. Due to a continuous noise source, the output Y is a continuous random variable value specified by two given conditional densities $p_{y|x_1} = \phi_1(y)$ and $p_{y|x_2} = \phi_2(y)$. A quantizer Q is used to quantize the continuous output Y to a discrete output set $Z = \{z_1, z_2, \ldots, z_N\}$. Let $C(\mathbf{p}_Z) = C(p_{z_1}, p_{z_2}, \ldots, p_{z_N})$ be an arbitrary cost function of $\mathbf{p}_Z$. Our objective is to find the solution to the following optimization problem.

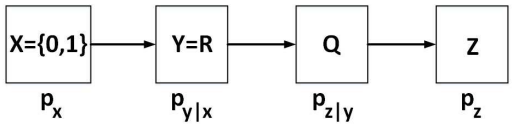

Figure 1: Problem setting: X is a binary random variable, Y is a continuous random variable, and Z is discrete random variable resulted from the quantization of Y.

$$\max_{Q} \beta I(X;Z) - C(\mathbf{p}_Z), \tag{2}$$

where β is pre-specified parameter to control the trade-off between maximizing $I(X;Z)$ and minimizing $C(\mathbf{p}_Z)$.

III. Preliminaries

A. Notations and definitions

In this paper, the capital letter denotes the set and the bold letter denotes the vector. For convenience, we use the following notations and definitions:

1) $r_y = p_{x_1|y}$ denotes the conditional probability of $X = x_1$ given $Y = y$. Let $\phi_1(y) = p_{y|x_1}$ and $\phi_2(y) = p_{y|x_2}$ then $r_y = \dfrac{p_1\phi_1(y)}{p_1\phi_1(y) + p_2\phi_2(y)}$.
2) $\mathbf{v}_y = \mathbf{p}_{x|y} = [p_{x_1|y}, p_{x_2|y}]$ denotes the conditional probability vector of X given Y, $\mathbf{v}_y = [r_y, 1 - r_y]$.
3) $\mu(y)$ denotes the density function of Y, $\mu(y) = p_1\phi_1(y) + p_2\phi_2(y)$.
4) Z_i denotes the set of y that is quantized to the i^{th} output z_i. $Z_i = \{y : Q(y) = z_i\}$.

Definition 1. (**Kullback-Leibler (KL) divergence**) KL divergence of two probability vectors $\mathbf{a} = (a_1, a_2, \ldots, a_J)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_J)$ is defined by

$$D(\mathbf{a}||\mathbf{b}) = \sum_{i=1}^{J} a_i \log(\frac{a_i}{b_i}). \tag{3}$$

Definition 2. (**Centroid**) Centroid of subset $Z_i \subset \mathbb{R}$ is a two dimensional vector $\mathbf{c}_i = [c_i, 1 - c_i]$ that globally minimizes the total KL divergence $\mathbf{v}_y$ to $\mathbf{c}_i$ from all $y \in Z_i$:

$$\mathbf{c}_i = \underset{\mathbf{c}}{\operatorname{argmin}} \int_{y \in Z_i} D(\mathbf{v}_y||\mathbf{c})\mu(y)dy. \tag{4}$$

Definition 3. (**Distortion measurement**) The total distortion of a quantizer Q with N output sets $(Z_1, Z_2, \ldots, Z_N)$ is denoted by:

$$D(Q) = \sum_{i=1}^{N} \int_{y \in Z_i} D(\mathbf{v}_y||\mathbf{c}_i)\mu(y)dy, \tag{5}$$

where $\mathbf{c}_i$ is the centroid of Z_i.

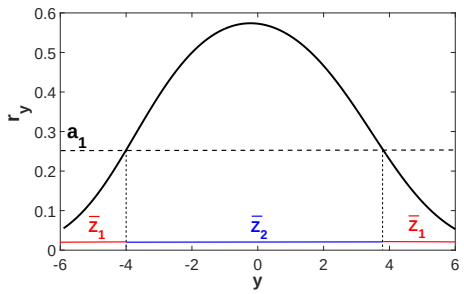

Figure 2: Illustration of a convex cell quantizer for $N = 2$ corresponding to two output sets $\bar{Z}_1$ and $\bar{Z}_2$. $\bar{Z}_1 \leq \bar{Z}_2$ and $r_{y\in\bar{Z}_1} \leq a_1 = 0.25 \leq r_{y\in\bar{Z}_2}$. $p_{x_1} = p_{x_2} = 0.5$, $\phi_1(y) = N(0,\sqrt{3})$ and $\phi_2(y) = 0.5N(-1,\sqrt{6}+0.5N(1,\sqrt{5}))$.

Definition 4. (**Vector order**) For two given conditional probability vectors $\mathbf{v}_{y_1}$ and $\mathbf{v}_{y_2}$, we define $\mathbf{v}_{y_1} \leq \mathbf{v}_{y_2}$ if and only if $p_{x_1|y_1} \leq p_{x_1|y_2}$ or $r_{y_1} \leq r_{y_2}$. Similarly, for two centroid vectors $\mathbf{c}_i = [c_i, 1-c_i]$ and $\mathbf{c}_j = [c_j, 1-c_j]$, $\mathbf{c}_i \leq \mathbf{c}_j$ if and only if $c_i \leq c_j$.

Definition 5. (**Set order**) Given two arbitrary sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, we define $A \leq B$ if and only if for all $y_a \in A$ and any $y_b \in B$, we have $\mathbf{v}_{y_a} \leq \mathbf{v}_{y_b}$. We define $A \equiv B$ if and only if $A \subset B$ and $B \subset A$.

As an illustration, Figure 2 shows r_y as a function of y for a *convex cell quantizer* $\bar{Q}(r_y)$ for $N = 2$. In this case, $\bar{Z}_1 \leq \bar{Z}_2$ since $\forall\ y_1 \in Z_1$ and $\forall\ y_2 \in Z_2$, we have $p_{x_1|y_1} = r_{y_1} \leq r_{y_2} = p_{x_1|y_2}$. Note that $\bar{Q}(r_y)$ is equivalent to the quantizer $Q(y)$ where $Q(y) = z_2$ if $y \in (-4, 3.8)$ and $Q(y) = z_1$, otherwise. As defined, $Q(y)$ is not a *convex cell quantizer* since Z_1 consists of two disjoint sets. On the other hand, using r_y as a variable in the quantizer, $\bar{Q}(r_y)$ is a *convex cell quantizer* since r_y is separated into two disjoint sets.

B. Optimal quantizer and Kullback-Leibler divergence

Interestingly, one can show that finding the optimal quantizer Q^* that maximizes the mutual information $I(X;Z)$ is equivalent to determining the optimal clustering that minimizes the distortion using KL divergence as the distance metric. The proof was given in [4] in discrete domain but it can be extended into continuous domain. For convenience, we just provide a sketch of proof using our notations. For a given y and a given quantizer that produces $z_i = Q(y)$ having the centroid $\mathbf{c}_i$, the KL-divergence between the conditional pmf $\mathbf{v}_y$ and $\mathbf{c}_i$ is denoted as $D(\mathbf{v}_y||\mathbf{c}_i)$. If the expectation is taken over Y, then from Lemma 1 in [4], we have:

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

Since $\mathbf{p}_X$ and $\phi_i(y)$, $i = 1,2$ are given, $I(X;Y)$ is given and independent of the quantizer Q. Thus, maximizing $I(X;Z)$ over Q is equivalent to minimizing $\mathbb{E}_Y[D(\mathbf{v}_y||\mathbf{c}_i)]$ with optimal quantizer:

$$Q^* = \min_Q \mathbb{E}_Y[D(\mathbf{v}_y||\mathbf{c}_i)] = \min_Q \sum_{i=1}^{N} \int_{y\in Z_i} D(\mathbf{v}_y||\mathbf{c}_i)\mu(y)dy.$$

Noting that for a given output set $Z_i \neq \emptyset$, the centroid $\mathbf{c}_i$ can be computed by a closed-form expression [26].

IV. Structure of Optimal Quantizers

In this section, we show that any arbitrary quantizer can be replaced by a better *convex cell quantizer* in the sense of maximizing $\beta I(X;Z) - C(\mathbf{p}_Z)$. Thus, an algorithm for finding the best quantizer in the set of all *convex cell quantizers* will find the globally 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 cell quantizers* than to search in through all the possible quantizers. We begin with a special case $N = 2$.

A. Structure of optimal quantizer for $N = 2$

In this section, we consider the quantizer for the case $N = 2$, i.e., output Y is quantized into a binary Z. We show that for any arbitrary quantizer $Q(y)$, there exists a *convex cell quantizer* $\bar{Q}(r_y)$ that: (1) has the same p_Z as $Q(y)$ and (2) the total distortion $D(\bar{Q})$ is less than or equal to $D(Q)$, or equivalently $I(X;Z)$ produced by $\bar{Q}(r_y)$ is at least as large as that produced by $Q(y)$. Thus, to find the optimal quantizer that maximizes $\beta I(X;Z) - C(\mathbf{p}_Z)$, it is sufficient to search in the set of the *convex cell quantizers*. The result is formally stated in Theorem 1.

Theorem 1. Let Q be an arbitrary quantizer that produces two disjoint output sets $\{Z_1, Z_2\}$ corresponding to two centroid vectors $\mathbf{c}_1, \mathbf{c}_2$ such that $\mathbf{c}_1 \leq \mathbf{c}_2$, there exists a *convex cell quantizer* $\bar{Q}$ with two output sets $\{\bar{Z}_1, \bar{Z}_2\}$ and the corresponding centroids $\{\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2\}$ such that $\bar{Z}_1 \leq \bar{Z}_2$, $p_{Z_i} = p_{\bar{Z}_i}$ for $i = 1,2$ and $D(\bar{Q}) \leq D(Q)$.

Proof. (Outline). Suppose a quantizer $Q(y)$ produces $\mathbf{p}_Z = (p_{Z_1}, p_{Z_2})$. Our first claim is that one can always find a *convex cell quantizer* $\bar{Q}$ that produces $\mathbf{p}_{\bar{Z}}$ such that $\mathbf{p}_{\bar{Z}} = \mathbf{p}_Z$. Due to limited space, we just outline the argument for this using Fig. 2. For any arbitrary r_y, as a_1 increases, $p_{\bar{Z}_1}$ increases and $p_{\bar{Z}_2} = 1 - p_{\bar{Z}_1}$ decreases. Thus, we can always choose an appropriate value of a_1 to make $p_{\bar{Z}_1} = p_{Z_1}$, and therefore $p_{\bar{Z}_2} = p_{Z_2}$. Also, in Fig. 2, we have $\bar{Z}_1 \leq \bar{Z}_2$ by Definition 5.

Now, let $A = \bar{Z}_1 \cap Z_2$ and $B = \bar{Z}_2 \cap Z_1$. Note that if A or B is empty set then we can show that $\bar{Z}_1 \equiv \bar{Z}_1$, i.e., the quantizer Q is already a *convex cell quantizer*. From $\bar{Z}_1 \leq \bar{Z}_2$, we have $A \leq B$. Moreover, let p_A and p_B be the probabilities that y is in these sets, respectively, then $p_A = p_B$ which is proven in Appendix A. Also, let $\mathbf{c}_1 = [c_1, 1-c_1] \leq \mathbf{c}_2 = [c_2, 1-c_2]$, $F(r_y) = D(\mathbf{v}_y||\mathbf{c}_1) - D(\mathbf{v}_y||\mathbf{c}_2)$ is a non-decreasing function in r_y. Please see a proof for this in the Appendix B. Next, let

$$D_A^{\max} = \max_{y\in A}[D(\mathbf{v}_y||\mathbf{c}_1) - D(\mathbf{v}_y||\mathbf{c}_2)],$$

$$D_B^{\min} = \min_{y\in B}[D(\mathbf{v}_y||\mathbf{c}_1) - D(\mathbf{v}_y||\mathbf{c}_2)].$$

Since $\int_{y\in A}\mu(y)dy = p_A = p_B = \int_{y\in B}\mu(y)dy$, we have:

$$\int_{y\in A}[D(\mathbf{v}_y||\mathbf{c}_1) - D(\mathbf{v}_y||\mathbf{c}_2)]\mu(y)dy \leq D_A^{\max}\int_{y\in A}\mu(y)dy = D_A^{\max}p_A. \tag{11}$$

$$\int_{y\in A} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy - \int_{y\in A} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy + \int_{y\in\{Z_1\cap\bar{Z}_1\}} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_2\cap\bar{Z}_2\}} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy$$
$$\leq \int_{y\in B} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy - \int_{y\in B} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy + \int_{y\in\{Z_1\cap\bar{Z}_1\}} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_2\cap\bar{Z}_2\}} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy. \quad (6)$$

$$\left(\int_{y\in A} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_1\cap\bar{Z}_1\}} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy\right) + \left(\int_{y\in\{Z_2\cap\bar{Z}_2\}} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy + \int_{y\in B} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy\right)$$
$$\leq \left(\int_{y\in B} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_1\cap\bar{Z}_1\}} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy\right) + \left(\int_{y\in\{Z_2\cap\bar{Z}_2\}} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy + \int_{y\in A} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy\right). \quad (7)$$

$$\int_{y\in\bar{Z}_1} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\bar{Z}_2} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy \leq \int_{y\in Z_1} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in Z_2} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy. \quad (8)$$

$$\int_{y\in\bar{Z}_1} D(\mathbf{v}_y||\bar{\mathbf{c}}_1)\mu(y)dy + \int_{y\in\bar{Z}_2} D(\mathbf{v}_y||\bar{\mathbf{c}}_2)\mu(y)dy \leq \int_{y\in\bar{Z}_1} D(_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\bar{Z}_2} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy. \quad (9)$$

$$\int_{y\in\bar{Z}_1} D(\mathbf{v}_y||\bar{\mathbf{c}}_1)\mu(y)dy + \int_{y\in\bar{Z}_2} D(\mathbf{v}_y||\bar{\mathbf{c}}_2)\mu(y)dy \leq \int_{y\in Z_1} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in Z_2} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy. \quad (10)$$

$$\int_{y\in B}[D(\mathbf{v}_y||\mathbf{c}_1)-D(\mathbf{v}_y||\mathbf{c}_2)]\mu(y)dy \geq D_B^{\min}\int_{y\in B}\mu(y)dy = D_B^{\min}p_B. \quad (12)$$

Since $A \leq B$, by Definition 5 and the monotonic non-decreasing property of $F(r_y)$, we have $D_A^{\max} \leq D_B^{\min}$. Thus, from (11) and (12),

$$\int_{y\in A}[D(\mathbf{v}_y||\mathbf{c}_1)-D(\mathbf{v}_y||\mathbf{c}_2)]\mu(y)dy \leq \int_{y\in B}[D(\mathbf{v}_y||\mathbf{c}_1)-D(\mathbf{v}_y||\mathbf{c}_2)]\mu(y)dy. \quad (13)$$

Adding $\int_{y\in\{Z_1\cap\bar{Z}_1\}} D(\mathbf{v}_y||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_2\cap\bar{Z}_2\}} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy$ to both sides of (13), we obtain (6).

By moving $-\int_{y\in A} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy$ to the RHS and $-\int_{y\in B} D(\mathbf{v}_y||\mathbf{c}_2)\mu(y)dy$ to the LHS of (6), we obtain (7).

Now, note that $A = \{\bar{Z}_1 \cap Z_2\}$ and $\{Z_1 \cap \bar{Z}_1\}$ are disjoint due to Z_1 and Z_2 are disjoint. Thus, $A \cap \{Z_1 \cap \bar{Z}_1\} = \emptyset$ and the integral over A and $\{Z_1 \cap \bar{Z}_1\}$ is equivalent to the integral over $A \cup \{Z_1 \cap \bar{Z}_1\} = \bar{Z}_1$. Similarly, using $B \cup \{Z_2 \cap \bar{Z}_2\} = \bar{Z}_2$, $B \cup \{Z_1 \cap \bar{Z}_1\} = Z_1$ and $A \cup \{Z_2 \cap \bar{Z}_2\} = Z_2$, (8) is obtained from (7).

Now, by using $\bar{\mathbf{c}}_1$ and $\bar{\mathbf{c}}_2$ are the new centroids of $\bar{Z}_1$ and $\bar{Z}_2$, from Definition 2, (9) is constructed.

Finally, from (8) and (9), (10) is established, i.e., $D(\bar{Q}) \leq D(Q)$. □

B. Structure of optimal quantizer for $N > 2$

Theorem 2. Let Q be a quantizer with arbitrary disjoint quantized-output sets $\{Z_1, Z_2, \ldots, Z_N\}$ corresponding to N centroids $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N$ such that $\mathbf{c}_i \leq \mathbf{c}_{i+1}\ \forall\ i$, there exists an other *convex cell quantizer* $\bar{Q}$ with the output sets $\{\bar{Z}_1, \bar{Z}_2, \ldots, \bar{Z}_N\}$ and the corresponding centroids $\{\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2, \ldots, \bar{\mathbf{c}}_N\}$ such that $\bar{Z}_i \leq \bar{Z}_{i+1}$, $p_{Z_i} = p_{\bar{Z}_i}\ \forall\ i$ and $D(\bar{Q}) \leq D(Q)$.

Proof. Due to limited space, we omit the proof and note that the proof can be accomplished using the induction method in which we show that for $N = 2$, it is true from Theorem 1, and assume that $N = k$ is true, we show that $N = k + 1$ is also true. □

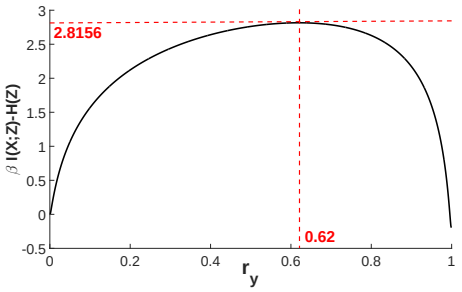

Figure 3: $\beta I(X;Z) - H(Z)$ as a function of r_y.

V. Applications

A. Finding optimal quantizer for $N = 2$ and single threshold quantizer

When $N = 2$, based on Theorem 1, the optimal quantizer has the structure

$$Q^*(r_y) = \begin{cases} z_1 \text{ if } r_y \leq a_1^*, \\ z_2 \text{ if } r_y > a_1^*, \end{cases}$$

for an optimal value $a_1^* \in (0, 1)$. Thus, the optimal quantizer can be found by an exhaustive searching over $a_1^* \in (0, 1)$. The complexity of this algorithm is $O(M)$ where $M = \frac{1}{\epsilon}$ and ϵ is a small number denotes the precise of the solution. If one wants to construct an equivalent classical quantizer $Q(y)$ which compares y to certain thresholds h_i's (rather than comparing r_y with a_1^*), then h_i's are the solutions of $r_y = a_1^*$. Thus, there might be multiple h_i's. However, one can show that if $\frac{\phi_2(y)}{\phi_1(y)}$ is a *strictly increasing/decreasing* function, then there is only one h_1, thus the classical one threshold quantizer is optimal.

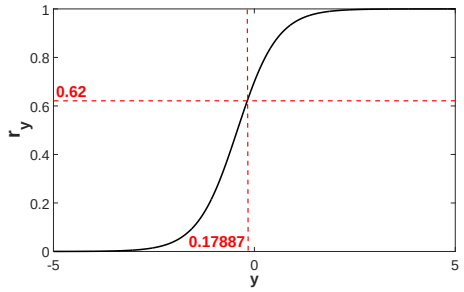

Figure 4: The optimal $r_y^* = 0.62$ corresponds to a single optimal threshold $y^* = -0.17887$.

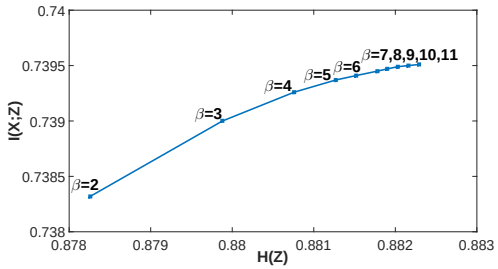

Figure 5: $I^*(X;Z)$ and $H^*(Z)$ for various values of β.

As an example, consider a channel having $p_1 = 0.7$, $p_2 = 0.3$, $\phi_1(y) = N(1,1)$, $\phi_2(y) = N(-1,1)$, using $\beta = 5$ and $C(\mathbf{p}_Z) = H(Z)$ is entropy function of output, Fig. 3 plots $\beta I(X;Z) - H(Z)$ as a function of r_y. As seen, the optimal $r_y^* = 0.62$ which corresponds to $\beta I(X;Z) - H(Z) = 2.8156$. Since $\dfrac{\phi_2(y)}{\phi_1(y)}$ is strictly monotonic, the optimal quantizer has a single threshold $y^* = -0.17887$ that corresponds to $r_y^* = a_1^* = 0.62$. Fig. 4 shows r_y vs. y. Fig. 5 shows the optimal value pairs of mutual information and entropy corresponding to $\beta = 2, 3, \ldots, 11$. As seen, for maximizing $I(X;Z)$ subject to $H(Z) \leq 0.882$, we can pick $\beta = 8$ which produces $I^*(X;Z) = 0.73947$ and $H^*(Z) = 0.8819 \approx 0.882$.

B. Finding optimal quantizer for $N > 2$

From the convex cells property of the optimal quantizer, finding the optimal quantizer is equivalent to finding $N+1$ scalar thresholds $a_0 = 0 < a_1 < \cdots < a_{N-1} < a_N = 1$ as the boundaries such that

$$Q(y) = z_i, \text{ if } a_{i-1} \leq r_y = p_{x_1|y} < a_i.$$

Now, if the constraint of output has the following structure

$$C(\mathbf{p}_Z) = g_1(p_{Z_1}) + g_2(p_{Z_2}) + \cdots + g(p_{Z_N}), \qquad (14)$$

where $g_i(.)$ can be an arbitrary function, then the problem of finding globally optimal quantizer can be cast as a 1-dimensional scalar quantization problem that can be solved efficiently using the well-known dynamic programming [3], [13], [27]. We note that the condition in (14) is not too restricted. In fact, many well-known constraints such as entropy-constrained satisfy this structure.

VI. Discussion

We list few open problems that may interest the reader.

1) Our preliminary results show that our proof technique might be applicable to the problems of maximizing $I(X;Z) - \beta C(p_Z)$ for K-input channels where $K > 2$. The *convex cell quantizer* can be constructed by hyperplane cuts in space of posterior probability $p_{X|Y}$. This result is similar to the result stated in [28] for maximizing $I(X;Z)$ without any constraint on the output.
2) The optimal quantizer that maximizes the mutual information can be solved using dynamic programming in a polynomial time complexity [3]. The time complexity can be further reduced by using SMAWK algorithm technique [5]. Is it possible to use the SMAWK algorithm [17] to find the optimal quantizer that maximizes mutual information under an output constraint? We believe that if the constraint has the structure in (14) and function $g_i(.)$ is convex, $\forall\ i$, then the SMAWK algorithm is applicable.

VII. Conclusion

We describes the structure of an optimal quantizer that maximizes $\beta I(X;Z) - C(\mathbf{p}_Z)$. Our result shows that the structure of the optimal quantizer separates r_y into convex cells. In other words, the optimal quantizer has the form: $Q^*(r_y) = z_i$, if $a_{i-1}^* \leq r_y < a_i^*$, for some optimal thresholds $a_0^* = 0 < a_1^* < a_2^* < \cdots < a_{N-1}^* < a_N^* = 1$. Based on this optimal structure, we described some fast algorithms for determining the optimal quantizers.

Appendix

A. Proof of $p_A = p_B$

$$\begin{aligned} p_B &= p_{\bar{Z}_2 \cap Z_1} = p_{\bar{Z}_2} p_{Z_1|\bar{Z}_2} = p_{\bar{Z}_2}(1 - p_{Z_2|\bar{Z}_2}) && (15)\\ &= p_{\bar{Z}_2} - p_{\bar{Z}_2} p_{Z_2|\bar{Z}_2} = p_{\bar{Z}_2} - p_{Z_2} p_{\bar{Z}_2|Z_2} && (16)\\ &= p_{\bar{Z}_2} - p_{Z_2}(1 - p_{\bar{Z}_1|Z_2}) && (17)\\ &= p_{\bar{Z}_2} - p_{Z_2} + p_{\bar{Z}_1 \cap Z_2} = p_{\bar{Z}_1 \cap Z_2} = p_A. && (18) \end{aligned}$$

with (15) and (17) are due to $p_{Z_1} + p_{Z_2} = 1$ and $p_{\bar{Z}_1} + p_{\bar{Z}_2} = 1$, (16) due to $p_{Z_2} p_{\bar{Z}_2|Z_2} = p_{\bar{Z}_2} p_{Z_2|\bar{Z}_2} = p_{Z_2 \cap \bar{Z}_2}$, (18) due to $p_{\bar{Z}_i} = p_{Z_i}$ for $\forall\ i = 1, 2$.

B. Proof of $F(r_y)$ is a non-decreasing function

We show that for $\mathbf{c}_1 = [c_1, 1-c_1] \leq \mathbf{c}_2 = [c_2, 1-c_2]$ then $F(r_y) = D(\mathbf{v}_y||\mathbf{c}_1) - D(\mathbf{v}_y||\mathbf{c}_2)$ is a non-decreasing function in $p_{x_1|y} = r_y$. Indeed, from Definition 1,

$$D(\mathbf{v}_y||\mathbf{c}_1) - D(\mathbf{v}_y||\mathbf{c}_2) = r_y \log \frac{c_2(1-c_1)}{c_1(1-c_2)} + \log(\frac{1-c_2}{1-c_1}). \qquad (19)$$

Due to $\mathbf{c}_1 \leq \mathbf{c}_2$ implies that $c_1 \leq c_2$, then $F'(r_y) = \log \dfrac{c_2(1-c_1)}{c_1(1-c_2)} \geq 0$. Thus, $F(r_y)$ is a non-decreasing function.

REFERENCES

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

[6] 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.

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

[8] Harish Vangala, Emanuele Viterbo, and Yi Hong. Quantization of binary input dmc at optimal mutual information using constrained shortest path problem. In *2015 22nd International Conference on Telecommunications (ICT)*, pages 151–155. IEEE, 2015.

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

[10] 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.

[11] Thuan Nguyen and Thinh Nguyen. Single-bit quantization capacity of binary-input continuous-output channels. *arXiv preprint arXiv:2001.01842*, 2020.

[12] 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.

[13] Xuan He, Kui Cai, Wentu Song, and Zhen Mei. Dynamic programming for discrete memoryless channel quantization. *arXiv preprint arXiv:1901.01659*, 2019.

[14] 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.

[15] 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.

[16] 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.

[17] Alok Aggarwal, Maria M Klawe, Shlomo Moran, Peter Shor, and Robert Wilber. Geometric applications of a matrix-searching algorithm. *Algorithmica*, 2(1-4):195–208, 1987.

[18] Daniel Marco and David L. Neuhoff. Performance of low rate entropy-constrained scalar quantizers. *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pages 495–, 2004.

[19] Dan Muresan and Michelle Effros. Quantization as histogram segmentation: Optimal scalar quantizer design in network systems. *IEEE Transactions on Information Theory*, 54(1):344–366, 2008.

[20] 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.

[21] 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.

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

[23] David Yuheng Zhao, Jonas Samuelsson, and Mattias Nilsson. On entropy-constrained vector quantization using. 2008.

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

[25] Thuan Nguyen and Thinh Nguyen. Minimizing impurity partition under constraints. *arXiv preprint arXiv:1912.13141*, 2019.

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

[27] Haizhou Wang and Mingzhou Song. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R journal*, 3(2):29, 2011.

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