Communication-Channel Optimized Impurity Partition

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

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322342

***Abstract*—Given an original discrete source X with the distribution p_X that is corrupted by noise to produce a noisy data Y with the given joint distribution $\mathrm{p}_{(X,Y)}$. A quantizer/classifier $Q : Y \to Z$ is then used to classify/quantize Y to a discrete partitioned output Z having probability distribution p_Z. Next, Z is transmitted over a discrete memoryless channel (DMC) with a given channel matrix A that produces the final discrete output T. One wants to design an optimal quantizer/classifier Q^* to minimize the end-to-end impurity/cost function $F(X,T)$ between the input X and the final output T. Our result generalizes some previous results. First, an iteration linear time complexity algorithm is proposed to find the locally optimal quantizer. Second, we show that the optimal quantizers produce the hard partitions that are equivalent to the cuts by hyper-planes in the space of the posterior distribution $\mathrm{p}_{X|Y}$. This result provides a polynomial-time complexity algorithm to find the globally optimal quantizer. Finally, in the special case where the source X is binary, an efficient algorithm is proposed to find the truly global optimal partition.**

Keyword: quantization, impurity, communication channel.

I. Introduction

Channel optimized partition/quantization is a common approach to lossy-compression data source-channel coding that aims to minimize the end-to-end distortion when the quantized/classified data is transmitted over a noisy channel. Due to the huge volume of data and the limited rate of the transmission channel, the data should be coded/quantized at the local stations/nodes before transmitted over a channel to the central station/node. The quality of the relay channel that is specified by its channel matrix, therefore, is important. Of course, one should design the partition/classification based on the channel matrix of the relay channel. From the source coding perspective, the quality of quantization/partition is normally measured by the end-to-end distortion between the input and the final output. While the squared-error distortion often uses to measure the distortion of scalar quantization, it is less appropriate for other problems in communication context, for example, maximizing the mutual information or minimizing the compression rate where other distortion measurements such as Kullback-Leiber divergence are more preferred.

In this paper, we consider the design of quantizer with the aim of minimizing the end-to-end impurity between the input and the final output produced by a relay channel. The impurity termed the loss function that measures the "impurity" of the partitioned sets. Some of the popular impurity functions are entropy function and Gini index [1], [2], [3], [4]. For example, when the empirical entropy of a set is large, this indicates a high level of non-homogeneity of the elements in the set, i.e., "impurity". Impurity function was vastly used in learning theory, decision tree and communication [1], [2], [3], [5], [6], [7]. Therefore, finding the optimal partition minimizing impurity function has various applications. For example, if the impurity is conditional entropy, minimizing impurity is equivalent to maximizing the mutual information between the input and the final output [5], [8], [9], [10], [11]. Therefore, partition/quantization that minimizes the entropy impurity has many applications in design of polar code and LDPC code decoder [12], [13].

To that end, the problem of finding the optimal quantizer that minimizes the end-to-end impurity between the input and the final output of a relay channel is an interesting problem that covers many sub-problems in [5], [8], [14], [15]. For example, if the relay channel matrix is an identity matrix, our setting is back to the model in [5], [8] using conditional entropy impurity function. If the impurity function is conditional entropy, our problem can be viewed as the problem in [14], [15]. It is worth noting that to solving these problems, the methods in [14] and [15] are based on the famous information bottleneck method (IBM) [16] while the results in [5] and [8] are based on the result in [2]. On the other hand, our approach is based on the method in [3] to characterize the necessary condition of the optimal partition. The more detail of these sub-problems can be viewed in Section II, Table I.

The outline of our paper is as follows. In Section II, we describe the problem formulation. In Section III, we provide the optimality condition for the optimal partition. In Section IV, we provide an iteration algorithm that can find a locally optimal solution and show that the optimal partition is equivalent to the cuts by hyper-planes in the probability space of the posterior probability. Based on the hyper-plane cuts, we describe a polynomial time algorithm that can determine the truly global optimal partition if the source X is binary. Finally, we provide a few concluding remarks in Section V.

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

II. Problem Formulation

Fig. 1 illustrates our model. The input set consists of N discrete symbols $X = \{X_1, X_2, \ldots, X_N\}$ with a given pmf $\mathbf{p}_X = [p_1, p_2, \ldots, p_N]$. X is sent over a channel modeled by a conditional distribution $\mathbf{p}_{Y|X}$. The received Y consists of M discrete points $Y = \{Y_1, Y_2, \ldots, Y_M\}$ having the pmf $\mathbf{p}_Y = [p_{Y_1}, p_{Y_2}, \ldots, p_{Y_M}]$ and the joint distribution $p_{(X_n,Y_m)}$, $\forall\ n = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, M$. Y is quantized to into the partitioned output $Z = \{Z_1, Z_2, \ldots, Z_K\}$ having the pmf $\mathbf{p}_Z = [p_{Z_1}, p_{Z_2}, \ldots, p_{Z_K}]$ using a quantizer $Q : Y \to Z$. Note that Q can be a stochastic quantizer i.e., $0 \leq p_{Z_k|Y_m} \leq 1$. The partitioned output Z is then transmitted over a relay channel having a channel matrix A to result in the final output $T = \{T_1, T_2, \ldots, T_H\}$ having the pmf $\mathbf{p}_T = [p_{T_1}, p_{T_2}, \ldots, p_{T_H}]$. A is a stochastic matrix where entry A_{kh} denotes the conditional probability $p_{T_h|Z_k}$ i.e., the probability of transmitter transmits Z_k but the receiver receives T_h. Quantizer Q is a mapping from Y to Z as illustrated in Fig. 2. Our goal is finding an optimal quantizer Q^* that minimizes the end-to-end impurity/cost function $F(X, T)$ between the input X and the final output T.

To design the optimal quantizer Q^* such that the end-to-end impurity function is minimized, we are interested in solving the following optimization problem:

$$Q^* = \min_{Q} F(X; T), \tag{1}$$

where the impurity/cost function takes the form:

$$F(X, T) = \sum_{h=1}^{H} F(X, T_h), \tag{2}$$

with $F(X, T_h)$ denotes the impurity/cost corresponding to the final output T_h.

$$F(X, T_h) = p_{T_h} f(p_{X_1|T_h}, p_{X_2|T_h}, \ldots, p_{X_N|T_h}). \tag{3}$$

The total impurity/cost $F(X, T)$, therefore, is the summation of impurity/cost in each final output $F(X, T_h)$. The factor p_{T_h} denotes the weight of the final output T_h, $f(.)$ is a *concave* function that measures the impurity/cost in final output T_h and $p_{X_n|T_h}$ denotes the conditional probability of X_n given T_h. For convenient, define:

$$\mathbf{p}_{(X,Y_m)} = [p_{(X_1,Y_m)}, p_{(X_2,Y_m)}, \ldots, p_{(X_N,Y_m)}], \tag{4}$$

$$\mathbf{p}_{(X,T_h)} = [p_{(X_1,T_h)}, p_{(X_2,T_h)}, \ldots, p_{(X_N,T_h)}], \tag{5}$$

$$\mathbf{p}_{X|T_h} = [p_{X_1|T_h}, p_{X_2|T_h}, \ldots, p_{X_N|T_h}]. \tag{6}$$

Now, suppose that a quantizer Q quantizes $Q(Y_m) \to Z_k$ with the probability $p_{Z_k|Y_m}$, then:

$$p_{(X_n,Z_k)} = \sum_{Y_m \in Y} p_{(X_n,Y_m)} p_{Z_k|Y_m}. \tag{7}$$

However, the final output T can be computed via the partitioned output Z and the given channel matrix A. Thus, $p_{(X_n,T_h)}$ can be determined by:

$$p_{(X_n,T_h)} = \sum_{k=1}^{K} p_{(X_n,Z_k)} A_{kh}. \tag{8}$$

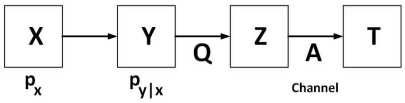

Figure 1: The quantizer/classifier Q is designed to minimize the impurity function between the input X and the final output T.

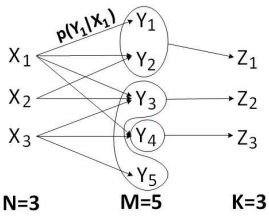

Figure 2: Quantizer Q is a mapping from Y to Z.

Now, the impurity function in each final output T_h can be rewritten by:

$$F(X, T_h) = (\sum_{n=1}^{N} p_{(X_n,T_h)}) f\big(\frac{p_{(X_1,T_h)}}{\sum_{n=1}^{N} p_{(X_n,T_h)}}, \ldots, \frac{p_{(X_N,T_h)}}{\sum_{n=1}^{N} p_{(X_n,T_h)}}\big)$$

where $\sum_{n=1}^{N} p_{(X_n,T_h)}$ is the weight of T_h and $\frac{p_{(X_n,T_h)}}{\sum_{n=1}^{N} p_{(X_n,T_h)}}$ denotes the conditional distribution $p_{X_n|T_h}$. *The function $F(X, T_h)$, therefore, is only the function of the joint distribution vector* $\mathbf{p}_{(X,T_h)} = [p_{(X_1,T_h)}, p_{(X_2,T_h)}, \ldots, p_{(X_N,T_h)}]$. For convenient, in the rest of this paper, we denote $F(X, T_h)$ by $F(\mathbf{p}_{(X,T_h)})$.

Noting that the impurity function $f(.)$ is *concave* and satisfies the following inequality:

$$f(\lambda \mathbf{a} + (1 - \lambda)\mathbf{b}) \geq \lambda f(\mathbf{a}) + (1 - \lambda) f(\mathbf{b}), \forall \lambda \in (0, 1), \tag{9}$$

for all probability vector $\mathbf{a} = [a_1, a_2, \ldots, a_N]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_N]$ with equality happens if and only if $\mathbf{a} = \mathbf{b}$.

Corresponding to the setting of $f(.)$, N and the channel matrix A, our problem generalizes many sub-problems as listed in Table I. Based on the concave property of $f(.)$, an iteration algorithm is proposed to find the locally optimal quantizer. Moreover, we show that the optimal quantizers (local and global) produce a hard partition that is equivalent to the cuts by hyper-planes in the space of the posterior probability $p_{X|Y}$. This interesting property finally yields a polynomial time algorithm to determine the truly global optimal quantizer.

Ref.	A	f(.)	N
[2]	Identity matrix	x	x
[5]	Identity matrix	Entropy	2
[8]	Identity matrix	Entropy	x
[15]	x	Entropy	x
[14]	x	Entropy	x
[17]	Identity matrix	Gini index	x
[18]	Identity matrix	Gini index	x

Table I: Our problem generalizes many existing sub-problems. Symbol "x" is used if there is no specific setting.

III. Properties of Optimal Partitions

We first begin with some properties of $F(X, T_h)$.

Proposition 1. The impurity/cost in each subset T_h is defined by $F(X, T_h)$ which has the following properties:

(i) The impurity/cost function is proportional increasing/ decreasing to its weight: if $\mathbf{p}_{(X,T_a)} = \lambda \mathbf{p}_{(X,T_b)}$, then

$$\frac{F(X, T_a)}{F(X, T_b)} = \lambda. \tag{10}$$

(ii) The impurity gain after partition is always non-negative: If $\mathbf{p}_{(X,T_a)} = \mathbf{p}_{(X,T_b)} + \mathbf{p}_{(X,T_c)}$, then

$$F(X, T_a) \geq F(X, T_b) + F(X, T_c). \tag{11}$$

Proof. (i) From $\mathbf{p}_{(X,T_a)} = \lambda \mathbf{p}_{(X,T_b)}$, then $\mathbf{p}_{X|T_a} = \mathbf{p}_{X|T_b}$ and $p_{T_a} = \lambda p_{T_b}$. Thus, using the definition of $F(X, T_h)$ in (3), it is obviously to prove the first property.

(ii) By dividing both side of $\mathbf{p}_{(X,T_a)} = \mathbf{p}_{(X,T_b)} + \mathbf{p}_{(X,T_c)}$ to p_{T_a}, we have

$$\mathbf{p}_{X|T_a} = \frac{p_{T_b}}{p_{T_a}} \mathbf{p}_{X|T_b} + \frac{p_{T_c}}{p_{T_a}} \mathbf{p}_{X|T_c}. \tag{12}$$

Now, using the original definition in (3),

$$\begin{aligned} F(X, T_a) &= p_{T_a} f(\mathbf{p}_{X|T_a}) \\ &= p_{T_a} f[\frac{p_{T_b}}{p_{T_a}} \mathbf{p}_{X|T_b} + \frac{p_{T_c}}{p_{T_a}} \mathbf{p}_{X|T_c}] && (13) \\ &\geq p_{T_a} [\frac{p_{T_b}}{p_{T_a}} f(\mathbf{p}_{X|T_b}) + \frac{p_{T_c}}{p_{T_a}} f(\mathbf{p}_{X|T_c})] && (14) \\ &= p_{T_b} f(\mathbf{p}_{X|T_b}) + p_{T_c} f(\mathbf{p}_{X|T_c}) \\ &= F(X, T_b) + F(X, T_c), && (15) \end{aligned}$$

with (13) is due to (12), (14) due to concave property of $f(.)$ which is defined in (9) using $\lambda = \frac{p_{T_b}}{p_{T_a}}$, $1 - \lambda = \frac{p_{T_c}}{p_{T_a}}$, (15) due to the definitions in (3) and (5). □

Now, we are ready to show the main result which characterizes the condition for the optimal partition Q^*.

Theorem 1. Suppose that the optimal quantizer Q^* yields the optimal partitioned output $Z = \{Z_1, Z_2, \ldots, Z_K\}$ and the optimal final output $T = \{T_1, T_2, \ldots, T_H\}$. We define vector $\mathbf{c}_k = [c_k^1, c_k^2, \ldots, c_k^N]$, $k = 1, 2, \ldots, T$ where:

$$c_k^n = \frac{\partial F(p_{(X,T_k)})}{\partial p_{(X_n,T_k)}}, \forall n \in \{1, 2, \ldots, N\}. \tag{16}$$

Define the "distance" from $Y_m \in Y$ to Z_k is:

$$D(Y_m, Z_k) = \sum_{h=1}^{H} \sum_{n=1}^{N} [c_k^n p_{(X_n,Y_m)}] A_{kh}. \tag{17}$$

Then,

(i) The globally optimal quantizer of the problem (1) is a deterministic quantizer (hard clustering) i.e., $p_{Z_i|Y_j} \in \{0, 1\}$, $\forall\ i, j$.

(ii) Data Y_m is quantized to Z_s if $D(Y_m, Z_s) \leq D(Y_m, Z_k)$ for $\forall k \in \{1, 2, \ldots, K\}$ and $s \neq k$.

Proof. Due to the limited space, we only provide the outline of proof. Suppose that $D(Y_m, Z_s) \leq D(Y_m, Z_k)$ for $\forall k \in \{1, 2, \ldots, K\}$ and $s \neq k$. Consider two arbitrary optimal partitioned outputs Z_q and Z_s and a trial data Y_m. Consider a *soft partition* optimal quantizer Q^* that allocates Y_m to Z_q with the probability of $p_{Z_q|Y_m} = v$, $0 < v < 1$. We remind that $\mathbf{p}_{(X,Y_m)} = [p_{(X_1,Y_m)}, p_{(X_2,Y_m)}, \ldots, p_{(X_N,Y_m)}]$ denotes the joint distribution in the sample Y_m. We will determine the change of impurity function $F(X, T)$ as a function of t when changing amount of $tv\mathbf{p}_{(X,Y_m)}$ from $\mathbf{p}_{(X,Z_q)}$ to $\mathbf{p}_{(X,Z_s)}$ where t is a scalar and $0 < t < 1$.

Now, by changing $tv\mathbf{p}_{(X,Y_m)}$, the new joint distributions in Z_q and Z_s are $\mathbf{p}_{(X,Z_q)} - tv\mathbf{p}_{(X,Y_m)}$ and $\mathbf{p}_{(X,Z_s)} + tv\mathbf{p}_{(X,Y_m)}$, respectively. Denote the new joint distribution in each final output T_h after changing $tv\mathbf{p}_{(X,Y_m)}$ as a function of variable t is $\mathbf{p}_{(X,T_h)_t}$, from (8):

$$\begin{aligned} \mathbf{p}_{(X,T_h)_t} &= \mathbf{p}_{(X,T_h)} - tv\mathbf{p}_{(X,Y_m)} A_{qh} + tv\mathbf{p}_{(X,Y_m)} A_{sh} \\ &= \mathbf{p}_{(X,T_h)} + tv\mathbf{p}_{(X,Y_m)} (A_{sh} - A_{qh}). \end{aligned}$$

Now, denote $tv\mathbf{p}_{(X,Y_m)}(A_{sh} - A_{qh}) = \delta_{th}$. The total change of impurity function $F(X, T)$ is:

$$I_t = \sum_{h=1}^{H} F(\mathbf{p}_{(X,T_h)} + \delta_{th}). \tag{18}$$

However, from (18) and (16):

$$\frac{\partial I_t}{\partial t}|_{t=0} = v \sum_{h=1}^{H} \sum_{n=1}^{N} (c_k^n p_{(X_n,Y_m)})(A_{sh} - A_{qh}). \tag{19}$$

From (19) and (17), we have:

$$\frac{\partial I_t}{\partial t}|_{t=0} = v[D(Y_m, Z_s) - D(Y_m, Z_q)].$$

From the assumption that $D(Y_m, Z_s) \leq D(Y_m, Z_k)$, $\forall k \in \{1, 2, \ldots, K\}$, then $D(Y_m, Z_s) \leq D(Y_m, Z_q)$. Thus,

$$\frac{\partial I_t}{\partial t}|_{t=0} \leq 0. \tag{20}$$

Proposition 2. Consider I_t which is defined in (18). For $0 < t < a < 1$, we have:

$$\frac{I_t - I_0}{t} \geq \frac{I_a - I_0}{a}. \tag{21}$$

Proof. Due to the limited space, we sketch the proof as following. First, (21) is equivalent to:

$$I_t \geq (1 - \frac{t}{a}) I_0 + \frac{t}{a} I_a. \tag{22}$$

Noting that I_t (in (18)) is the summation of the impurity in each partition i.e., $I_t = \sum_{h=1}^{H} F(\mathbf{p}_{(X,T_h)} + \delta_{th})$ and $F(.)$ admits the properties in Proposition 1-(ii). Thus,

$$F(\mathbf{p}_{(X,T_h)} + \delta_{th}) \geq (1 - \frac{t}{a}) F(\mathbf{p}_{(X,T_h)} + \delta_{0h}) + \frac{t}{a} F(\mathbf{p}_{(X,T_h)} + \delta_{ah}), \tag{23}$$

where δ_{0h} denotes δ_{th} at $t = 0$. Summing up (23) for $h = 1, 2, \ldots, H$ and using a bit of algebra, (22) follows. Please see the full proof in our extension version. □

Now, we continue to the proof of Theorem 1. From Proposition 2 and the assumption in (20), we have:

$$0 \geq \frac{\partial I_t}{\partial t}|_{t=0} = \lim \frac{I_t - I_0}{t} \geq \frac{I_1 - I_0}{1}.$$

Thus, $I_0 \geq I_1$ which obviously implies that by completely changing amount of $v\mathbf{p}_{(X,Y_m)}$ from $\mathbf{p}_{(X,Z_q)}$ to $\mathbf{p}_{(X,Z_s)}$, the total of the impurity is obviously non-increasing. After that, $p_{Z_q|Y_m} = 0$ while $p_{Z_s|Y_m}$ increases an amount of v. By the induction method and the assumption that $D(Y_m, Z_s) \leq D(Y_m, Z_k)$, $\forall k \in \{1, 2, \ldots, K\}$, by completely changing Y_m from Z_k to Z_s, $\forall k \in \{1, 2, \ldots, K\}$, $k \neq s$, the total of the impurity is obviously non-increasing. Therefore, if $D(Y_m, Z_s) \leq D(Y_m, Z_k)$ $\forall k$, a new quantizer having $p_{Z_k|Y_m} = 0$, $\forall\ k \neq s$ and $p_{Z_s|Y_m} = 1$ provides the impurity at least as the impurity of the *soft partition* quantizer Q^*. Thus, (i) the globally optimal quantizer of the problem (1) is a deterministic quantizer (hard clustering) i.e., $p_{Z_i|Y_j} \in \{0, 1\}$, $\forall\ i, j$ and (ii) data Y_m is quantized to Z_s if $D(Y_m, Z_s) \leq D(Y_m, Z_k)$ for $\forall k \in \{1, 2, \ldots, K\}$ and $s \neq k$. □

Remark: To find the optimal quantizer, we only need to search over all the possible hard quantizers.

IV. Algorithms

A. Practical Algorithm

From the optimality condition in Theorem 1, we should allocate the data Y_m to the partitioned output Z_k if and only if the "distance" $D(Y_m, Z_k)$ is shortest. Therefore, a simple alternative optimization algorithm that is very similar to the k-means algorithm can be applied to find the locally optimal solution. Our algorithm is proposed in Algorithm 1. We also note that the distance $D(Y_m, Z_k)$ is:

$$\begin{aligned} D(Y_m, Z_k) &= \sum_{h=1}^{H}\sum_{n=1}^{N}[c_k^n p_{(X_n,Y_m)}]A_{kh} \\ &= p_{Y_m}\sum_{h=1}^{H}\sum_{n=1}^{N}[c_k^n p_{X_n|Y_m}]A_{kh}. \end{aligned}$$

Therefore, one can ignore the constant p_{Y_m} while comparing the distance $D(Y_m, Z_k)$ and use a simpler version distance $D'(Y_m, Z_k)$ as follows:

$$D'(Y_m, Z_k) = \sum_{h=1}^{H}\sum_{n=1}^{N}[c_k^n p_{X_n|Y_m}]A_{kh}. \quad (24)$$

Algorithm 1 works similarly to the k-means algorithm. The distances from each data point $Y_m \in Y$ to each partitioned output $Z_k \in Z$ are updated per iterations. Next, Y_m will be assigned to Z_k if $D(Y_m, Z_k)$ is the shortest distance. The complexity of this algorithm, therefore, is $O(TNKM)$ where T is the number of iterations, N, K, M are the size of data dimensional, the size of partitioned set Z and the size of data set Y. Noting that in the case of the impurity function is entropy and the channel matrix is identity matrix, our algorithm is identical to the algorithm in [8].

Algorithm 1 Finding the communication optimized partition

1: **Input**: $\mathbf{p}_X$, $\mathbf{p}_Y$, $\mathbf{p}_{(X,Y)}$, $f(.)$.
2: **Output**: $Z = \{Z_1, Z_2, \ldots, Z_K\}$.
3: **Initialization**: Randomly hard cluster Y into K clusters.
4: **Step 1**: Updating $\mathbf{p}_{(X,Z_k)}$ and $\mathbf{p}_{(X,T_h)}$ for $\forall\ k \in \{1, 2, \ldots, K\}$ and $h \in \{1, 2, \ldots, H\}$:

$$\begin{aligned} p_{(X_n,Z_k)} &= \sum_{Y_m \in Z_k} p_{(X_n,Y_m)}, \\ p_{(X_n,T_h)} &= \sum_{k=1}^{K} p_{(X_n,Z_k)}A_{kh}, \\ c_k^n &= \frac{\partial F(\mathbf{p}_{(X,T_k)})}{\partial p_{(X_n,T_k)}}, \forall n \in \{1, 2, \ldots, N\}. \end{aligned}$$

5: **Step 2**: Updating the membership by measurement the distance from each $Y_m \in Y$ to each $Z_k \in Z$:

$$Z_k = \{Y_m | D(Y_m, Z_k) \leq D(Y_m, Z_s)\}, \forall s \neq k, \quad (25)$$

where $D(Y_m, Z_k)$ is defined in (17) or in (24).
6: **Step 3**: Go to Step 1 until all the partitioned outputs $\{Z_1, Z_2, \ldots, Z_K\}$ stop changing or the maximum number of iterations has been reached.

B. Hyper-plane separation

Similar to the work in [2], it is possible to show that the optimal partition is equivalent to the cuts by hyper-planes in the space of the posterior distribution. Therefore, existing a polynomial time algorithm that can find the globally optimal quantizer. The proof is based on the optimality condition in Theorem 1 which states that if the quantizer Q^* is optimal then $\forall Y_m \in Z_k$, $D'(Y_m, Z_k) \leq D'(Y_m, Z_s)$, $\forall s \neq k$. We refer the reader to a similar proof in [6]. Due to the limited space, the details of proof will be presented in our extension version.

C. Globally optimal quantizer for binary input data

Algorithm 1 is possible to find the optimal quantizer in a linear time complexity of $O(TNKM)$ where T is the number of iterations, N, K, M are the size of data dimensional, the size of partitioned set Z and the size of data set Y. Unfortunately, this algorithms can get stuck at a locally optimal solution which can be far away from the globally optimal solutions. In this section, we show that if the data is binary ($N = 2$), then the global optimal quantizer can be found efficiently in a polynomial time complexity of $O(M^3)$.

Theorem 2. If the input data is binary ($|X| = N = 2$), the globally optimal partition can be found in polynomial time complexity of $O(M^3)$.

Proof. (Sketch). From the result in Sec. IV-B, the optimal partition is equivalent to the cuts by hyper-planes in the space of the posterior distribution. If $N = 2$, $X = \{X_1, X_2\}$, then:

$$\mathbf{p}_{X|Y_j} = [p_{X_1|Y_j}, p_{X_2|Y_j}] = [p_{X_1|Y_j}, 1 - p_{X_1|Y_j}]. \quad (26)$$

Thus, the space of the posterior distribution is 1-dimensional space respect to the unique variable $p_{X_1|Y_j}$ for $j = 1, 2, \ldots, M$ and a hyper-plane in this space is obviously a scalar between zero and one. To achieve K partitioned outputs, one requires $K-1$ hyper-planes or $K-1$ scalars $a_1, a_2, \ldots, a_{K-1}$ such that:

$$0 \leq a_1 < a_2 < \ldots, a_{K-1} \leq 1.$$

Now, each partitioned output Z_i is separated by two scalars a_{i-1} and a_i such that if $a_{i-1} \leq p_{X_1|Y_j} \leq a_i$ then Y_j is quantized to Z_i ($Q(Y_j) = Z_i$). Therefore, the problem of finding the optimal partition can be cast as the problem of 1-dimensional quantization that can be solved efficiently using dynamic programming algorithm [5]. The time complexity of the dynamic programming is $O(M^3)$ in the worst case [5]. Thus, the globally optimal partition can be found in polynomial time complexity of $O(M^3)$. For the detail algorithm, we refer the reader to a very similar work in [5] for quantization that maximizes mutual information of binary input channels. □

D. Numerical results

We end this section by presenting an example of binary input channel. Consider a binary input $X = \{X_1 = -1, X_2 = 1\}$ having $\mathbf{p}_X = [0.7, 0.3]$ which is transmitted over an additive noisy channel with a normal distribution $N(0, 1)$. Due to the additive property, the received output $Y \in \mathbb{R}$ is a continuous-valued signal with the conditional distributions $p_{Y|X_1} = N(-1, 1)$ and $p_{Y|X_2} = N(1, 1)$. Y then is quantized into $K = 3$ partitioned outputs $Z = \{Z_1, Z_2, Z_3\}$. Next, Z is transmitted over a relay channel with channel matrix A to result in a final output T.

$$A = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}.$$

Our goal is to design an optimal quantizer Q^* that maximizes the mutual information $I(X;T)$ between the input X and the final output T. Noting that $\mathbf{p}_X$ is given and $I(X;T) = H(X) - H(X|T)$, maximizing $I(X;T)$ is equivalent to minimizing the conditional entropy $H(X|T)$. It is possible to verify that the conditional entropy satisfies conditions of a impurity function [5]. Thus, all of the proposed algorithms can be applied to design the optimal quantizer Q^* that maximizes $I(X;T)$. To use the proposed algorithms, we first discretize Y to $M = 200$ pieces $Y = [Y_1, Y_2, \ldots, Y_{200}]$ from $[-10, 10]$ with the same interval width of $\epsilon = 0.1$. Next, $Y = [Y_1, Y_2, \ldots, Y_{200}]$ is sorted according to the increasing order of $p_{X_1|Y}$. The dynamic programming is used to find the optimal partition. Finally, the optimal partition is achieved at $Z_1^* = \{Y| -10 \leq Y < -0.5\}$, $Z_2^* = \{Y| -0.5 \leq Y < 0.2\}$ and $Z_3^* = \{Y|0.2 \leq Y \leq 10\}$ which produces $I^*(X;T) = 0.20953$. The running time of the dynamic programming is 2.32 seconds. Noting that a smaller value of ϵ (or a higher value of M) results in a higher accuracy of the optimal partition at the expense of a larger time and memory complexities.

V. Conclusion

The problem of designing the optimal quantizer that minimizes the end-to-end impurity function between the input and the final output is investigated. Our result generalizes some previous results. An iteration algorithm is proposed to find the locally optimal quantizer in a linear time complexity. In additional, we also show that the optimal quantizer produces a hard partition that is equivalent to hyper-plane cuts in the probability space of the posterior probability. Thus, there exists a polynomial time algorithm that can determine the globally optimal quantizer.

References

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

[2] David Burshtein, Vincent Della Pietra, Dimitri Kanevsky, and Arthur Nadas. Minimum impurity partitions. *The Annals of Statistics*, pages 1637–1646, 1992.

[3] 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.

[4] 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.

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

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

[7] Thuan Nguyen and Thinh 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. IEEE, 2020.

[8] 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.

[9] 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.

[10] Thuan Nguyen and Thinh Nguyen. On binary quantizer for maximizing mutual information. *IEEE Transactions on Communications*, 2020.

[11] Thuan Nguyen and Thinh Nguyen. Thresholding quantizer design for mutual information maximization under output constraint.

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

[13] 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.

[14] 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.

[15] S. Hassanpour, D. Wubben, and A. Dekorsy. On the equivalence of two information bottleneck-based routines devised for joint source-channel coding. In *2018 25th International Conference on Telecommunications (ICT)*, pages 253–258, 2018.

[16] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[17] 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.

[18] Eduardo Laber and Lucas Murtinho. Minimization of gini impurity: Np-completeness and approximation algorithm via connections with the k-means problem. *Electronic Notes in Theoretical Computer Science*, 346:567–576, 2019.