AN ABSTRACT OF THE DISSERTATION OF

Thuan Nguyen for the degree of Doctor of Philosophy in Electrical and Computer Engineering presented on February 22, 2021.

Title: INFORMATION THEORETIC APPROACH TO QUANTIZATION AND CLASSIFICATION FOR SIGNAL PROCESSING, COMMUNICATIONS, AND MACHINE LEARNING APPLICATIONS

Abstract approved: _____

Thinh P. Nguyen

There are five main contributions of this dissertation. The first contribution is new closed-form expressions for channel capacity of a new class of channel matrices. The second contribution is the discovery of the structure for optimal binary quantizer and the associated methods for finding an optimal quantizer that maximizes mutual information between the input and output for a given input distribution. The third contribution is the discovery of the structure for an optimal K-ary quantizer that maximizes the mutual information subject to an arbitrary constraint on the output distribution. The fourth contribution is the joint design of an optimal quantizer that maximizes the mutual information over both the input distribution and the quantization parameters for an arbitrary binary noisy channel with a given noise density. The last contribution is the development and analysis of novel efficient classification algorithms for finding the minimum impurity partition using mutual information as the metric.

INFORMATION THEORETIC APPROACH TO QUANTIZATION AND CLASSIFICATION FOR SIGNAL PROCESSING, COMMUNICATIONS, AND MACHINE LEARNING APPLICATIONS

by

Thuan Nguyen

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented February 22, 2021
Commencement June 2021

Doctor of Philosophy dissertation of Thuan Nguyen presented on February 22, 2021.

APPROVED:

Major Professor, representing Electrical and Computer Engineering

---

Head of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Thuan Nguyen, Author

ACKNOWLEDGEMENTS

This dissertation is dedicated to my dad, my mom, my younger brother, and my beloved family. I would first like to thank my supervisor, Prof. Thinh Nguyen, who has been supported me over the past five years. Thank you so much from the bottom of my heart for your encouragement and patience and for everything you did for me like a father treats his son. I would like to thank my dad, my mom, and my younger brother who always facilitate and support me along my journey. Especially, I would like to express my deep appreciation to my wife and my two lovely daughters. Everything will be extremely hard without your love and sacrifice. I am also thankful to all of my friends for all the unconditional support. The warmth from the friendship makes me love this land and feel it as my second hometown. Finally, I would like to thank all of my graduate committee members, who give me their valuable time to improve my dissertation and assist me during my job search.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF ALGORITHMS

Chapter 1: Introduction

Information theory is the scientific study of the quantification, storage, and transmission of information. Built on the previous ideas of Nyquist and Hartley, Claude Shannon formulated the basics of information theory in his ground-breaking 1948 paper titled "A Mathematical Theory of Communication" [2]. Historically, Shannon's information theory was developed to study the fundamental limits of communications from an engineering perspective. Over the past 70 years, information theory has provided a powerful foundation and paved the way for all modern communication and signal processing technologies. Importantly, information theory has also been playing key roles and inspirations in the development of many other scientific disciplines beyond engineering such as computer science, physics, statistics, and biology, just to name a few. A key measure in information theory is mutual information which quantifies the amount of the shared information between two random variables. The maximum value of mutual information between two random variables that model the input and the output of a communication channel is defined as the channel capacity. Based on the well-known Shannon's channel coding theorem, the highest achievable rate that information can be transmitted over a noisy channel with arbitrarily small error probability, is the channel capacity. Therefore, maximizing mutual information between the input and the output of a channel is one of the most useful and interesting problems in communications and signal processing. That said, the significance of mutual information extends beyond the fields of communications and signal processing. Indeed, mutual information is a key metric behind many successful algorithms in statistics and machine learning. To that end, the central theme of this dissertation is the development of novel information-theoretic based approach for quantization algorithms (from a signal and communication perspective) and classi-

fication algorithms (from a statistics and machine learning perspective) that aims to maximize mutual information between the input and the quantized/classified output.

There are five main contributions of this dissertation. Each contribution is presented in a separate chapter and can be read as a self-contained article. The first contribution is new closed-form expressions for channel capacity of a new class of channel matrices. The second contribution is the discovery of the structure for optimal binary quantizer and the associated methods for finding an optimal quantizer that maximizes mutual information between the input and output for a given input distribution. The third contribution is the discovery of the structure for an optimal K-ary quantizer that maximizes the mutual information subject to an arbitrary constraint on the output distribution. The fourth contribution is the joint design of an optimal quantizer that maximizes the mutual information over both the input distribution and the quantization parameters for an arbitrary binary noisy channel with a given noise density. The last contribution is the development and analysis of novel efficient classification algorithms for finding the minimum impurity partition using mutual information as the metric. We now provide an abstract for each contribution.

First contribution. While capacities of discrete memoryless channels are well studied, it is still not possible to obtain a closed-form expression for the capacity of an arbitrary discrete memoryless channel (DMC). In this contribution, we study a class of DMCs whose channel matrix is an invertible positive matrix. This class of channel matrices can be used to model many real-world settings. Next, an elementary technique based on Karush-Kuhn-Tucker (KKT) conditions is used to obtain (1) a good upper bound of channel capacity of a discrete memoryless channel having an invertible positive channel matrix and (2) a closed-form expression for the capacity if the channel matrix satisfies certain conditions related to its singular value and its Gershgorin's disk.

Second contribution. We consider a channel with a binary input X being corrupted by a

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

Third contribution. We consider a channel whose the input contains K discrete symbols modeled as a discrete random variable X having a probability mass function $\mathbf{p}(x) = [p(x_1), p(x_2), \ldots, p(x_K)]$ 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 $Z = \{z_1, \ldots, z_N\}$ such that the mutual information $I(X;Z)$ is maximized subject to an arbitrary constraint on $\mathbf{p}(z) = [p(z_1), p(z_2), \ldots, p(z_N)]$. Formally, we are interested in designing an optimal quantizer Q^* that maximizes $\beta I(X;Z) - C(\mathbf{p}(z))$ where β is a positive number that controls the trade-off between maximizing $I(X;Z)$ and minimizing an arbitrary cost function $C(\mathbf{p}(z))$. 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 given value 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.

Four contribution. We consider a communication channel with a binary input X being distorted by an arbitrary continuous-valued noise which results in a continuous-valued signal Y at the receiver. A quantizer Q is used to quantize Y back to a binary output Z. Our goal is to determine the optimal quantizer Q^* and the corresponding input probability mass function $\mathbf{p}_X^*$ that achieve the capacity. We present two new lower and upper bounds on the capacity in terms of quantization parameters, based on which we propose an efficient algorithm for finding the optimal quantizer.

Fifth contribution. Partitioning algorithms play a key role in many scientific and engineering disciplines. A partitioning algorithm divides a set into a number of disjoint subsets or partitions. Often, the quality of the resulted partitions is measured by the amount of impurity in each partition, the smaller impurity the higher quality of the partitions. In general, for a given impurity measure specified by a function of the partitions, finding the minimum impurity partitions is an NP-hard problem. Let M be the number of N-dimensional elements in a set and K be the number of desired partitions, then an exhaustive search over all the possible partitions to find a minimum partition has the complexity of $O(K^M)$ which quickly becomes impractical for many applications with modest values of K and M. Thus, many approximate algorithms with polynomial time complexity have been proposed, but few provide the bounded guarantee. In this paper, we propose a linear time algorithm with bounded guarantee based on the maximum likelihood principle. Furthermore, the guarantee bound of the proposed algorithm is better than the state-of-the-art method in [1] for many impurity functions, and at the same time, for $K \geq N$, the computational complexity is reduced from a polynomial time complexity $O(M^3)$ to a linear time complexity $O(NM)$. Both theoretical and practical results are provided to illustrate the advantages of the proposed algorithm.

Chapter 2: Bounds and Closed-Form Expressions for Capacities of Discrete Memoryless Channels with Invertible Positive Matrices

2.1 Introduction

Discrete memoryless channels (DMC) play a critical role in the early development of information theory and its applications. DMCs are especially useful for studying many well-known modulation/demodulation schemes (e.g., PSK and QAM) in which the continuous inputs and outputs of a channel are quantized into discrete symbols. Thus, there exists a rich literature on the capacities of DMCs [3–9]. In particular, capacities of many well-known channels such as (weakly) symmetric channels can be written in elementary formulas [3]. However, it is often not possible to express the capacity of an arbitrary DMC in a closed-form expression [3]. Recently, several papers have been able to obtain closed-form expressions for a small class of DMCs with small alphabets. For example, Martin et al. established closed-form expression for a general binary channel [10]. Liang showed that the capacity of channels with two inputs and three outputs can be expressed as an infinite series [11]. Paul Cotae et al. found the capacity of two input and two output channels in term of the eigenvalues of the channel matrices [12]. In [13], the authors used geometric programming to construct a simple closed-form expression for the upper bound of the capacity of an arbitrary DMC. It is worth noting that the approach in [13] based on elementary Lagrange functions is similar to our approach in this chapter. On the other hand, the problem of finding the capacity of a discrete memoryless channel can be formulated as a convex optimization problem [14], [15]. Thus, efficient algorithmic solutions exist. There are also iterative algorithms such

as Arimoto-Blahut algorithm [4], [5] and other variants for computing channel capacities [16–20]. Even though there exist efficient algorithms for finding the capacity of an arbitrary DMC, there are a number of reasons why channel capacity or bounds expressed in closed-form expression can be very useful. These include (1) formulas can often provide a good intuition about the relationship between the capacity and different channel parameters, (2) formulas offer a faster way to determine the capacity than that of algorithms, and (3) formulas are useful for analytical derivations where closed-form expression of the capacity is needed in the intermediate steps. Moreover, the channel capacity or bounds expressed in closed-form expression might be particularly useful for channels having large alphabet sizes since the well-known Arimoto-Blahut algorithm already provides the capacity values fairly quickly for channels with small alphabet sizes. In fact, our work is motivated by our current work on a prototype of a Free Space Optical communication system called WiFO [21]. WiFO's transceiver is capable of adjusting transmitting and receiving parameters for power and coverage optimization. The result is that the channel matrix can be changed dynamically. For a given channel matrix, we want to know the closed-form expression of the channel capacity so that a trade-off among power consumption, coverage, and capacity can be optimized quickly.

To that end, in this chapter we investigate the closed-form expressions for the capacities and their upper bounds of an important class of DMCs whose channel matrices are invertible positive matrices. An invertible positive matrix is a square matrix whose entries are strictly greater than zero and invertible. There are a number of reasons for using an invertible positive matrix to model many communication channels in real-world settings. First, in most digital communication systems, the transmitter sends a set of transmitted symbols (inputs) and the receiver aims to decode the received signals into one of the transmitted symbols (outputs). Consequently, the channel matrix is a square matrix consisting of the same number of inputs and outputs. Second, since it is physically impossible to design a communication channel without error, the assumption

on the entries in the channel matrix to be strictly greater than zero is reasonable. In the case when an entry is truly zero, it is always possible to approximate the zero with a small positive number. Third, for a $n \times n$ matrix, if the entries are drawn uniformly from a real set (or more precisely in (0,1) and the rows form a valid conditional pmf), then it can be shown that the probability of the matrix being invertible is approaching 1 with increasing n. Thus, invertible matrices are arguably useful to model many communication channels in real-world settings.

Building on the work in [9], our contributions include: (1) we describe an elementary technique based on the theory of convex optimization, to find the closed-form expression for a good upper bound on capacities of discrete memoryless channels with positive invertible channel matrix, and (2) we find the optimality conditions of the channel matrix for which the upper bound is precisely the capacity. We refine the optimality conditions in [9] and provide additional easy-to-use conditions for obtaining closed-form expression for capacities. In particular, the optimality conditions establish a relationship between the singular value and the Gershgorin's disk of the channel matrix. Intuitively, this optimality condition of a channel matrix corresponds to the channel matrix belonging to a subclass of strictly diagonally dominant matrices. Since strictly diagonally dominant matrices represent reliable channels (to be discussed), our results could be useful since most communication systems are designed to achieve a certain level of reliability. Furthermore, our results extend the class of channel matrices, especially the symmetric and weakly symmetric matrices whose channel capacities can be found in closed-form expressions.

2.2 Preliminaries

In this section, we provide definitions together with elementary results that will aid our discussions. In particular, we will discuss (1) the optimality KKT conditions and (2) linear algebra results which we use to derive the closed-form expressions for both the capacity upper bound and exact capacity.

2.2.1 Convex Optimization and KKT Conditions

A DMC is characterized by a random variable $X \in \{x_1, x_2, \ldots, x_m\}$ for the inputs, a random variable $Y \in \{y_1, y_2, \ldots, y_n\}$ for the outputs, and a channel matrix $A \in \mathbf{R}^{m \times n}$. In this chapter, we consider DMCs with equal number of inputs and outputs n, thus $A \in \mathbf{R}^{n \times n}$. The matrix entry A_{ij} represents the conditional probability that given x_i is transmitted, y_j is received. Let $p = (p_1, p_2, \ldots, p_n)^T$ be the input probability mass vector (pmf) of X, where p_i denotes the probability of x_i to be transmitted, then the pmf of Y is $q = (q_1, q_2, \ldots, q_n)^T = A^T p$ and A^T denotes the transpose of A. The mutual information between X and Y is:

$$I(X;Y) = H(Y) - H(Y|X), \tag{2.1}$$

where

$$H(Y) = -\sum_{j=1}^{n} q_j \log q_j \tag{2.2}$$

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_i A_{ij} \log A_{ij}. \tag{2.3}$$

The mutual information function can be written as:

$$I(X;Y) = -\sum_{j=1}^{n} (A^T p)_j \log (A^T p)_j + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i A_{ij} \log A_{ij}, \tag{2.4}$$

where $(A^T p)_j$ denotes the j^{th} component of the vector $q = (A^T p)$. The capacity C associated with a channel matrix A is the theoretical maximum rate at which information can be transmitted over the channel without the error [7], [22], [23]. It is obtained using the optimal pmf p^* such that $I(X;Y)$ is maximized. For a given channel matrix A, $I(X;Y)$ is a concave function of p [3].

Therefore, maximizing $I(X;Y)$ is equivalent to minimizing $-I(X;Y)$, and finding the capacity can be cast as the following convex problem:

Minimize:

$$\sum_{j=1}^{n} (A^T p)_j \log (A^T p)_j - \sum_{i=1}^{n} \sum_{j=1}^{n} p_i A_{ij} \log A_{ij}.$$

Subject to:

$$\begin{cases} p \succeq \mathbf{0} \\ \mathbf{1}^T p = 1. \end{cases}$$

The optimal p^* can be found efficiently using various algorithms such as gradient methods [24], but in a few cases, p^* can be found directly using the Karush-Kuhn-Tucker (KKT) conditions [24]. To explain the KKT conditions, we first state the canonical convex optimization problem below:

Problem **P1**: Minimize: $f(x)$

Subject to:

$$\begin{cases} g_i(x) \leq 0, i = 1, 2, \ldots n, \\ h_j(x) = 0, j = 1, 2, \ldots, m, \end{cases}$$

where $f(x)$, $g_i(x)$ are convex functions and $h_j(x)$ is a linear function.

Define the Lagrangian function as:

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \nu_j h_j(x), \tag{2.5}$$

then the KKT conditions [24] states that, the optimal point x^* must satisfy:

$$\begin{cases} g_i(x^*) \leq 0, \\ h_j(x^*) = 0, \\ \frac{dL(x,\lambda,\nu)}{dx}|_{x=x^*,\lambda=\lambda^*,\nu=\nu^*} = 0, \\ \lambda_i^* g_i(x^*) = 0, \\ \lambda_i^* \geq 0, \end{cases} \tag{2.6}$$

for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

2.2.2 Elementary Linear Algebra Results

We first begin with some definitions and preliminaries that will be used to derive our results.

Definition 2.1. *Let $A \in \mathbf{R}^{n\times n}$ be an invertible channel matrix and $H(A_i) = -\sum_{k=1}^{n} A_{ik} \log A_{ik}$ be the entropy of i^{th} row, define*

$$K_j = -\sum_{i=1}^{n} A_{ji}^{-1} \sum_{k=1}^{n} A_{ik} \log A_{ik} = \sum_{i=1}^{n} A_{ji}^{-1} H(A_i),$$

where A_{ji}^{-1} denotes the entry (j, i) of the inverse matrix A^{-1}. $K_{\max} = \max_j K_j$ and $K_{\min} = \min_j K_j$ are called the maximum and minimum inverse row entropies of A, respectively.

Definition 2.2. *Let $A \in \mathbf{R}^{n\times n}$ be a square matrix. The Gershgorin radius of i^{th} row of A [25] is defined as:*

$$R_i(A) = \sum_{j\neq i}^{n} |A_{ij}|. \tag{2.7}$$

The Gershgorin ratio of i^{th} row of A is defined as:

$$c_i(A) = \frac{A_{ii}}{R_i(A)}, \tag{2.8}$$

and the minimum Gershgorin ratio of A is defined as:

$$c_{\min}(A) = \min_i \frac{A_{ii}}{R_i(A)}. \tag{2.9}$$

We note that since the channel matrix is a stochastic matrix, therefore

$$c_{\min}(A) = \min_i \frac{A_{ii}}{R_i(A)} = \min_i \frac{A_{ii}}{1 - A_{ii}}. \tag{2.10}$$

Definition 2.3. *Let $A \in \mathbf{R}^{n\times n}$ be a square matrix.*

(a) A is called a positive matrix if $A_{ij} > 0 \ \forall \ i, j$.

(b) A is called a strictly diagonally dominant positive matrix [26] if A is a positive matrix and

$$A_{ii} > \sum_{j\neq i} A_{ij}, \forall i, j. \tag{2.11}$$

Lemma 2.1. *Let $A \in \mathbf{R}^{n\times n}$ be a strictly diagonally dominant positive channel matrix then (a) it is invertible; (b) the eigenvalues of A^{-1} are $\frac{1}{\lambda_i} \ \forall \ i$ where λ_i are eigenvalues of A, (c) $A_{ii}^{-1} > 0$ and the largest absolute element in the i^{th} column of A^{-1} is A_{ii}^{-1}, i.e., $A_{ii}^{-1} \geq |A_{ji}^{-1}| \ \forall \ j$.*

Proof. The proof is shown in Appendix 2.6.1. □

Lemma 2.2. *Let $A \in \mathbf{R}^{n\times n}$ be a strictly diagonally dominant positive matrix, then:*

$$c_i(A^{-T}) \geq \frac{c_{\min}(A) - 1}{(n-1)}, \forall i. \tag{2.12}$$

Moreover, for any rows k and l,

$$|A^{-1}_{ki}| + |A^{-1}_{li}| \leq A^{-1}_{ii} \frac{c_{\min}(A)}{c_{\min}(A) - 1}, \forall i. \tag{2.13}$$

Proof. The proof is shown in Appendix 2.6.2. □

Lemma 2.3. *Let $A \in \mathbf{R}^{n \times n}$ be a strictly diagonally dominant positive matrix, then:*

$$\max_{i,j} A^{-1}_{ij} \leq \frac{1}{\sigma_{\min}(A)}, \tag{2.14}$$

where $\max_{i,j} A^{-1}_{ij}$ is the largest entry in A^{-1} and $\sigma_{\min}(A)$ is the minimum singular value of A.

Proof. The proof is shown in Appendix 2.6.3. □

Lemma 2.4. *Let $A \in \mathbf{R}^{n \times n}$ be an invertible channel matrix, then*

$$A^{-1}\boldsymbol{1} = \boldsymbol{1},$$

i.e., the sum of any row of A^{-1} equals to 1. Furthermore, for any probability mass vector x, sum of the vector $y = A^{-T}x$ equal to 1.

Proof. The proof is shown in Appendix 2.6.4. □

2.3 Main Results

Our first main result is an upper bound on the capacity of discrete memoryless channels having invertible positive channel matrices.

Proposition 2.1 (Main Result 1)**.** *Let $A \in \mathbf{R}^{n\times n}$ be an invertible positive channel matrix and*

$$q_j^* = \frac{2^{-K_j}}{\sum_{i=1}^n 2^{-K_i}}, \tag{2.15}$$

$$p' = A^{-T} q^*, \tag{2.16}$$

then the capacity C associated with the channel matrix A is upper bounded by:

$$C \leq -\sum_{j=1}^n q_j^* \log q_j^* + \sum_{i=1}^n \sum_{j=1}^n p_i' A_{ij} \log A_{ij}. \tag{2.17}$$

Proof. Let q be the pmf of the output Y, then $q = A^T p$. Thus,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= -\sum_{j=1}^n q_j \log q_j + \sum_i^n (A^{-T}q)_i \sum_k^n A_{ik} \log A_{ik}. \end{aligned} \tag{2.18}$$

We construct the Lagrangian in (2.5) using $-I(X;Y)$ as the objective function and optimization variable q_j:

$$L(q_j, \lambda_j, \nu_j) = -I(X;Y) - \sum_{j=1}^n q_j \lambda_j + \nu(\sum_{j=1}^n q_j - 1), \tag{2.19}$$

where the constraints $g(x)$ and $h(x)$ in problem **P1** are translated into $-q_j \leq 0$ and $\sum_{j=1}^n q_j = 1$, respectively.

Using the KKT conditions in (2.6), the optimal points q_j^*, λ_j^*, ν^* for all j, must satisfy:

$$q_j^* \geq 0, \tag{2.20}$$

$$\sum_{j=1}^{n} q_j^* = 1, \tag{2.21}$$

$$\nu^* - \lambda_j^* - \frac{dI(X;Y)}{dq_j^*} = 0, \tag{2.22}$$

$$\lambda_j^* \geq 0, \tag{2.23}$$

$$\lambda_j^* q_j^* = 0. \tag{2.24}$$

Since $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$, there exists at least one $p_i > 0$. Since $A_{ij} > 0 \ \forall i, j$, we have:

$$q_j^* = \sum_{i=1}^{n} p_i A_{ij} > 0, \forall j. \tag{2.25}$$

Based on (2.24) and (2.25), we must have $\lambda_j^* = 0, \forall j$. Therefore, all five KKT conditions (2.20-2.24) are reduced to the following two conditions:

$$\sum_{j=1}^{n} q_j^* = 1, \tag{2.26}$$

$$\nu^* - \frac{dI(X;Y)}{dq_j^*} = 0. \tag{2.27}$$

Next,

$$\begin{aligned} \frac{dI(X;Y)}{dq_j} &= \sum_{i=1}^{n} A_{ji}^{-1} \sum_{k=1}^{n} A_{ik} \log A_{ik} - (1 + \log q_j) \\ &= -K_j - (1 + \log q_j). \end{aligned} \tag{2.28}$$

Using (2.27) and (2.28), we have:

$$q_j^* = 2^{-K_j-\nu^*-1}. \tag{2.29}$$

Plugging (2.29) to (2.26), we have:

$$\sum_{j=1}^{n} 2^{-K_j-\nu^*-1} = 1,$$

$$\nu^* = \log \sum_{j=1}^{n} 2^{-K_j-1}.$$

From (2.29),

$$q_j^* = 2^{-K_j-\nu^*-1} = \frac{2^{-K_j}}{2^{\nu^*+1}} = \frac{2^{-K_j}}{\sum_{j=1}^{n} 2^{-K_j}}, \forall j. \tag{2.30}$$

We know that a valid optimal input distribution has to satisfy $0 \leq p_i^* \leq 1$ and $\sum_{i=1}^{n} p_i^* = 1$. If q^* is such that $p' = A^{-T}q^* \succeq 0$ and $(A^{-T}q^*)^T\mathbf{1} = \sum_{i=1}^{n} p_i' = 1$, then $p' = p^*$ is a valid and optimal pmf, and Proposition 2.1 will hold with equality by the KKT conditions. Now, the condition $\sum_{i=1}^{n} p_i' = 1$ holds by Lemma 2.4. However, the condition $0 \leq p_i' \leq 1$ may not satisfy. In this case, maximizing $I(X;Y)$ in terms of q and ignoring this constraint is equivalent to enlarging the feasible region. Since $\max_{x\in A} f(x) \geq \max_{x\in B} f(x)$ if $B \subset A$ for any arbitrary $f(x)$, the upper bound of channel capacity in Proposition 2.1 is achieved by plugging q^* from (2.15) into (2.16) to obtain p', and plugging p' and q^* into (2.4). □

We note that the closed-form expressions for channel capacity are also described in [6] and [8] (Section 3.3). However in both [6] and [8], the sufficient conditions for the closed-form expressions are not fully characterized. We now show another contribution that characterizes the sufficient conditions on the channel matrix A such that its capacity can be written in closed-form expression,

specifically the upper bound in (2.17).

Proposition 2.2 (Main Result 2)**.** *Let $A \in \mathbf{R}^{n\times n}$ be a strictly diagonally dominant positive matrix, if $\forall i$,*

$$c_i(A^{-T}) \geq (n-1)2^{K_{\max}-K_{\min}}, \tag{2.31}$$

then the capacity of the channel having channel matrix A admits a closed-form expression which is exactly the upper bound in Proposition 2.1.

Proof. Based on the discussion of the KKT conditions, it is sufficient to show that if $p^* = A^{-T}q^* \succeq 0$ and $\sum_{i=1}^n p_i^* = (A^{-T}q^*)^T\mathbf{1} = 1$ then C has a closed-form expression. The condition $(A^{-T}q^*)^T\mathbf{1} = 1$ is always true as shown in Lemma 2.4 in the Appendix 2.6.4. Thus, we only need to show that if $c_i(A^{-T}) \geq 2^{K_{\max}-K_{\min}}$, then $p^* = A^{-T}q^* \succeq 0$.

Let $q^*_{\min} = \min_j q_j^*$ and $q^*_{\max} = \max_j q_j^*$, we have:

$$\begin{aligned}
p_i^* &= \sum_j q_j^* A_{ji}^{-1} \\
&= q_i^* A_{ii}^{-1} + \sum_{j\neq i} q_j^* A_{ji}^{-1} \\
&\geq q^*_{\min} A_{ii}^{-1} - (\sum_{j\neq i} q_j^*)(\sum_{j\neq i} |A_{ji}^{-1}|) \qquad (2.32)\\
&\geq q^*_{\min} A_{ii}^{-1} - (n-1)q^*_{\max}(\sum_{j\neq i} |A_{ji}^{-1}|), \qquad (2.33)
\end{aligned}$$

with (2.32) due to $A_{ii}^{-1} > 0$ which follows by Lemma 2.1-(c), (2.33) is due to $q^*_{\max} \geq q_j^* \ \forall\ j$. Now

if we want $p_i^* \geq 0$, $\forall\ i$, from (2.33), it is sufficient to require that, $\forall i$,

$$\begin{aligned} c_i(A^{-T}) = \frac{A_{ii}^{-1}}{\sum_{j\neq i}|A_{ji}^{-1}|} &\geq \frac{(n-1)q_{\max}^*}{q_{\min}^*} \\ &= (n-1)\frac{\frac{2^{-K_{\min}}}{\sum_{j=1}^n 2^{-K_j}}}{\frac{2^{-K_{\max}}}{\sum_{j=1}^n 2^{-K_j}}} \\ &= (n-1)2^{K_{\max}-K_{\min}}, \end{aligned} \tag{2.34}$$

with (2.34) due to (2.30) and $q_{\max}^*$, $q_{\min}^*$ are corresponding to $K_{\min}$, $K_{\max}$, respectively. Thus, Proposition 2.2 is proven.

□

We are now ready to state and prove the third main result that characterizes the sufficient conditions on a channel matrix so that the upper bound in Proposition 2.1 is precisely the capacity.

Proposition 2.3. *Let $A \in \mathbf{R}^{n\times n}$ be a strictly diagonally dominant positive channel matrix and $H_{\max}(A)$ be the maximum row entropy of A. The capacity C is the upper bound in Proposition 2.1 i.e., hold with equality if*

$$\sqrt[V]{\frac{c_{\min}(A)-1}{(n-1)^2}} \geq 2^{\frac{nH_{\max}(A)}{\sigma_{\min}(A)}}, \tag{2.35}$$

where $\sigma_{\min}(A)$ is the minimum singular value of channel matrix A, and

$$V = \frac{c_{\min}(A)}{c_{\min}(A)-1}. \tag{2.36}$$

Proof. From (2.12) in Lemma 2.2 and Proposition 2.2, if we can show that

$$\frac{c_{\min}(A)-1}{(n-1)} \geq (n-1)2^{K_{\max}-K_{\min}}, \tag{2.37}$$

then Proposition 2.3 is proven. Suppose that $K_{\max}$ and $K_{\min}$ are obtained at rows $j = L$ and $j = S$, respectively. We note that from (2.30), $q_{\max} = \max_j q_j$ and $q_{\min} = \min_j q_j$ correspond to $K_{\min}$ and $K_{\max}$, respectively. Thus, from the Definition **1**, we have:

$$\begin{aligned}
K_{\max}-K_{\min} &= \sum_{i=1}^{n} A_{Li}^{-1}H(A_i) - \sum_{i=1}^{n} A_{Si}^{-1}H(A_i) \\
&\leq |\sum_{i=1}^{n} A_{Li}^{-1}H(A_i)| + |\sum_{i=1}^{n} A_{Si}^{-1}H(A_i)| && (2.38)\\
&\leq \sum_{i=1}^{n} |A_{Li}^{-1}||H(A_i)| + \sum_{i=1}^{n} |A_{Si}^{-1}||H(A_i)| && (2.39)\\
&\leq H_{\max}(A)\sum_{i=1}^{n}(|A_{Li}^{-1}| + |A_{Si}^{-1}|) && (2.40)\\
&\leq H_{\max}(A)\sum_{i=1}^{n} A_{ii}^{-1}\frac{c_{\min}(A)}{c_{\min}(A)-1} && (2.41)\\
&\leq nH_{\max}(A)(\max_{i,j} A_{ij}^{-1})\frac{c_{\min}(A)}{c_{\min}(A)-1} && (2.42)\\
&\leq \frac{nH_{\max}(A)V}{\sigma_{\min}(A)}, && (2.43)
\end{aligned}$$

where (2.38) due to the property of absolute value function, (2.39) due to Schwarz inequality, (2.40) due to $H_{\max}(A)$ is the maximum row entropy of A, (2.41) due to (2.13), (2.42) due to $\max_{i,j} A_{ij}^{-1}$ is the largest entry in A^{-1} and (2.43) is due to Lemma 2.3. Thus,

$$(n-1)2^{\frac{nH_{\max}(A)V}{\sigma_{\min}(A)}} \geq (n-1)2^{K_{\max}-K_{\min}}. \tag{2.44}$$

From (2.37) and (2.44), if

$$\frac{c_{\min}(A)-1}{(n-1)} \geq (n-1)2^{\frac{nH_{\max}(A)V}{\sigma_{\min}(A)}}, \tag{2.45}$$

then the capacity C is the upper bound in Proposition 2.1. (2.45) is equivalent to (2.35). Thus Proposition 2.3 is proven. □

We note that the condition in Proposition 2.3 is easier to verify than the condition in Proposition 2.2 since it can be performed without requiring matrix inverse. Other easy-to-use versions of checking condition are stated in Proposition 2.4 and Corollary 2.1.

Proposition 2.4. *The capacity C is the upper bound in Proposition 2.1 if*

$$\frac{c_{\min}(A)-1}{(n-1)^2} \geq 2^{\frac{2n\log n}{\sigma_{\min}(A)}}. \tag{2.46}$$

Proof. Similar to Proposition 2.3,

$$\begin{aligned} K_{\max}-K_{\min} &\leq H_{\max}(A)\sum_{i=1}^{n}(|A_{Li}^{-1}|+|A_{Si}^{-1}|) && (2.47)\\ &\leq H_{\max}(A)n(2\max_{i,j}A_{ij}^{-1}) && (2.48)\\ &\leq \frac{2n\log n}{\sigma_{\min}(A)}, && (2.49) \end{aligned}$$

with (2.47) is identical to (2.40), (2.48) is due to $\max_{i,j}A_{ij}^{-1}$ is the largest entry in A^{-1}, (2.49) due to $H_{\max}(A)\leq \log n$ and Lemma 2.3. Thus, by changing $\frac{nH_{\max}(A)V}{\sigma_{\min}(A)}$ in (2.45) by $\frac{2n\log n}{\sigma_{\min}(A)}$, the Proposition 2.4 is proven. □

A direct result of Proposition 2.3 without using singular value is shown in Corollary 2.1.

Corollary 2.1. *The capacity C is the upper bound in Proposition 2.1 if*

$$\sqrt[V]{\frac{c_{\min}(A)-1}{(n-1)^2}} \geq 2^{\frac{nH^*_{\max}(A)}{\sigma^*}}, \tag{2.50}$$

where,

$$V = \frac{c_{\min}(A)}{c_{\min}(A)-1}, \tag{2.51}$$

$$\sigma^* = \frac{c_{\min}(A)-n/2}{c_{\min}(A)+1}, \tag{2.52}$$

$$H^*_{\max}(A) = \log(c_{\min}(A)+1) + \frac{\log(n-1) - c_{\min}(A)\log c_{\min}(A)}{c_{\min}(A)+1}. \tag{2.53}$$

Proof. We will construct the lower bound for $\sigma_{\min}(A)$ and the upper bound for $H_{\max}(A)$. From Lemma 2.5 in Appendix 2.6.5

$$\sigma_{\min}(A) \geq \frac{c_{\min}(A)-n/2}{c_{\min}(A)+1} = \sigma^*, \tag{2.54}$$

and

$$\begin{aligned} H_{\max}(A) &\leq \log(c_{\min}(A)+1) + \frac{\log(n-1) - c_{\min}(A)\log c_{\min}(A)}{c_{\min}(A)+1} \\ &= H^*_{\max}(A). \end{aligned} \tag{2.55}$$

Therefore

$$\frac{nH_{\max}(A)V}{\sigma_{\min}(A)} \leq \frac{nH^*_{\max}(A)}{\sigma^*}. \tag{2.56}$$

Thus, by changing $\frac{nH_{\max}(A)}{\sigma_{\min}(A)}$ in (2.35) by $\frac{nH^*_{\max}(A)}{\sigma^*}$, the Corollary 2.1 is proven.

We note that, when $c_{\min}(A)$ is relatively larger than the size of matrix n, the lower bound of $\sigma_{\min}(A)$ goes to 1. We also note that (2.50) can be checked efficiently without requiring both $H_{\max}(A)$ and $\sigma_{\min}(A)$ at the expense of a looser upper bound as compare to (2.35). □

2.4 Examples and Numerical Results

2.4.1 Example 1: Cooperative Relay-MISO Channels

In this example, we investigate the channel capacity for a class of channels named Relay-MISO (Relay - Multiple Input Single Output). Relay-MISO channel [27] can be constructed by the combination of a relay channel [28] [29] and a Multiple Input Single Output channel, as illustrated in Fig. 2.1.

In a Relay-MISO channel, n senders want to transmit data to a same receiver via n relay base station nodes. The uplink of these senders using wireless links that are prone to transmission errors. Each sender can transmit bit "0" or "1" with the probability of bit flipping is α, $0 \leq \alpha \leq 1$. For a simplicity, suppose that n relay channels have the same error probability α. Next, all of the relay base station nodes will relay the signal by a reliable channel such as optical fiber cable to a same receiver. The receiver adds all the relay signals (symbols) to produce a single output symbol.

It can be shown that the channel matrix of this Relay-MISO channel [27] is an invertible matrix of size $(n+1) \times (n+1)$ whose A_{ij} can be computed as:

$$A_{ij} = \sum_{s=\max(i-j,0)}^{s=\min(n+1-j,i-1)} \binom{j-i+s}{n+1-i} \binom{s}{i-1} \alpha^{j-i+2s} (1-\alpha)^{n-(j-i+2s)}.$$

We note that this Relay-MISO channel matrix is invertible and the inverse matrix has the closed-form expression which is characterized in [27]. For example, the channel matrix of a Relay-MISO

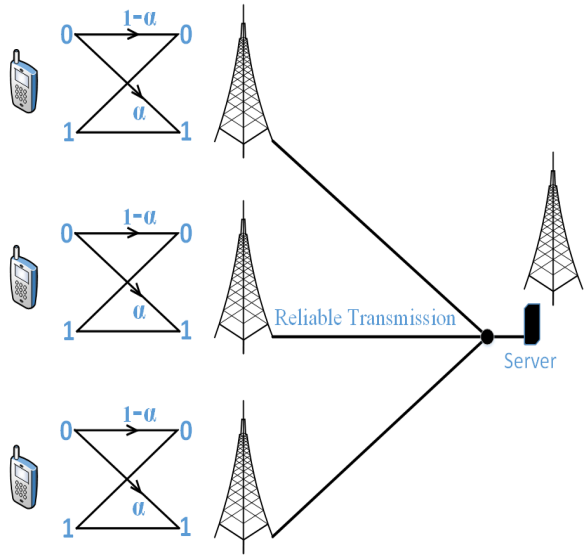


Figure 2.1: Relay-MISO channel

channel with $n = 3$ is given as follows:

$$\begin{bmatrix} (1-\alpha)^3 & 3(1-\alpha)^2\alpha & 3(1-\alpha)\alpha^2 & \alpha^3 \\ \alpha(1-\alpha)^2 & 2\alpha^2(1-\alpha)+(1-\alpha)^3 & 2(1-\alpha)^2\alpha+\alpha^3 & (1-\alpha)\alpha^2 \\ (1-\alpha)\alpha^2 & 2(1-\alpha)^2\alpha+\alpha^3 & 2\alpha^2(1-\alpha)+(1-\alpha)^3 & \alpha(1-\alpha)^2 \\ \alpha^3 & 3(1-\alpha)\alpha^2 & 3(1-\alpha)^2\alpha & (1-\alpha)^3 \end{bmatrix},$$

where $0 \leq \alpha \leq 1$. We note that this channel matrix is strictly diagonally dominant matrix when α is close to 0 or α is close to 1. In addition, for α values that are close to 0 or 1, it can be shown that channel matrix A satisfies the conditions in Proposition 2.3. Thus, the channel capacity admits a closed-form expression in Proposition 2.1. For other values of α, e.g. closer to 0.5, the optimality conditions in Proposition 2.3 no longer holds. In this case, Proposition 2.1 can still be used as a good upper bound on the capacity.

We show that our upper bound is tighter than existing upper bounds. In particular, Fig.

2.2 shows the actual capacity and the known upper bounds as functions of parameter α for Relay-MISO channels having $n = 3$. The green curve depicts the actual capacity computed using convex optimization algorithm. The red curve is constructed using our closed-form expression in Proposition 2.1, and the blue dotted curve is the constructed using the well-known upper bound result of channel capacity in [13], [30]. Specifically, this upper bound is:

$$C \leq \log(\sum_{j=1}^{n} \max_{i} A_{ij}). \tag{2.57}$$

Finally, the red dotted curve shows another well-known upper bound by Arimoto [5] which is:

$$C \leq \log(n) + \max_{j}[\sum_{i=1}^{n} A_{ji} \log(\frac{A_{ji}}{\sum_{k=1}^{n} A_{ki}})]. \tag{2.58}$$

We note that the second term is negative.

Fig. 2.2 shows that our closed-form upper bound is precisely the capacity (the red and green graphs are overlapped) when α values are close to 0 or 1 as predicted by the optimality conditions in Proposition 2.3. On the other hand, when α values are closer to 0.5, our optimality conditions no longer hold. In this case, we can only determine the upper bound. However, it is interesting to note that our upper bound in this case is tighter than both the Boy-Chiang [13] and Arimoto [5] upper bounds.

2.4.2 Example 2: Symmetric and Weakly Symmetric Channels

Our results confirm the capacity of the well known symmetric and weakly symmetric channel matrices. In particular, when the channel matrix is symmetric and positive definite, all our results are applicable. Indeed, since the channel matrix is symmetric and positive definite, the

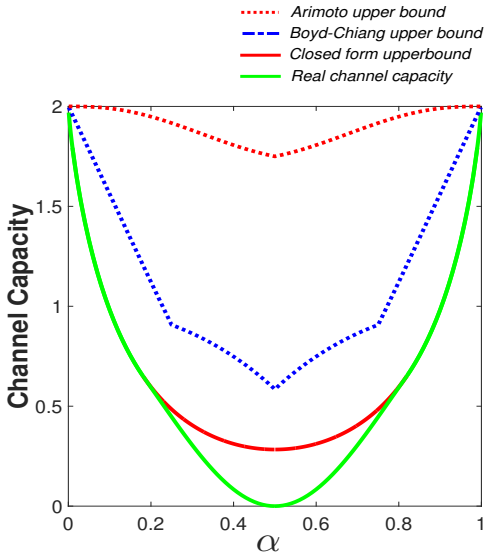


Figure 2.2: Channel capacity and various upper bounds as functions of α

inverse channel matrix exists and also is symmetric. From Definition 2.1, all values of K_j is the same since they are the same sum of permutation entries. Therefore, from Proposition 2.1, the optimal output probability mass vector

$$q_j^* = \frac{2^{-K_j}}{\sum_{i=1}^{n} 2^{-K_i}} \tag{2.59}$$

are equal each other for all j. As a result, the input probability mass function $p^* = A^{-T}q^*$ is the uniform distribution, and the channel capacity is upper bounded by:

$$\begin{aligned} C &\leq -\sum_{j=1}^{n} q_j^* \log q_j^* + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i^* A_{ij} \log A_{ij} && (2.60) \\ &= \log n - H(A_{row}). && (2.61) \end{aligned}$$

Interestingly, our result also shows the capacities of many channels that are *not* weakly sym-

metric, but admits the closed-form formula of weakly symmetric channels. In particular, consider a channel matrix called semi-weakly symmetric whose all rows are permutations of each other, but the sum of entries in each column might not be the same. Furthermore, if the optimal condition is satisfied (Proposition 2.3), then the channel has closed-form capacity which is identical to the capacity of a symmetric and weakly symmetric channel:

$$C = \log n - H(A_{row}). \tag{2.62}$$

Note that every row of a quasi-symmetric matrix is a permutation of the first row [31]. Thus, a quasi-symmetric matrix is an example of a semi-weakly symmetric matrix. For example, the following channel matrix:

$$A = \begin{bmatrix} 0.93 & 0.04 & 0.03 \\ 0.04 & 0.93 & 0.03 \\ 0.04 & 0.03 & 0.93 \end{bmatrix}$$

is not a weakly symmetric channel even though its rows are permutations of each other since the column sums are different. However, this channel matrix satisfies Proposition 2.3 and Corollary 2.1 since $n = 3$, $\sigma_{\min}(A) = 0.88916$, $\sigma^* = 0.825$, $H_{\max}(A) = 0.43489$, $H^*_{\max}(A) = 0.43592$ and $c_{\min}(A) = 13.286$. Thus, it has closed-form formula for capacity, and can be easily shown to be $C = \log 3 - H(0.93, 0.04, 0.03) = 1.1501$. The optimal output and input probability mass vectors can be shown to be:

$$q^T = \begin{bmatrix} 0.33333 & 0.33333 & 0.33333 \end{bmatrix},$$

$$p^T = \begin{bmatrix} 0.32959 & 0.33337 & 0.33704 \end{bmatrix},$$

respectively.

The following channel matrix is another example of semi-weakly symmetric matrix whose

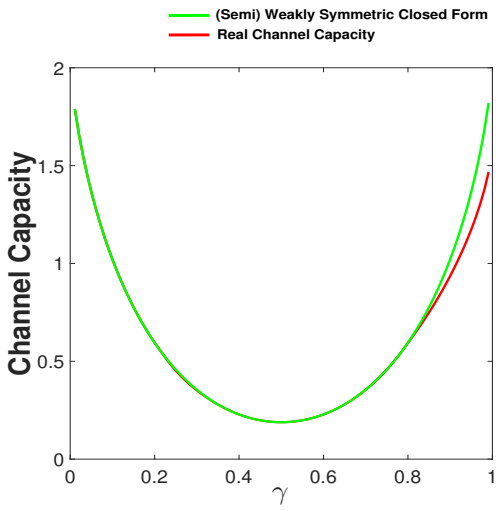


Figure 2.3: Channel capacity of (semi) weakly symmetric channel as a function of γ

entries are controlled by a parameter γ in the range of $(0, 1)$ and given by the following form:

$$\begin{bmatrix} (1-\gamma)^3 & 3(1-\gamma)^2\gamma & 3(1-\gamma)\gamma^2 & \gamma^3 \\ 3(1-\gamma)^2\gamma & (1-\gamma)^3 & \gamma^3 & 3(1-\gamma)\gamma^2 \\ \gamma^3 & 3(1-\gamma)\gamma^2 & (1-\gamma)^3 & 3(1-\gamma)^2\gamma \\ \gamma^3 & 3(1-\gamma)\gamma^2 & 3(1-\gamma)^2\gamma & (1-\gamma)^3 \end{bmatrix}.$$

Fig. 2.3 shows the capacity upper bound of the semi-weakly symmetric channel and the actual channel capacity as function of γ. Theoretically, the conditions in Proposition 2.3 and Proposition 2.4 can be shown to hold for $\gamma \leq 0.02$. However, for much values of γ, the upper bound is identical to the actual channel capacity which can be numerically determined using CVX [14]. This happens because these conditions are sufficient but not necessary.

2.4.3 Example 3: Unreliable Channels

We now consider an unreliable channel whose channel matrix is:

$$A = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.7 & 0.1 & 0.2 \\ 0.5 & 0.05 & 0.45 \end{bmatrix}.$$

In this case, our optimality conditions do not satisfy, and the Arimoto upper bound is tightest (0.17083) as compared to our upper bound (0.19282) and Boyd-Chiang upper bound (0.848).

2.4.4 Example 4: Bounds as Function of Channel Reliability

Since we know that our proposed bounds are tight if the channel is reliable, we want to examine quantitatively how channel reliability affects various bounds. In this example, we consider a special class of channel whose channel matrix entries are controlled by a reliability parameter β for $0 \leq \beta \leq 1$ as shown below:

$$A = \begin{bmatrix} 1-\beta & 0.3\beta & 0.4\beta & 0.3\beta \\ 0.4\beta & 1-\beta & 0.3\beta & 0.3\beta \\ 0.5\beta & 0.4\beta & 1-\beta & 0.1\beta \\ 0.1\beta & 0.2\beta & 0.7\beta & 1-\beta \end{bmatrix}.$$

When β is small, the channel tends to be reliable and when β is large, the channel tends to be unreliable. Fig. 2.4 shows various upper bounds as a function of β together with the actual capacity. The actual channel capacities for various β are numerically computed using a convex optimization algorithm [14]. As seen, our closed-form upper bound expression for capacity (red

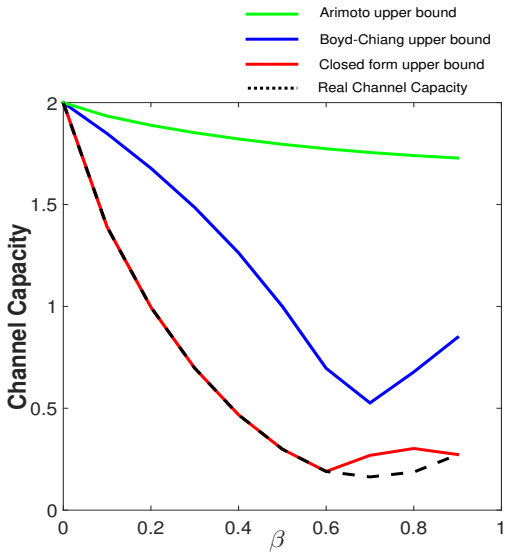


Figure 2.4: Channel capacity and various upper bounds functions of β

curve) from Proposition 2.1 is much closer to the actual capacity (black dash curve) than other bounds for most values of β. When β is small ($\beta \leq 0.6$) or channel is reliable, the closed-form upper bound is precise the real channel capacity, and we can verify that the optimal conditions in Proposition 2.3 holds. When the channel becomes unreliable, i.e., $\beta \geq 0.6$, our upper bound is no longer tight, however, it is still the tightest among all the existing upper bounds. We note that when the β is small, the channel matrix becomes a nearly diagonally dominant matrix, and our upper bound is tightest.

2.5 Conclusion

In this chapter, we describe an elementary technique based on Karush-Kuhn-Tucker (KKT) conditions to obtain (1) a good upper bound of a discrete memoryless channel having an invertible positive channel matrix and (2) a closed-form expression for the capacity if the channel matrix

satisfies certain conditions related to its singular value and its Gershgorin's disk. We provide a number of channels where the proposed upper bound becomes precisely the capacity. We also demonstrate that our proposed bounds are tighter than other existing bounds for these channels.

2.6 Appendix

2.6.1 Proof of Lemma 2.1

For claim (a), since the channel matrix is strictly diagonally dominant, using Gershgorin circle theorem [25] that for any eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, we must have:

$$\lambda_i \geq A_{ii} - \sum_{j \neq i} |A_{ij}| > 0.$$

Thus, $det(A) = \lambda_1 \lambda_2 \ldots \lambda_n > 0$. Therefore, A is invertible.

Claim (b) is a well-known algebra result [32].

For claim (c), due to $AA^{-1} = I$ and $A_{ij} > 0 \ \forall \ i, j$, therefore, $\forall \ j$ exists at least i such that $A_{ij}^{-1} \neq 0$. Therefore the largest absolute entry in each column $\neq 0$. Claim (c) can be obtained by contradiction. Suppose that the largest absolute entry in j^{th} column of A^{-1} is A_{ij}^{-1} in i^{th} row, that said $|A_{ij}^{-1}| \geq |A_{kj}^{-1}| \ \forall \ k$. We suppose that $A_{ij}^{-1} < 0$. Thus:

$$\begin{aligned} \sum_{k=1}^{n} A_{ik} A_{kj}^{-1} &\leq -A_{ii}|A_{ij}^{-1}| + \sum_{k=1, k \neq i}^{n} A_{ik}|A_{ij}^{-1}| && (2.63) \\ &= (-A_{ii} + \sum_{k=1, k \neq i}^{n} A_{ik})|A_{ij}^{-1}| \\ &< 0, && (2.64) \end{aligned}$$

which contradicts with $\sum_{k=1}^{n} A_{ik}A_{kj}^{-1} = I_{ij} \geq 0$. Thus, the largest absolute value in each column of A^{-1} is positive. That said in j^{th} column, if $|A_{ij}^{-1}| \geq |A_{kj}^{-1}| \; \forall \; k$, then $A_{ij}^{-1} > 0$.

Now, suppose that the largest absolute element in j^{th} column of A^{-1}, is A_{ij}^{-1} with $i \neq j$ and $A_{ij}^{-1} > 0$. Then:

$$\begin{aligned} 0 &= \sum_{k=1}^{n} A_{ik}A_{kj}^{-1} \\ &\geq A_{ii}|A_{ij}^{-1}| - \sum_{k=1,k\neq i}^{n} A_{ik}|A_{ij}^{-1}| \qquad (2.65) \\ &= (A_{ii} - \sum_{k=1,k\neq i}^{n} A_{ik})A_{ij}^{-1} \\ &> 0, \qquad (2.66) \end{aligned}$$

with (2.65) due to A_{ij}^{-1} is the largest absolute element in j^{th} column and (2.66) due to A is strictly diagonally dominant matrix. This is a contradiction. Therefore, the largest absolute entry in j^{th} column of A^{-1} should be A_{jj}^{-1} and $A_{jj}^{-1} > 0$.

2.6.2 Proof of Lemma 2.2

First, let's show that the second largest absolute value in each column of A^{-1} is a negative entry by contradiction method. Suppose that the second largest absolute value in j^{th} column of A^{-1} is

positive and in k^{th} row ($k \neq j$), $A_{kj}^{-1} \geq 0$. Consider,

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} A_{ki} A_{ij}^{-1} \\
&\geq A_{kj} A_{jj}^{-1} + A_{kk} A_{kj}^{-1} - |\sum_{i=1, i\neq k; i\neq j}^{n} A_{ki} A_{ij}^{-1}| && (2.67) \\
&\geq A_{kj} A_{jj}^{-1} + A_{kk} A_{kj}^{-1} - \sum_{i=1, i\neq k; i\neq j}^{n} |A_{ki} A_{ij}^{-1}| && (2.68) \\
&\geq A_{kj} A_{jj}^{-1} + A_{kk} A_{kj}^{-1} - \sum_{i=1, i\neq k; i\neq j}^{n} A_{ki} |A_{ij}^{-1}| && (2.69) \\
&\geq A_{kj} A_{jj}^{-1} + A_{kk} A_{kj}^{-1} - \sum_{i=1, i\neq k; i\neq j}^{n} A_{ki} |A_{kj}^{-1}| && (2.70) \\
&= A_{kj} A_{jj}^{-1} + A_{kj}^{-1} (A_{kk} - \sum_{i=1, i\neq k; i\neq j}^{n} A_{ki}) && (2.71) \\
&> 0, && (2.72)
\end{aligned}
$$

with (2.67) due to the fact that $C \geq -|C| \; \forall \; C$, (2.68) due to the triangle inequality, (2.69) due to A_{ki} is positive, (2.70) due to A_{kj}^{-1} is the second largest absolute value in j^{th} column of A^{-1}, (2.71) due to the assumption that $A_{kj}^{-1} \geq 0$ and (2.72) due to (2.11) such that $A_{kk} \geq \sum_{i=1, i\neq k}^{n} A_{ki} \geq \sum_{i=1, i\neq k; i\neq j}^{n} A_{ki}$. Thus, the second largest absolute value in column of A^{-1} is negative ($A_{kj}^{-1} < 0$).

Due to Lemma 2.1 part (c), A_{jj}^{-1} is the largest absolute value entry and $A_{jj}^{-1} > 0$. Similarly,

$$\begin{aligned}
0 &= \sum_{i=1}^{n} A_{ki}A_{ij}^{-1} \\
&\leq A_{kj}A_{jj}^{-1} + A_{kk}A_{kj}^{-1} + |\sum_{i=1,i\neq k;i\neq j}^{n} A_{ki}A_{ij}^{-1}| && (2.73)\\
&\leq A_{kj}A_{jj}^{-1} + A_{kk}A_{kj}^{-1} + \sum_{i=1,i\neq k;i\neq j}^{n} |A_{ki}A_{ij}^{-1}| && (2.74)\\
&\leq A_{kj}A_{jj}^{-1} + A_{kk}A_{kj}^{-1} + \sum_{i=1,i\neq k;i\neq j}^{n} A_{ki}|A_{ij}^{-1}| && (2.75)\\
&\leq A_{kj}A_{jj}^{-1} - A_{kk}|A_{kj}^{-1}| + \sum_{i=1,i\neq k;i\neq j}^{n} A_{ki}|A_{kj}^{-1}|, && (2.76)
\end{aligned}$$

with (2.73) due to the fact that $C \leq |C| \ \forall \ C$, (2.74) due to the triangle inequality, (2.75) due to $A_{ki} \geq 0, \ \forall \ i$ and (2.76) due to $A_{kj}^{-1} < 0$ and A_{kj}^{-1} is the second largest absolute value in j^{th} column. Hence,

$$\begin{aligned}
A_{kj}A_{jj}^{-1} &\geq A_{kk}|A_{kj}^{-1}| - \sum_{i=1,i\neq k;i\neq j}^{n} A_{ki}|A_{kj}^{-1}| \\
A_{jj}^{-1} &\geq \frac{|A_{kj}^{-1}|(A_{kk} - \sum_{i=1,i\neq k;i\neq j}^{n} A_{ki})}{A_{kj}} \\
A_{jj}^{-1} &\geq |A_{kj}^{-1}|\frac{A_{kk} - \frac{A_{kk}}{c_{\min}(A)}}{\frac{A_{kk}}{c_{\min}(A)}} && (2.77)\\
A_{jj}^{-1} &\geq |A_{kj}^{-1}|[c_{\min}(A) - 1], \forall j, && (2.78)
\end{aligned}$$

with (2.77) due to Definition 2.2 and (2.9) such that $\frac{A_{kk}}{c_{\min}(A)} \geq \sum_{i=1,i\neq k}^{n} A_{ki} \geq \sum_{i=1,i\neq k,i\neq j}^{n} A_{ki}$. Thus, we have:

$$c_j(A^{-T}) = \frac{A_{jj}^{-1}}{\sum_{k\neq j} |A_{kj}^{-1}|} \geq \frac{c_{\min}(A)-1}{n-1}. \tag{2.79}$$

Thus, (2.12) is proven.

Next, we note that from (2.78)

$$\frac{A_{jj}^{-1}}{c_{\min}(A)-1} \geq |A_{kj}^{-1}|, \forall k. \tag{2.80}$$

Moreover, from Lemma 2.1, $A_{jj}^{-1} \geq 0$ and is the largest entry in j^{th} row. Thus, for an arbitrary L and S,

$$\begin{aligned} |A_{Lj}^{-1}| + |A_{Sj}^{-1}| &\leq A_{jj}^{-1} + \frac{A_{jj}^{-1}}{c_{\min}(A)-1} \\ &= A_{jj}^{-1}\frac{c_{\min}(A)}{c_{\min}(A)-1}, \forall j. \end{aligned} \tag{2.81}$$

Thus, (2.13) is proven.

2.6.3 Proof of Lemma 2.3

Consider the matrix $B = A^{-1}A^{-T}$, B is symmetric, all its eigenvalues are real and satisfy the Rayleigh quotient [33]. Let λ_B^{max} be the maximum eigenvalue of B then from [33]

$$R(B,x) = \frac{x^*Bx}{x^*x} \leq \lambda_B^{max}. \tag{2.82}$$

Consider the unit vector $e = [0, \ldots, 1, \ldots, 0]^T$ with entry "1" is in the i^{th} column. Let $x = e$ in (2.82), we have:

$$B_{ii} \leq \lambda_B^{max}. \tag{2.83}$$

Thus,

$$\begin{aligned} \lambda_B^{max} &\geq B_{ii} \\ &= \sum_{j=1}^{n} A_{ij}^{-1} A_{ij}^{-1} \\ &\geq (A_{ii}^{-1})^2. \end{aligned} \tag{2.84}$$

Now since B is a symmetric matrix $\lambda_B^{max} = \sigma_{\max}(B)$ [32]. However, from [32], $\sigma_{\max}(B) = \sigma_{\max}(A^{-1}A^{-T}) = \sigma_{\max}^2 A^{-1}$ and $\sigma_{\max} A^{-1} = \dfrac{1}{\sigma_{\min}(A)}$. Thus:

$$\frac{1}{\sigma_{\min}(A)} \geq A_{ii}^{-1}. \tag{2.85}$$

From Lemma 2.1-(c), the largest entry in A^{-1} must be a diagonal element, thus

$$\max_{i,j} A_{ij}^{-1} \leq \frac{1}{\sigma_{\min}(A)}.$$

2.6.4 Proof of Lemma 2.4

For the first claim, since A is a stochastic matrix,

$$A\mathbf{1} = \mathbf{1}.$$

Left multiply both sides by A^{-1} results in $\mathbf{1} = A^{-1}\mathbf{1}$. For the second claim, left multiplying $y = A^{-T}x$ by $\mathbf{1}^T$, we have:

$$\mathbf{1}^T y = \mathbf{1}^T A^{-T} x = x^T A^{-1}\mathbf{1} = x^T \mathbf{1} = 1,$$

where we use $A^{-1}\mathbf{1} = \mathbf{1}$ in the previous claim.

Thus, we have $\sum_{i=1}^{n} p_i^* = 1$ since from (2.30), q^* is a probability mass vector.

2.6.5 Proof of Corollary 2.1

Lemma 2.5. *Lower bound of $\sigma_{\min}(A)$ and upper bound of $H_{\max}(A)$ are σ^* and $H^*_{\max}(A)$, respectively*

$$\sigma_{\min}(A) \geq \sigma^* = \frac{c_{\min}(A) - n/2}{c_{\min}(A) + 1}, \tag{2.86}$$

and

$$H_{\max}(A) \leq H^*_{\max}(A), \tag{2.87}$$

where

$$H^*_{\max}(A) = \log(c_{\min}(A) + 1) + \frac{\log(n-1) - c_{\min}(A)\log c_{\min}(A)}{c_{\min}(A)+1}. \tag{2.88}$$

Proof. Due to the channel matrix is a strictly diagonally dominant positive matrix. Thus, we have

$$A_{kk} \geq \frac{c_{\min}(A)}{c_{\min}(A) + 1}, \tag{2.89}$$

$$R_k(A) = 1 - A_{kk} \leq 1 - \frac{c_{\min}(A)}{c_{\min}(A) + 1} = \frac{1}{c_{\min}(A) + 1}, \tag{2.90}$$

$$C_k(A) = \sum_{j=1, j\neq k}^{j=n} A_{jk} \leq \sum_{j=1, j\neq k}^{j=n} R_j(A) \leq \frac{n-1}{c_{\min}(A)+1}, \forall k, \tag{2.91}$$

with (6.65) due to (2.10), (2.90) due to (6.65), (2.91) due to the fact that $\forall\ j \neq k$, $A_{jk} \leq \sum_{j\neq k} A_{jk} = R_j(A)$ and each $R_j(A) \leq \frac{1}{c_{\min}(A)+1}$ which is proven in (6.65). Now, we are ready to establish the upper bound of $H_{\max}(A)$ and the lower bound of $\sigma_{\min}(A)$, respectively.

- Suppose that $H_{\max}(A)$ achieves at k^{th} row, then

$$\begin{aligned}
H_{\max}(A) &= -(\sum_{i=1}^{n} A_{ki} \log A_{ki}) \\
&= -(A_{kk} \log A_{kk} + \sum_{i=1, i\neq k}^{n} A_{ki} \log A_{ki}) \\
&= -A_{kk} \log A_{kk} \\
&- (1-A_{kk}) \sum_{i=1, i\neq k}^{n} \frac{A_{ki}}{1-A_{kk}} (\log \frac{A_{ki}}{1-A_{kk}} + \log(1-A_{kk})) \\
&= -A_{kk} \log A_{kk} \\
&- (1-A_{kk}) \sum_{i=1, i\neq k}^{n} \frac{A_{ki}}{1-A_{kk}} \log \frac{A_{ki}}{1-A_{kk}} \\
&- (1-A_{kk}) \log(1-A_{kk}) \\
&\leq -A_{kk} \log A_{kk} + (1-A_{kk}) \log(n-1) \\
&- (1-A_{kk}) \log(1-A_{kk}) \qquad (2.92) \\
&= -(A_{kk} \log A_{kk} + (1-A_{kk}) \log(\frac{1-A_{kk}}{n-1})) \\
&\leq -(\frac{c_{\min}(A)}{c_{\min}(A)+1} \log \frac{c_{\min}(A)}{c_{\min}(A)+1} \\
&+ (1-\frac{c_{\min}(A)}{c_{\min}(A)+1}) \log \frac{1-\frac{c_{\min}(A)}{c_{\min}(A)+1}}{n-1}) \qquad (2.93) \\
&= \log(c_{\min}(A)+1) + \frac{\log(n-1) - c_{\min}(A) \log c_{\min}(A)}{c_{\min}(A)+1},
\end{aligned}$$

with (2.92) is due to $-\sum_{i=1, i\neq k}^{n} \frac{A_{ki}}{1-A_{kk}} \log \frac{A_{ki}}{1-A_{kk}}$ is the entropy of $n-1$ elements which is bounded by $\log(n-1)$. For (2.93), first we show that $f(x) = -(x \log x + (1-x) \log(\frac{1-x}{n-1}))$ is

monotonically decreasing function for $\frac{x}{1-x} \geq n-1$. Indeed,

$$\begin{aligned} \frac{d(f(x))}{d(x)} &= \log x - \log(1-x) - \log(n-1) \\ &= -(\log \frac{x}{1-x} - \log(n-1)). \end{aligned}$$

Thus, if $\frac{x}{1-x} \geq n-1$ then $\frac{d(f(x))}{d(x)} \leq 0$. However, from (6.65),

$$\frac{A_{kk}}{1-A_{kk}} \geq \frac{\frac{c_{\min}(A)}{c_{\min}(A)+1}}{1-\frac{c_{\min}(A)}{c_{\min}(A)+1}} = c_{\min}(A). \tag{2.94}$$

From (2.50)

$$c_{\min}(A) \geq 1 + (n-1)^2 2^{\frac{nH^*_{\max}(A)}{\sigma^*}} \geq 1 + (n-1)^2 > n-1, \tag{2.95}$$

due to $\frac{nH^*_{\max}(A)}{\sigma^*} \geq 0$ and $n \geq 2$. Thus, $\frac{A_{kk}}{1-A_{kk}} > n-1$. From (2.94) and (2.95), $f(x)$ is decreasing function and (2.93) is constructed by plugging the lower bound of A_{kk} in (6.65).

- Secondly, the lower bound of $\sigma_{\min}(A)$ can be found in [34] (Theorem 3)

$$\sigma_{\min}(A) \geq \min_{1 \leq k \leq n} |A_{kk}| - \frac{1}{2}(R_k(A) + C_k(A)), \tag{2.96}$$

or in [35] (Theorem 0)

$$\sigma_{\min}(A) \geq \min_{1 \leq k \leq n} \frac{1}{2}(\{4|A_{kk}|^2 + (R_k(A) - C_k(A))^2\}^{1/2} - [R_k(A) + C_k(A)]), \tag{2.97}$$

with $R_k(A) = \sum_{j=1, j \neq k}^{j=n} |A_{kj}|$ and $C_k(A) = \sum_{j=1, j \neq k}^{j=n} |A_{jk}|$, respectively. Thus, if we use the lower

bound established in (2.97),

$$\begin{aligned}\sigma_{\min}(A) &\geq \frac{1}{2}(\{4[\frac{c_{\min}(A)}{c_{\min}(A)+1}]^2\}^{1/2} \\ &- [\frac{1}{c_{\min}(A)+1}+\frac{n-1}{c_{\min}(A)+1}]) \\ &= \frac{c_{\min}(A)-n/2}{c_{\min}(A)+1}=\sigma^*,\end{aligned} \tag{2.98}$$

with (2.98) due to (6.65), (2.90), (2.91) and the fact that $\{R_k(A)-C_k(A)\}^2 \geq 0$.

A similar lower bound can be constructed using (2.96)

$$\begin{aligned}\sigma_{\min}(A) &\geq \frac{c_{\min}(A)}{c_{\min}(A)+1} \\ &- \frac{1}{2}(\frac{1}{c_{\min}(A)+1}+\frac{n-1}{c_{\min}(A)+1}) \\ &= \frac{c_{\min}(A)-n/2}{c_{\min}(A)+1}=\sigma^*,\end{aligned} \tag{2.99}$$

with (2.99) due to (6.65), (2.90) and (2.91). As seen, both our approaches yield a same lower bound of $\sigma_{\min}(A)$. However, (2.97) is tighter than (2.96) due to $\{R_k(A)-C_k(A)\}^2$. □

Chapter 3: Binary Quantizer Designing For Maximizing Mutual Information

3.1 Introduction

Quantization techniques play a vital role in signal processing, communication, and information theory. A classical quantization technique maps a given real number to an element in a given finite discrete set that minimizes/maximizes a certain objective. In compression, quantization is often used to minimize the distortion (e.g. mean square error (MSE)) between the original data and its quantized version [36, 37]. In graphics, color quantization is used to reduce the number of colors in the images for displays with various capabilities [38]. In communication, quantization is often used to minimize the decoding errors. Broadly, any conversion of a high-resolution signal to a low-resolution signal requires quantization. In this chapter, we consider the quantization in the context of a communication channel where the transmitted binary signal is corrupted by a continuous noise, resulting in a continuous-valued signal at the receiver. To recover the transmitted signal, the receiver performs a quantization algorithm that maps the received continuous-valued signal to the quantized signal such that the objective function between the input and the quantized output is maximized/minimized. There is a rich literature on quantizer design that minimizes various objectives. One popular objective is to minimize the average decoding error. Another fundamental objective is to maximize the mutual information between the discrete transmitted inputs and the quantized outputs. Equivalently, this objective minimizes the information loss between the inputs and the outputs, and is related to the capacity of the channel. Specifically,

for a given discrete memoryless channel (DMC) specified by a channel matrix M, its capacity is found by maximizing the mutual information between the input and the output with respect to the input distribution p [3], [9]. On the other hand, our work is focused on maximizing the mutual information with respect to the quantization parameters, i.e, it is equivalent to designing a channel matrix M for a fixed distribution p that maximizes the capacity. This situation often arises in real-world scenarios where the distribution of input is already given. In addition, many recent works have proposed to use quantization strategies that maximize the mutual information in the designs of low density parity check codes (LDPC) [39, 40] and polar codes [41].

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

The outline of this chapter is as follows. First, we discuss a few related works in Section 3.2. In Section 3.3, we formulate the problem of designing the optimal quantizer that maximizes the mutual information. In Section 3.4, we describe the structure of optimal quantizers. In Section 3.5, we describe the sufficient conditions via the fixed point theorem for the optimality and uniqueness of a quantizer, together with an efficient procedure for finding the optimal quantizer.

3.2 Related Work

Research on quantization techniques has a long history, including many earliest works in 1960s [42] that aim to minimize the distortion between the original signal and the quantized signal. From a communication perspective, designing the quantizers that maximize the information capacity for Gaussian channels have also been proposed in 1970s [43]. Recently, in constructing efficient codes such as LDPC and polar codes, a number of works have made use of quantizers that maximize the mutual information [39–41]. Many advanced quantization algorithms have also been proposed to maximize the mutual information between the input and the quantized output over the past decade [44–49]. In [44], the channel is assumed to have discrete input and discrete output, and the optimal quantizers can be found efficiently using dynamic programming that has polynomial time complexity [50]. On the other hand, we study the channels with discrete binary inputs and continuous-valued outputs which are then quantized to binary outputs. The continuous-valued output is a direct result of the conditional channel density. We note that it is possible to first discretize the continuous-valued output, then use the existing quantization algorithms for the discrete input-discrete output channels [44]. However, in many scenarios, this may result in loss of efficiencies. In particular, many analytical and computational techniques for dealing with continuous-valued functions are more efficient than their discrete counterparts.

Our work is also related to the classification problem in learning theory. Burshtein et al. gave the condition on the existence of an optimal quantizer which minimizes the impurity of partitions [51]. Because of the similarity between maximizing mutual information and minimizing conditional entropy function [44], [52], the result in [51] can be applied for finding the optimal quantizer. A similar result also can be found in [53]. In [54], Zhang et al. show that finding an optimal quantizer is equivalent to finding an optimal clustering. Therefore, a locally optimal solution can be found using k-means algorithm with the Kullback-Leibler (KL) divergence as

the distance metric. Recently, there have also been many works on approximating the optimal clustering that minimize the impurity function for high dimensional data [55], [56], [1].

There are also works on finding channel capacity by maximizing the mutual information over both input probability mass function (pmf) and thresholds variables. This problem remains a hard problem [45], [57], [58], [59], [60]. Although the mutual information is a convex function in the input pmf, it is not a convex function in the quantization parameters. As such, many successful convex optimization techniques for finding the optimal solution are not applicable. In [57], a heuristic near optimal quantization algorithm is proposed. However, the algorithm only works well when the SNR ratio is high. In [45], R. Mathar et al. investigated an optimal quantization strategy for binary input-multiple output channels using two support points. These results are only applicable to approximate the optimal point between two supporting points. In [52], Kurkoski et al. constructed a sufficient condition such that a single threshold quantizer is optimal for arbitrary binary-input, continuous-output channels based on Burshtein et al.'s theorem on optimal classification [51]. On the other hand, our work describes the generalized conditions for the existence of a single threshold optimal quantizer together with a simple procedure that is able to find the globally optimal quantizer efficiently.

3.3 Problem description

We consider the channel shown in Fig. 3.1 where the binary signals $x \in X = \{0, 1\}$ are transmitted and corrupted by a continuous noise source to produce a continuous-valued output $y \in \mathbb{R}$ at the receiver. Specifically, y is specified by the a channel conditional density $p(y|x)$. $p(y|x)$ models the distortion caused by noise. The receiver recovers the original binary signal x by decoding the received continuous-valued signal y to $z \in Z = \{0, 1\}$ using a quantizer Q. Since $y \in \mathbb{R}$, the

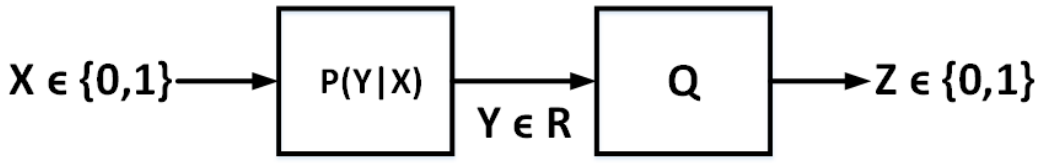


Figure 3.1: Channel model: binary input X is corrupted by continuous noise to result in continuous-valued Y at the receiver. The receiver attempts to recover X by quantizing Y into binary signal Z.

quantization parameters can be specified by a thresholding vector

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

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

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

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

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

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

$$Z = \begin{cases} 0 & \text{if } Y \in \mathbb{H}, \\ 1 & \text{if } Y \in \bar{\mathbb{H}}. \end{cases} \tag{3.1}$$

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

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} I(X;Z). \tag{3.2}$$

We note that both the values of thresholds h_i's and the number of thresholds n are the optimization variables. The maximization in (3.2) assumes that the input probability mass function $p(x)$ and the channel conditional density $p(y|x)$ are given.

3.4 Optimal Quantizer Structure

For convenience, we use the following notations:

1. $\mathbf{p} = (p_0, p_1)$ denotes the probability mass function for the input X, with $p_0 = P(X = 0)$ and $p_1 = P(X = 1)$.

2. $\mathbf{q} = (q_0, q_1)$ denotes probability mass function for the output Z, with $q_0 = P(Z = 0)$ and $q_1 = P(Z = 1)$.

3. $\phi_0(y) = p(y|x = 0)$ and $\phi_1(y) = p(y|x = 1)$ denote conditional density functions of the received signal Y given the input signal $X = 0$ and $X = 1$, respectively.

Furthermore, we make two following assumptions:

Assumptions:

1. $r(y) = \dfrac{\phi_0(y)}{\phi_1(y)}$ will play a central role in this chapter. All the results assume that $r(y)$ is a continuous function, and has a finite number of stationary points. Equivalently, $r(y) = r'$

has a finite number of solutions for any constant $r' > 0$. Note that this assumption will hold for most $\phi_0(y)$ and $\phi_1(y)$.

2. Both $\phi_0(y)$ and $\phi_1(y)$ are differentiable everywhere.

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

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

where

$$A_{11} = \int_{y \in \mathbb{H}} \phi_0(y) dy, \tag{3.3}$$

$$A_{22} = \int_{y \in \bar{\mathbb{H}}} \phi_1(y) dy. \tag{3.4}$$

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

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

In general, this quantizer is not optimal, i.e., does not maximize the mutual information $I(X; Z)$. Using the results of Burshtein et al. [51], Kurkoski et al. [52] showed a sufficient condition on $p(y|x)$ for which the single threshold quantizer is indeed an optimal quantizer. Our first contribution is to show that the optimal binary quantizer with multiple thresholds, specified by a thresholding vector $\mathbf{h}^* = (h_1^*, h_2^*, \ldots, h_n^*)$ with $h_i^* < h_{i+1}^*$, must satisfy the conditions stated in the Theorem 3.1.

Theorem 3.1. *Let $\boldsymbol{h}^* = (h_1^*, \ldots, h_n^*)$ be a thresholding vector of an optimal quantizer Q^*, then:*

$$\frac{\phi_0(h_i^*)}{\phi_1(h_i^*)} = \frac{\phi_0(h_j^*)}{\phi_1(h_j^*)} = r^*, \tag{3.5}$$

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

Proof. We note that using the optimal thresholding vector $\mathbf{h}^*$, the quantization mapping follows (3.1). $\mathbf{h}^*$ divides $\mathbb{R}$ into $n+1$ contiguous disjoint segments, each maps to either 0 or 1 alternatively. The overall DMC in Fig. 3.1 has the channel matrix

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

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

$$I(\mathbf{h}) = H(Z) - H(Z|X) = H(q_0) - [p_0 H(A_{11}) + p_1 H(A_{22})], \tag{3.6}$$

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

This is an optimization problem that maximizes $I(\mathbf{h})$. The theory of optimization requires that an optimal point must satisfy the KKT conditions [24]. In particular, define the Lagrangian function as:

$$L(\mathbf{h}, \lambda) = I(\mathbf{h}) + \sum_{i=1}^{n-1} \lambda_i (h_i - h_{i+1}), \tag{3.7}$$

then the KKT conditions [24] states that, an optimal point $\mathbf{h}^*$ and $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_{n-1}^*)$ must

satisfy:

$$\begin{cases} \frac{\partial L(\mathbf{h},\lambda)}{\partial h_i}|_{\mathbf{h}=\mathbf{h}^*,\lambda=\lambda^*} = 0, i = 1, 2, \ldots, n-1, \\ \lambda_i^*(h_i - h_{i+1}) = 0, i = 1, 2, \ldots, n-1, \\ \lambda_i^* \geq 0, i = 1, 2, \ldots, n-1. \end{cases} \tag{3.8}$$

Since the structure of the quantizer requires that $h_i < h_{i+1}$, the second and the third conditions in (3.8) together imply that $\lambda_i^* = 0, i = 1, 2, \ldots, n-1$. Consequently, from (3.7) and the first condition in (3.8), we have:

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

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

$$\begin{aligned} \frac{\partial I(\mathbf{h})}{\partial h_i} &= (\log\frac{1-q_0}{q_0})\frac{\partial q_0}{\partial h_i} - p_0(\log\frac{1-A_{11}}{A_{11}})\frac{\partial A_{11}}{\partial h_i} \\ &- p_1(\log\frac{1-A_{22}}{A_{22}})\frac{\partial A_{22}}{\partial h_i} \\ &= (\log\frac{1-q_0}{q_0})(p_0\frac{\partial A_{11}}{\partial h_i} - p_1\frac{\partial A_{22}}{\partial h_i}) \\ &- p_0(\log\frac{1-A_{11}}{A_{11}})\frac{\partial A_{11}}{\partial h_i} - p_1(\log\frac{1-A_{22}}{A_{22}})\frac{\partial A_{22}}{\partial h_i} \qquad (3.9) \\ &= p_0\frac{\partial A_{11}}{\partial h_i}(\log\frac{1-q_0}{q_0} - \log\frac{1-A_{11}}{A_{11}}) \\ &- p_1\frac{\partial A_{22}}{\partial h_i}(\log\frac{1-q_0}{q_0} + \log\frac{1-A_{22}}{A_{22}}) = 0, \qquad (3.10) \end{aligned}$$

with (3.9) due to $q_0 = p_0A_{11} + p_1A_{21} = p_0A_{11} + p_1(1 - A_{22})$.

Since $\frac{\partial A_{11}}{\partial h_i} = \phi_0(h_i)$ and $\frac{\partial A_{22}}{\partial h_i} = -\phi_1(h_i)$, from (3.10), we have:

$$\frac{\phi_0(h_i^*)}{\phi_1(h_i^*)} = -\frac{p_1}{p_0}\frac{\log\frac{1-q_0}{q_0} + \log\frac{1-A_{22}}{A_{22}}}{\log\frac{1-q_0}{q_0} - \log\frac{1-A_{11}}{A_{11}}} = r^*. \tag{3.11}$$

Since $r^* > 0$ (please see Appendix 3.7.5) and (3.11) holds for $\forall\ i$, the RHS of (3.11) equals to some constant $r^* > 0$ for a quantizer Q^*. Theorem 3.1 follows. □

Remark: The importance of Theorem 3.1 is as follows. Suppose the optimal value r^* is given and the equation $r(y) = r^*$ has m solutions: $y_1 < y_2 < \cdots < y_m$. Then, Theorem 3.1 says that the optimal quantizer must either have its thresholding vector be $(y_1, y_2, \ldots, y_m)$ or one of its ordered subsets, e.g., $(h_1^*, h_2^*) = (y_1, y_3)$, or both. In Theorem 3.2 below, we will show that the quantizer whose thresholding vector is all the solutions of $r(y) = r^*$, will be at least as good as any quantizer whose thresholding vector is an ordered subset of the set of all solutions. Moreover, we will show that under some sufficient conditions via Banach's fixed point theorem, r^* is unique, and describe an efficient procedure for finding r^* in Section 3.5.

Theorem 3.2. *Let $y_1^* < y_2^* < \cdots < y_n^*$ be the solutions of $r(y) = r^*$ for the optimal constant $r^* > 0$. Let $Q_{r^*}^n$ be the quantizer whose thresholding vector is all the solutions, i.e., $h_i^* = y_i^*, i = 1, 2, \ldots, n$, then for $k < n$, $Q_{r^*}^n$ is at least as good as any quantizer $Q_{r^*}^k$ whose thresholding vector is an ordered subset of k elements of the set of $(h_1^*, h_2^*, \ldots, h_n^*)$.*

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

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

$$I(z_1^*, z_2^*, \ldots, z_{m-1}^*) = I(h_1^*, h_2^*, \ldots, h_m^*) + \delta, \tag{3.12}$$

where δ is a positive constant.

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

$$I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(h_1, h_2, \ldots, h_{m-1}, h_m), \tag{3.13}$$

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

Now replacing $h_i = z_i^*$, for $i = 1, 2, \ldots, m-1$ into (3.13), we have:

$$I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(z_1^*, z_2^*, \ldots, z_{m-1}^*, h_m). \tag{3.14}$$

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

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

Consequently,

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

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

$$|I(z_1^*, z_2^*, \ldots, z_{m-1}^*, h_m) - I(z_1^*, z_2^*, \ldots, z_{m-1}^*)| \leq \epsilon, \tag{3.15}$$

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

$$\begin{aligned} & I(h_1^*, h_2^*, \ldots, h_m^*) && (3.16) \\ = \; & I(z_1^*, \ldots, z_{m-1}^*) + I(h_1^*, h_2^*, \ldots, h_m^*) - I(z_1^*, z_2^*, \ldots, z_{m-1}^*) \\ \geq \; & I(z_1^*, \ldots, z_{m-1}^*) - |I(h_1^*, h_2^*, \ldots, h_m^*) - I(z_1^*, z_2^*, \ldots, z_{m-1}^*)| \\ \geq \; & I(h_1^*, h_2^*, \ldots, h_m^*) + \delta - \epsilon, && (3.17) \end{aligned}$$

where (3.17) is due to (3.12) and (3.15). Since $\delta - \epsilon > 0$ by assumption, (3.17) indicates that $I(h_1^*, h_2^*, \ldots, h_m^*)$ is strictly greater than itself which is a contradiction. Thus, $I(h_1^*, h_2^*, \ldots, h_m^*) \geq I(z_1^*, z_2^*, \ldots, z_{m-1}^*)$.

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

□

Corollary 3.1. *If*

$$r(y) = \frac{\phi_0(y)}{\phi_1(y)} \tag{3.18}$$

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

Proof. Noting that since $r(y)$ is a strictly increasing or decreasing function. Therefore, $r(y_1) \neq r(y_2)$ for $y_1 \neq y_2$. Thus, (3.5) will not hold for $h_1^* \neq h_2^*$. Consequently, the optimal quantizer has

only a single threshold. □

We note that in a previous result [52], an optimality condition for a single threshold quantizer is that:

$$s(y) = \log \frac{\phi_0(y)}{\phi_1(y)} \tag{3.19}$$

is a monotonic function. If $\frac{\phi_0(y)}{\phi_1(y)}$ is a strictly monotonic function, then previous result is a consequence of Corollary 3.1 since log(.) is a strictly monotonic function, any strictly monotonic function $\frac{\phi_0(y)}{\phi_1(y)}$ results in a strictly monotonic function $s(y)$.

Corollary 3.2. *If*

$$\phi_0(y-\mu) = \phi_1(y) \text{ for some constant } \mu, \tag{3.20}$$

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

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

$$\frac{dr(y)}{dy} = \frac{\phi_0'(y)\phi_1(y) - \phi_0(y)\phi_1'(y)}{\phi_1(y)^2} > 0, \tag{3.21}$$

which is equivalent with:

$$\frac{\phi_0'(y)}{\phi_0(y)} > \frac{\phi_1'(y)}{\phi_1(y)}. \tag{3.22}$$

Using (3.20), we have:

$$\frac{\phi_0'(y)}{\phi_0(y)} > \frac{\phi_0'(y-\mu)}{\phi_0(y-\mu)}. \tag{3.23}$$

Now, a function $f(x)$ is strictly log-convex if and only if $\frac{f'(x)}{f(x)}$ is a strictly increasing function

[24]. Thus, if $\phi_0(y)$ is strictly log-convex, then

$$\frac{\phi_0'(y)}{\phi_0(y)} > \frac{\phi_0'(y-\mu)}{\phi_0(y-\mu)}. \tag{3.24}$$

Thus, $r'(y) > 0$ or $r(y)$ is a strictly increasing function which satisfies the condition for having an optimal single threshold quantizer in Corollary 3.1. A similar proof can be established for log-concave functions. □

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

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

3.5.1 Necessary Conditions for Optimality via Fixed Point Theorem

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

$$a = \frac{p_1\phi_1(y)}{p_0\phi_0(y) + p_1\phi_1(y)} = \frac{1}{1 + \frac{p_0\phi_0(y)}{p_1\phi_1(y)}} = \frac{1}{1 + (\frac{p_0}{p_1})r}, \tag{3.25}$$

where

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

We note that $a \in (0,1)$. In addition, the mapping from r to a is a one-to-one mapping. Furthermore, each value of a corresponds to a different value of r which in turn, corresponds to a quantizer in a set of possible quantizers that contains an optimal quantizer. As an example, Fig. 3.2 shows two conditional densities $\phi_0(y)$ and $\phi_1(y)$, and the corresponding $r(y)$ and $u(y)$ are shown in Fig. 3.3 and Fig. 3.4, respectively. Now, the mutual information $I(X;Z)$ can be re-written as a function of a, and is denoted as $I(X;Z)_a$. Thus, finding the optimal r^* is equivalent to finding the optimal a^* that maximizes $I(X;Z)_a$. Furthermore, the optimal thresholds $\mathbf{h}^* = (h_1^*, \ldots, h_n^*)$ can be directly determined as the solutions of

$$\frac{p_1\phi_1(h)}{p_0\phi_0(h) + p_1\phi_1(h)} = a^*. \tag{3.26}$$

First, let

$$u(y) = \frac{p_1\phi_1(y)}{p_0\phi_0(y) + p_1\phi_1(y)}. \tag{3.27}$$

For given a, define $\mathbb{H}_a = \{y : u(y) < a\}$ and $\bar{\mathbb{H}}_a = \{y : u(y) \geq a\}$. The sets $\mathbb{H}_a$ and $\bar{\mathbb{H}}_a$ together specify a binary quantizer that maps y to $z \in \{0,1\}$, depending on whether y belongs to $\mathbb{H}_a$ or $\bar{\mathbb{H}}_a$ as shown in Fig. 3.4.

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

$$z = \begin{cases} 0 & y \in \mathbb{H}_a, \\ 1 & y \in \bar{\mathbb{H}}_a, \end{cases} \tag{3.28}$$

then the channel matrix of the overall DMC is:

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

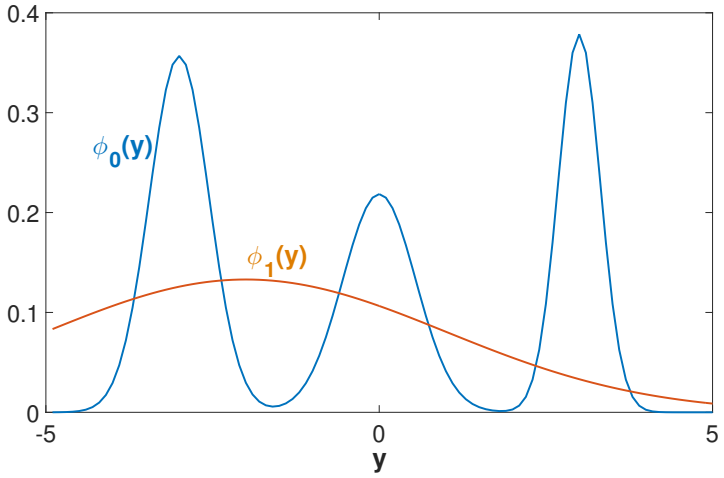


Figure 3.2: Conditional densities $\phi_0(y) = 0.3N(0, \sqrt{0.3}) + 0.4N(-3, \sqrt{0.2}) + 0.3N(3, \sqrt{0.1})$ and $\phi_1(y) = N(-2, 3)$. They are used in Fig. 3.3 and Fig. 3.4.

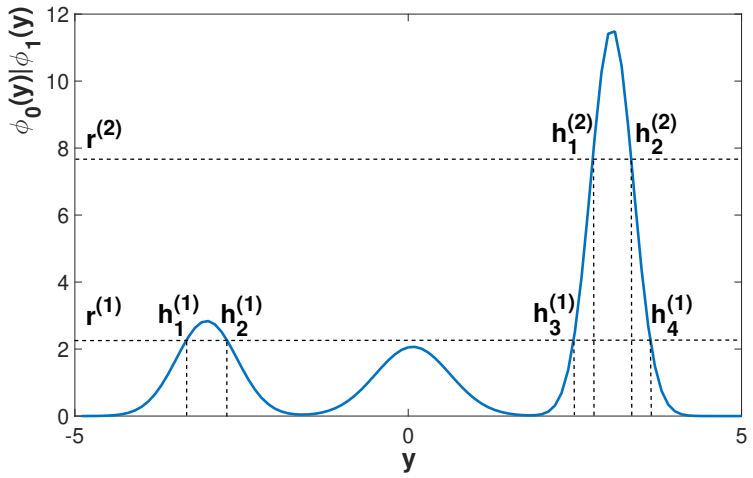


Figure 3.3: Two thresholding vectors: $\mathbf{h}^{(1)} = (h_1^{(1)}, h_2^{(1)}, h_3^{(1)}, h_4^{(1)})$ and $\mathbf{h}^{(2)} = (h_1^{(2)}, h_2^{(2)})$ correspond to two different values of r are shown. $\phi_0(y) = 0.3N(0, \sqrt{0.3}) + 0.4N(-3, \sqrt{0.2}) + 0.3N(3, \sqrt{0.1})$, $\phi_1(y) = N(-2, 3)$.

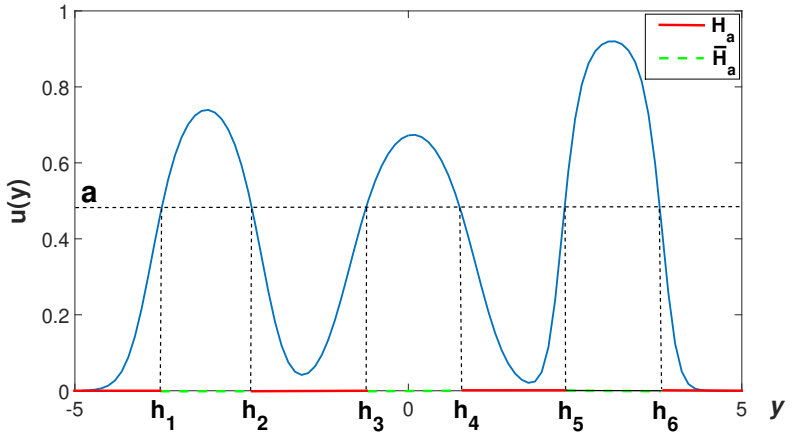


Figure 3.4: Illustration of the sets $\mathbb{H}_a$ and $\bar{\mathbb{H}}_a$. $\mathbb{H}_a$ consists of solid red segments while $\bar{\mathbb{H}}_a$ consists of green dotted segments. In this example, there exists a quantizer with 6 thresholds $h_1, h_2, \ldots, h_6$ that correspond to a specific value of $a = 0.5$. $p_0 = p_1 = 0.5$, $\phi_0(y) = 0.3N(0, \sqrt{0.3}) + 0.4N(-3, \sqrt{0.2}) + 0.3N(3, \sqrt{0.1})$, $\phi_1(y) = N(-2, 3)$.

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

$$f(a) = \int_{y \in \mathbb{H}_a} \phi_0(y) dy, \tag{3.29}$$

$$g(a) = \int_{y \in \bar{\mathbb{H}}_a} \phi_1(y) dy. \tag{3.30}$$

Now, let us consider the special cases where $a = 1$ or $a = 0$. In these cases, $I(X; Z) = 0$ due to $f(a) = 1$ and $g(a) = 0$ or vice-versa. Therefore, $a = 1$ and $a = 0$ cannot be the optimal points. Thus, we can assume that $a \in (0, 1)$ and $0 < f(a), g(a) < 1$. Lemmas 3.1 and 3.2 below provide the properties of $f(a)$ and $g(a)$ and the relationship with each other.

Lemma 3.1. *Derivatives of $f(a)$ and $g(a)$ are related through the following equation:*

$$\frac{dg(a)}{da} = -\frac{ap_0}{(1-a)p_1} \frac{df(a)}{da}. \tag{3.31}$$

Proof. Please see the proof in Appendix 3.7.1. □

Lemma 3.2. *For* $\forall\ a \in (0,1)$,

(1) $g'(a) < 0$ *and* $f'(a) > 0$.

(2) $f(a) + g(a) > 1$.

Proof. Please see the proof in Appendix 3.7.2. □

Define

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

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

$$D_{KL}(\mathbf{x}||\mathbf{y}) = x \log(\frac{x}{y}) + (1-x) \log(\frac{1-x}{1-y}). \tag{3.32}$$

Lemma 3.3. *Each optimal quantizer* Q^* *(local or global) corresponds to an optimal* a^* *such that*

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

Proof. Using Lemma 3.1, setting derivative of $I(X;Z)_a$ to zero, we have:

$$\begin{aligned}\frac{dI(X;Z)_a}{da} &= p_1 g'(a)\big[\frac{a-1}{a}\big(\log(\frac{f(a)}{1-f(a)}) \qquad (3.33)\\ &- \log(\frac{p_0 f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1 g(a)})\big)\\ &+ \log(\frac{g(a)}{1-g(a)}) + \log(\frac{p_0 f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1 g(a)})\big]\\ &= p_1 g'(a) F(a) = 0, \qquad (3.34)\end{aligned}$$

where

$$\begin{aligned}F(a) &= \frac{a-1}{a}\log(\frac{f(a)}{1-f(a)}) + \log(\frac{g(a)}{1-g(a)})\\ &+ \frac{1}{a}\log(\frac{p_0 f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1 g(a)}). \qquad (3.35)\end{aligned}$$

From Lemma 3.2 $g'(a) < 0$ and $p_1 > 0$, thus, the stationary points of $I(X;Z)_a$ must occur at $F(a) = 0$. Applying definitions of $\mathbf{a}, \mathbf{l}_a, \mathbf{r}_a$, and KL divergence, it can be shown that

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

Please see the proof in Appendix 3.7.3. Thus,

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

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

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

□

Lemma 3.4. *Let* $\boldsymbol{c}_a = [1 - c(a), c(a)]$ *then*

$$c(a) = \frac{\log(\frac{1-f(a)}{f(a)}\frac{p_0 f(a) + p_1(1-g(a))}{p_0(1-f(a)) + p_1 g(a)})}{\log(\frac{1-f(a)}{f(a)}\frac{1-g(a)}{g(a)})} \tag{3.36}$$

if and only if

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

Proof. By using the definitions of $\mathbf{c}_a, \mathbf{l}_a, \mathbf{r}_a$, and KL divergence, (3.37) follows. Now, $(1 - f(a))(1 - g(a)) = 1 - f(a) - g(a) + f(a)g(a) < f(a)g(a)$ due to $f(a) + g(a) > 1$. Thus, $\log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)})) \neq 0$. Therefore, $D_{KL}(\mathbf{c}_a||\mathbf{l}_a) - D_{KL}(\mathbf{c}_a||\mathbf{r}_a) = 0$ if and only if $c(a)$ satisfies (3.36). □

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

Theorem 3.3. *Let a quantizer* Q^* *be an optimal quantizer with an optimal* a^**, then* $c(a^*) = a^*$ *where* $c(a)$ *is defined in (3.36).*

Proof. From Lemma 3.3, the optimal quantizer Q^* corresponds to an optimal vector $\mathbf{a}^* = [1 - a^*, a^*]$ must have $D_{KL}(\mathbf{a}^*||\mathbf{l}_{a^*}) = D_{KL}(\mathbf{a}^*||\mathbf{r}_{a^*})$. Now, from Lemma 3.4 for given $\mathbf{l}_{a^*}$ and $\mathbf{r}_{a^*}$, there exists a unique vector $\mathbf{c}_{a^*} = [1 - c(a^*), c(a^*)]$ such that $D_{KL}(\mathbf{c}_{a^*}||\mathbf{l}_{a^*}) = D_{KL}(\mathbf{c}_{a^*}||\mathbf{r}_{a^*})$ where $c(a)$ is defined in (3.36). Combining Lemma 3.3 and 3.4, we have $c(a^*) = a^*$. □

We will use Theorem 3.3 in our algorithm for finding optimal quantizers. To do that, we will show some interesting properties of $c(a)$ in Theorem 3.4 and Theorem 3.5 below.

Theorem 3.4. $c(a) \in (0, 1)$ *and is a smooth (derivative exists), non-decreasing function of* a*.*

Proof. Please see Appendix 3.7.4 for the proof. □

$$
\begin{aligned}
& D_{KL}(\mathbf{c}_a||\mathbf{l}_a) - D_{KL}(\mathbf{c}_a||\mathbf{r}_a) \\
= \quad & \big(c(a)\log(\frac{c(a)}{\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}}) + (1-c(a))\log(\frac{1-c(a)}{\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}})\big) \\
- \quad & \big(c(a)\log(\frac{c(a)}{\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}) + (1-c(a))\log(\frac{1-c(a)}{\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}})\big) \\
= \quad & c(a)\log(\frac{\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}{\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}}) + (1-c(a))\log(\frac{\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}}{\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}}) \\
= \quad & \log(\frac{\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}}{\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}}) - c(a)\big(\log(\frac{\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}}{\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}}) \\
& - \log(\frac{\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}{\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}})\big) \\
= \quad & \log((\frac{1-f(a)}{f(a)})(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)})) - c(a)\log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)})) \\
= \quad & \log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)}))\big(\frac{\log((\frac{1-f(a)}{f(a)})(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)}))}{\log((\frac{1-f(a)}{f(a)})(\frac{1-g(a)}{g(a)}))} - c(a)\big).
\end{aligned}
\tag{3.37}
$$

Lemma 3.5. *The sequence $a^{i+1} = c(a^i)$ must converge to a fixed point a^* for any initial point $a^0 \in (0, 1)$.*

Proof. From Theorem 3.4, $c(a)$ is a non-decreasing function and $c(a) \in (0, 1)$. Thus, the sequence generated by $a^{i+1} = c(a^i)$, starting from any a^0 is monotone, i.e., $a^{i+1} \geq a^i \ \forall \ i$ or $a^{i+1} \leq a^i \ \forall \ i$. Specifically, if $a^1 \leq a^0$, then $a^2 = c(a^1) \leq c(a^0) = a^1$, therefore, $a^2 \leq a^1$. By induction method, if $a^1 \leq a^0$ then $a^{i+1} \leq a^i \ \forall \ i$. Similarly, if $a^1 \geq a^0$ then $a^{i+1} \geq a^i \ \forall \ i$. Thus, the sequence a^i is monotone. From Theorem 3.4, $c(a^i) \in (0, 1)$ or the sequence a^i is bounded in $(0, 1)$. Thus, sequence a^i has a limit a^* such that $a^* = c(a^*)$. □

Theorem 3.5. *For any initial point $a^0 \in (0, 1)$, if $\lim_{i \to +\infty} a^i = a^*$ where $a^{i+1} = c(a^i)$, then there is no other solution a' such that $a' = c(a')$ between a_0 and a^*.*

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

Fig. 3.5 illustrates the convergence of sequence a^i to a^* from the initial point a^0. □

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

A straightforward way of computing the optimal a^* is the iteration method by starting with a^0. However, depending on the starting point a^0, the iterations may lead to a local optimal solution. In other words, when the equation $a = c(a)$ has more than one solution, we need a procedure capable of finding all the solutions of $a = c(a)$. Using Theorem 3.5, we outline an efficient procedure that can find all the solutions to $a = c(a)$. A global solution then can be chosen among

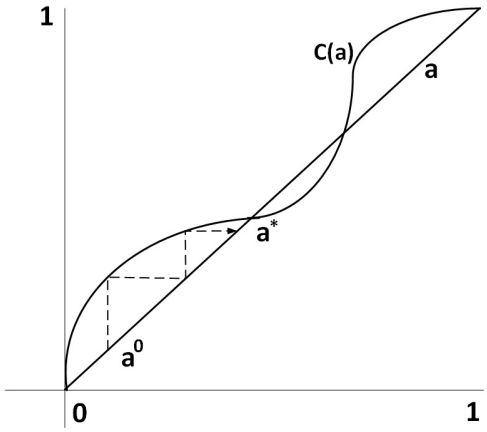


Figure 3.5: Illustration of the convergence of sequence a^i to a^* from the initial point a^0.

these solutions that maximize the mutual information.

Our procedure initiates two iteration loops using two starting points $a_l^0 = \epsilon$ and $a_r^0 = 1 - \epsilon$ where ϵ is a small number. Suppose that the first iteration loop converges to a_l^*, and the second iteration loop converges to a_r^*. If $a_l^* = a_r^*$, then the procedure terminates with $a^* = a_r^*$ being the optimal point. This is due to Theorem 3.5 which states that there is no solution of $a = c(a)$ in either (ϵ, a^*) or $(a^*, 1 - \epsilon)$. We assume that the optimal solution is not in $(0, \epsilon)$ or $(1 - \epsilon, 1)$ since we can make ϵ arbitrarily small. Otherwise, if $a_l^* < a_r^*$, we need to check whether or not there exists some other solutions in the interval (a_l^*, a_r^*). In order to find them, the procedure initiates another iteration loop using a starting point $a^0 = (a_l^* + a_r^*)/2$. After this iteration loop converges to a_c^*, one needs to run the iterations over two intervals $(a_l^*, \min(a^0, a_c^*))$ and $(\max(a^0, a_c^*), a_r^*)$. If any of these intervals is nonempty, then the procedure recursively repeats the previous steps until the whole interval $(0, 1)$ has been completely searched. When all a^*'s are found, we pick the one that maximizes the mutual information. Note that this fixed point method is much faster than an exhaustive search through all the values of a. Finally, we note that our procedure is based on the algorithm in [61].

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

Corollary 3.3. *Let $d(x, y)$ is an arbitrary distance metric between x and y. If there exists a $q \in [0, 1)$ such that for all $x, y \in (0, 1)$*

$$d(c(x), c(y)) \leq qd(x, y), \tag{3.38}$$

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

Proof. From Theorem 3.4, obviously that $a \in (0, 1)$ and $c(a) \in (0, 1)$. Thus, $c(a)$ maps to itself. If existing q and $d(,)$ such that $d(c(x), c(y)) \leq qd(x, y)$ for all $x, y \in (0, 1)$ then $c(.)$ is a contraction mapping. From Banach's fixed point theorem [62], there exists a unique a^* such that $c(a^*) = a^*$. □

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

3.6 Conclusion

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

3.7 Appendix

3.7.1 Proof for Lemma 3.1

From (3.25), we have:

$$\phi_1(h_i) = \frac{ap_0}{(1-a)p_1}\phi_0(h_i), \forall i \in \{1, 2, \ldots, n\}. \tag{3.39}$$

Now, suppose that $u(y) = a$ having n solutions $\{h_1, h_2, \ldots, h_n\}$. Without loss of generality, suppose that $\mathbb{H}_a = \{(-\infty, h_1) \cup [h_2, h_3) \cup \cdots \cup [h_n, +\infty)\}$ and $\bar{\mathbb{H}}_a = \mathbb{R} \setminus \mathbb{H}_a = \{[h_1, h_2) \cup [h_3, h_4) \cup \cdots \cup [h_{n-1}, h_n)\}$. From (3.29) and (3.30)

$$\frac{df(a)}{da} = \frac{\partial f(a)}{\partial h}\frac{\partial h}{\partial a} = +\phi_0(h_1)\frac{\partial h_1}{\partial a} - \phi_0(h_2)\frac{\partial h_2}{\partial a} + \ldots - \phi_0(h_n)\frac{\partial h_n}{\partial a}, \tag{3.40}$$

$$\frac{dg(a)}{da} = \frac{\partial g(a)}{\partial h}\frac{\partial h}{\partial a} = -\phi_1(h_1)\frac{\partial h_1}{\partial a} + \phi_1(h_2)\frac{\partial h_2}{\partial a} - \ldots + \phi_1(h_n)\frac{\partial h_n}{\partial a}. \tag{3.41}$$

Combining Eqs. (3.39), (3.40) and (3.41), we have the desired proof. We note that $f'(a)$ and $g'(a)$ have the opposite sign. As a result, if $f(a)$ increases, then $g(a)$ decreases and vice-versa.

■

3.7.2 Proof for Lemma 3.2

(1) From (3.26), $f(a)$ represents the quantized bit "0" which is the area of $u(y)$ (defined in (3.27)) where $u(y) < a$. Therefore, if a is increasing, $f(a)$ is obviously increasing. Thus, $f'(a) > 0$. A similar proof can be established for $g(a)$ which corresponds to the area of $u(y)$ where $u(y) \geq a$.

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

From (3.26)

$$ap_0\phi_0(y) > (1-a)p_1\phi_1(y), \forall y \in \mathbb{H}_a, \tag{3.42}$$

$$ap_0\phi_0(y) \leq (1-a)p_1\phi_1(y), \forall y \in \bar{\mathbb{H}}_a. \tag{3.43}$$

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

• If $a < p_1$ then $1-a > 1-p_1 = p_0$. Thus, from (3.42), $\phi_0(y) > \phi_1(y)$ for $\forall\ y \in \mathbb{H}_a$. Therefore,

$$\begin{aligned} f(a) + g(a) &= \int_{y\in\mathbb{H}_a} \phi_0(y)dy + \int_{y\in\bar{\mathbb{H}}_a} \phi_1(y)dy & (3.44)\\ &> \int_{y\in\mathbb{H}_a} \phi_1(y)dy + \int_{y\in\bar{\mathbb{H}}_a} \phi_1(y)dy & (3.45)\\ &= 1. & (3.46) \end{aligned}$$

• If $a \geq p_1$ then $1-a \leq 1-p_1 = p_0$. Thus, from (3.43), $\phi_0(y) \leq \phi_1(y)$ for $\forall\ y \in \bar{\mathbb{H}}_a$. Therefore,

$$\begin{aligned} f(a) + g(a) &= \int_{y\in\mathbb{H}_a} \phi_0(y)dy + \int_{y\in\bar{\mathbb{H}}_a} \phi_1(y)dy & (3.47)\\ &\geq \int_{y\in\mathbb{H}_a} \phi_0(y)dy + \int_{y\in\bar{\mathbb{H}}_a} \phi_0(y)dy & (3.48)\\ &= 1. & (3.49) \end{aligned}$$

■

Remark: The necessary condition for inequality (3.49) becomes equality is $\phi_0(y) = \phi_1(y)$ for $\forall\ y \in \bar{\mathbb{H}}_a$ that contradicts to the assumption that $r(y)$ has a finite number of stationary points. Thus, $f(a) + g(a) > 1$.

$$
\begin{aligned}
& \frac{1}{a}\big[D_{KL}(\mathbf{a}||\mathbf{l}_a) - D_{KL}(\mathbf{a}||\mathbf{r}_a)\big] \\
= \; & \frac{1}{a}\Big[\big(a\log(\frac{a}{\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}}) + (1-a)\log(\frac{1-a}{\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))}})\big) \\
- \; & \big(a\log(\frac{a}{\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}}) + (1-a)\log(\frac{1-a}{\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)}})\big)\Big] \\
= \; & \frac{1}{a}\Big[-\big(a\log(\frac{p_1(1-g(a))}{p_0f(a)+p_1(1-g(a))}) + (1-a)\log(\frac{p_0f(a)}{p_0f(a)+p_1(1-g(a))})\big) \\
+ \; & \big(a\log(\frac{p_1g(a)}{p_0(1-f(a))+p_1g(a)}) + (1-a)\log(\frac{p_0(1-f(a))}{p_0(1-f(a))+p_1g(a)})\big)\Big] \qquad (3.50) \\
= \; & \frac{1}{a}\big[(1-a)\log(\frac{p_0(1-f(a))}{p_0f(a)}) + a\log(\frac{p_1g(a)}{p_1(1-g(a))}) + \log(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)})\big] \\
= \; & \frac{1}{a}\big[(a-1)\log(\frac{f(a)}{1-f(a)}) + a\log(\frac{g(a)}{1-g(a)}) + \log(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)})\big] \qquad (3.51) \\
= \; & \frac{a-1}{a}\log(\frac{f(a)}{1-f(a)}) + \log(\frac{g(a)}{1-g(a)}) + \frac{1}{a}\log(\frac{p_0f(a)+p_1(1-g(a))}{p_0(1-f(a))+p_1g(a)}) \qquad (3.52) \\
= \; & F(a), \qquad (3.53)
\end{aligned}
$$

3.7.3 Proof of Lemma 3

By using the definitions of $\mathbf{a}, \mathbf{l}_a, \mathbf{r}_a$ and KL divergence, it can be shown that (3.53) holds with (3.50) due to $a\log a + (1-a)\log(1-a)$ is cancelled after summing up, (3.51) and (3.52) due to a bit of algebra, (3.53) due to the definition of $F(a)$ in (3.35).

3.7.4 Proof Theorem 3.4

We will use the following lemmas and the order notion of 2-dimensional vector below to prove Theorem 3.4.

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

Lemma 3.6. *For any three binary probabiity vectors* $\boldsymbol{a} = [1-a, a]$, $\boldsymbol{b} = [1-b, b]$ *and* $\boldsymbol{c} = [1-c, c]$ *such that* $\boldsymbol{a} \leq \boldsymbol{b} \leq \boldsymbol{c}$ *(or* $a \leq b \leq c$*), then*

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

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

$$D'(b) = \frac{1-a}{1-b} - \frac{a}{b}. \tag{3.54}$$

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

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

□

Lemma 3.7. *If* $D_{KL}(\boldsymbol{c}_a||\boldsymbol{l}_a) = D_{KL}(\boldsymbol{c}_a||\boldsymbol{r}_a)$*, then* $\boldsymbol{l}_a \leq \boldsymbol{c}_a \leq \boldsymbol{r}_a$

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

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

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

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

Lemma 3.8. *Consider a_1 and a_2 such that $0 < a_1 \leq a_2 < 1$, then $\boldsymbol{l}_{a_1} \leq \boldsymbol{l}_{a_2}$ and $\boldsymbol{r}_{a_1} \leq \boldsymbol{r}_{a_2}$.*

Proof. First, we show that $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2}$. Indeed, consider the function $s(a)$ as the ratio of the second entry over the first entry of $\mathbf{l}_a$, i.e., $s(a) = \dfrac{p_1(1-g(a))}{p_0 f(a)}$. We have

$$\begin{aligned} s'(a) &= \frac{-p_1 g'(a) p_0 f(a) - p_1(1-g(a)) p_0 f'(a)}{(p_0 f(a))^2} \\ &= p_0 p_1 f'(a)\big(\frac{a p_0}{(1-a)p_1} f(a) - (1-g(a))\big), \end{aligned} \tag{3.55}$$

with (3.55) due to Lemma 3.2. Also from (3.42),

$$\phi_1(y) < \frac{a p_0}{(1-a)p_1}\phi_0(y), \forall y \in \mathbb{H}_a.$$

Moreover, from the definitions of $f(a)$ and $g(a)$ in (3.29) and (3.30), $f(a)$ and $1-g(a)$ are the

integrals of $\phi_0(y)$ and $\phi_1(y)$, respectively over $\mathbb{H}_a$, respectively. Thus, $\frac{ap_0}{(1-a)p_1}f(a)-(1-g(a)) > 0$. From Lemma 3.2 $f'(a) > 0$, thus $s'(a) > 0$. That said, the ratio of the second entry over the first entry of $\mathbf{l}_a$ is an increasing function of a. Furthermore, $\mathbf{l}_a$ is a probability vector, i.e., the summation of the first entry and the second entry equals one. Therefore, the second entry of $\mathbf{l}_a$ is an increasing function of a or $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2}$.

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

Lemma 3.9. *Consider 4 vectors* $\boldsymbol{a} = [1-a, a]$, $\boldsymbol{b} = [1-b, b]$, $\boldsymbol{c} = [1-c, c]$ *and* $\boldsymbol{d} = [1-d, d]$ *such that* $\boldsymbol{a} \leq \boldsymbol{b} \leq \boldsymbol{c} \leq \boldsymbol{d}$ *(or* $a \leq b \leq c \leq d$*), then*

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

Proof. Proof of (a). We have $D_{KL}(\mathbf{c}||\mathbf{b}) \leq D_{KL}(\mathbf{c}||\mathbf{a})$ and $D_{KL}(\mathbf{c}||\mathbf{a}) \leq D_{KL}(\mathbf{d}||\mathbf{a})$ due to Lemma 3.6 part (b) and (c), respectively. Thus, $D_{KL}(\mathbf{d}||\mathbf{a}) \geq D_{KL}(\mathbf{c}||\mathbf{b})$. The equality happens if and only if $a = b$ and $c = d$.

Proof of (b). Similar to proof of part (a), $D_{KL}(\mathbf{a}||\mathbf{d}) \geq D_{KL}(\mathbf{b}||\mathbf{d})$ and $D_{KL}(\mathbf{b}||\mathbf{d}) \geq D_{KL}(\mathbf{b}||\mathbf{c})$ due to Lemma 3.6 part (a) and (d), respectively. Thus, $D_{KL}(\mathbf{a}||\mathbf{d}) \geq D_{KL}(\mathbf{b}||\mathbf{c})$. The equality happens if and only if $a = b$ and $c = d$. □

Now, we are ready to prove Theorem 3.4.

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

From Lemma 3.7, we have $\mathbf{l}_a \leq \mathbf{c}_a \leq \mathbf{r}_a$. Equivalently,

$$0 < \frac{p_1(1-g(a))}{p_0 f(a) + p_1(1-g(a))} \leq c(a) \leq \frac{p_1 g(a)}{p_0(1-f(a)) + p_1 g(a)} < 1. \tag{3.56}$$

Proof for the smoothness of $c(a)$. Since $0 < f(a), g(a) < 1$, $p_0(1-f(a)) + p_1 g(a) > 0$ and $f(a)g(a) > 0$, thus all of the denominators of (3.36) is positive. In addition, one can verify that

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

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

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

Suppose that there exists $a_1 \leq a_2$ such that $D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) = D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1})$ and $D_{KL}(\mathbf{c}_{a_2}||\mathbf{l}_{a_2}) = D_{KL}(\mathbf{c}_{a_2}||\mathbf{r}_{a_2})$ but $c(a_1) > c(a_2)$. From Lemma 3.8, $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2}$, $\mathbf{r}_{a_1} \leq \mathbf{r}_{a_2}$. From Lemma 3.7, $\mathbf{l}_{a_1} \leq \mathbf{c}_{a_1} \leq \mathbf{r}_{a_1}$ and $\mathbf{l}_{a_2} \leq \mathbf{c}_{a_2} \leq \mathbf{r}_{a_2}$. From the assumption that $c_{a_1} > c_{a_2}$, $\mathbf{c}_{a_1} > \mathbf{c}_{a_2}$. Therefore,

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

Now, using Lemma 3.9 part (a) for $\mathbf{l}_{a_1} \leq \mathbf{l}_{a_2} \leq \mathbf{c}_{a_2} < \mathbf{c}_{a_1}$,

$$D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) > D_{KL}(\mathbf{c}_{a_2}||\mathbf{l}_{a_2}). \tag{3.57}$$

Similarly, using Lemma 3.9 part (b) for $\mathbf{c}_{a_2} < \mathbf{c}_{a_1} \leq \mathbf{r}_{a_1} \leq \mathbf{r}_{a_2}$,

$$D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1}) < D_{KL}(\mathbf{c}_{a_2}||\mathbf{r}_{a_2}). \tag{3.58}$$

From (3.57) and (3.58),

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

that contradicts to our assumption that $D_{KL}(\mathbf{c}_{a_1}||\mathbf{l}_{a_1}) = D_{KL}(\mathbf{c}_{a_1}||\mathbf{r}_{a_1})$. By contradiction method,

$c(a_1) \leq c(a_2)$ if $a_1 \leq a_2$. Thus, $c(a)$ is a non-decreasing function of a. Combining with (3.56), we have the proof for Theorem 3.4.

3.7.5 Proof of $r^* > 0$

From (3.11) and $p_0 > 0$, $p_1 > 0$, $r^* > 0$ is equivalent to

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

Thus, we need to show that

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

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

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

Using a bit of algebra, (3.59) is equivalent to

$$(A_{11} - q_0)(A_{22} - q_1) > 0. \tag{3.59}$$

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

$$q_0 = p_0A_{11} + p_1A_{21} < p_0A_{11} + p_1A_{11} = A_{11}, \tag{3.60}$$

$$q_1 = p_0 A_{12} + p_1 A_{22} < p_0 A_{22} + p_1 A_{22} = A_{22}. \tag{3.61}$$

Combining (3.60) and (3.61), (3.59) follows. The proof is complete.

Chapter 4: Optimal Quantizer Structure for Maximizing Mutual Information Under Constraints

4.1 Introduction

Motivated by the development of polar codes [63] and LDPC codes [39], 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 [44, 45, 52, 54, 57, 64, 65]. Finding an optimal quantizer that maximizes the mutual information in a general setting is an NP-hard problem [60]. 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 [54], [45], [57]. On the other hand, under certain restrictions e.g., binary input channel, there exist polynomial-time algorithms [44], [52], [66] 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 chapter, we are interested in studying the optimal quantizers in the following communication setting. We consider a sender transmits K discrete symbols modeled as a discrete random variable X having a probability mass function $\mathbf{p}(x) = [p(x_1), p(x_2), \ldots, p(x_K)]$ 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 condi-

tional 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 $Z = \{z_1, z_2, \ldots, z_N\}$ such that the mutual information $I(X;Z)$ is maximized subject to an arbitrary constraint on $\mathbf{p}(z) = [p(z_1), p(z_2), \ldots, p(z_N)]$. Formally, we are interested in designing an optimal quantizer Q^* that maximizes $\beta I(X;Z) - C(\mathbf{p}(z))$ where β is a positive number that controls the trade-off between maximizing $I(X;Z)$ and minimizing an arbitrary cost function $C(\mathbf{p}(z))$.

This problem is a generalized version of the Deterministic Information Bottleneck [67], 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 chapter 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 [44,51]. 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 $\mathbf{p}(z)$ as that of $Q(y)$, therefore, the same cost function $C(\mathbf{p}(z))$, and (2) $I(X;Z)$ 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.

4.2 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 [44], [52], [66]. Based on this optimality structure, an optimal quantizer can be found efficiently in polynomial time via dynamic programming technique [44]. In particular, the structure of optimal binary-input quantizers in [44] and [52] is constructed based on the well-known result in [51] for the K-ary inputs. The results in [51] and [68] 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 [51].

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 [69, 70] and entropy-constrained vector quantization [71], [72] 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. [67] 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 [51] for the problem of minimizing impurity without constraints. Specifically, the result in [51] 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. Finally, we note that this work generalized our previous result in [73] for the case of binary input channels.

4.3 Problem Formulation

We consider a discrete input source modeled as a discrete random variable X consisting K discrete symbols $\{x_1, x_2, \ldots, x_K\}$ with a given p.m.f $\mathbf{p}(x) = [p(x_1), p(x_2), \ldots, p(x_K)]$. 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 random variable Z consisting of N discrete outcomes $z_1, z_2, \ldots, z_N$ with the p.m.f $\mathbf{p}(z) = [p(z_1), p(z_2), \ldots, p(z_N)]$. We note that $\mathbf{p}(z)$ depends on Q. Let $C(\mathbf{p}(z))$ be an arbitrary cost function of $\mathbf{p}(z)$. 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 $C(\mathbf{p}(z))$. Formally, we want to solve the following optimization problem:

$$Q^* = \max_Q \beta I(X;Z) - C(\mathbf{p}(z)), \tag{4.1}$$

where β is a pre-specified positive number that controls the trade-off between maximizing $I(X;Z)$ and minimizing $C(\mathbf{p}(z))$. A well-known constraint $C(\mathbf{p}(z))$ on the quantized-output 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 4.7.

4.4 Preliminaries

4.4.1 Notations and definitions

For convenience, we use the following notations and definitions:

1. $\mathbf{p}(x) = [p(x_1), p(x_2), \ldots, p(x_K)] = [p_1, p_2, \ldots, p_K]$ denotes the p.m.f of the input source 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. $\mu(y)$ denotes the density function of y. Specifically,

$$\mu(y) = \sum_{i=1}^{K} p_i \phi_i(y). \tag{4.2}$$

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,

$$p(x_i|y) = \frac{p_i \phi_i(y)}{\sum_{j=1}^{K} p_j \phi_j(y)}. \tag{4.3}$$

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

$$Z_i = \{y : Q(y) = z_i\}. \tag{4.4}$$

Definition 4.1. *(***Convex quantizer** ***(quantizer)****) Let $Z_1, Z_2, \ldots, Z_N$ be the N sets induced by a quantizer $Q(y)$. $Q(y)$ is a convex quantizer (denoted by quantizer) if for any Z_i and Z_j, $i \neq j$, there exists a hyper-plane that separates the two conditional probability vectors $\mathbf{p}(\mathbf{x}|y_i)$ and $\mathbf{p}(\mathbf{x}|y_j)$, $\forall y_i \in Z_i, \forall y_j \in Z_j$.*

We note that a *quantizer* 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 4.2. *(***Kullback-Leibler (KL) divergence***) The KL divergence of two probability vectors $\boldsymbol{a} = (a_1, a_2, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_K)$ is defined by:*

$$D(\boldsymbol{a}||\boldsymbol{b}) = \sum_{i=1}^{K} a_i \log(\frac{a_i}{b_i}). \tag{4.5}$$

Definition 4.3. *(***Centroid***) The centroid of output set 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 $\mathbf{c}_i$, $\forall y \in Z_i$. Formally,*

$$\boldsymbol{c}_i = \arg\min_{\boldsymbol{c}} \int_{y \in Z_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}) \mu(y) dy. \tag{4.6}$$

Definition 4.4. *(****Distortion measurement****) The distortion of a quantizer Q that induces N*

output sets $\{Z_1, Z_2, \ldots, Z_N\}$ *is:*

$$D(Q) = \sum_{i=1}^{N} D(Q_{Z_i}) = \sum_{i=1}^{N} \int_{y \in Z_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)\mu(y)dy, \tag{4.7}$$

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

$$D(Q_{Z_i}) = \int_{y \in Z_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)\mu(y)dy. \tag{4.8}$$

4.4.2 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 [74]. In a special case where the impurity function is the entropy, minimizing entropy impurity is equivalent to maximizing mutual information [44], [54]. 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 [54]. The result in [54] 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 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 [54], we have:

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

Since $\mathbf{p}(x)$ 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:

$$Q^* = \min_Q \mathbb{E}_Y[D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)] \tag{4.9}$$

$$= \min_Q \sum_{i=1}^{N} \int_{y \in Z_i} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_i)\mu(y)dy, \tag{4.10}$$

where $\mu(y)$ is the density of 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 chapter, 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 Z_i, its centroid $\mathbf{c}_i$ can be computed by a closed-form expression (Proposition 1, [75]).

4.5 Structure of optimal quantizer

We show that an optimal quantizer can be found within a class of convex quantizers as defined in Definition 4.1. Our approach is to show that any quantizer can be replaced by an equal or better convex quantizer that maximizes the objective function $\beta I(X;Z) - C(\mathbf{p}(z))$. 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 $\beta I(X;Z) - C(\mathbf{p}(z))$ 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.

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

Theorem 4.1. *Let Q be an arbitrary quantizer that induces two disjoint discrete output sets Z_1 and 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 $\{\tilde{Z}_1, \tilde{Z}_2\}$ having the corresponding centroids $\{\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2\}$ such that (1) $p(Z_i) \stackrel{\triangle}{=} P(y \in Z_i) = p(\tilde{Z}_i) \stackrel{\triangle}{=} P(y \in \tilde{Z}_i), i = 1, 2$, and (2) $D(\tilde{Q}) \leq D(Q)$.*

Proof. Let Q be a given arbitrary quantizer. Q induces Z_1, 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:

$$\begin{aligned} F(\mathbf{p}(\mathbf{x}|y)) &= D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2) \\ &= \sum_{i=1}^{K} p(x_i|y) \log \frac{p(x_i|y)}{c_1^i} - \sum_{i=1}^{K} p(x_i|y) \log \frac{p(x_i|y)}{c_2^i} \\ &= \sum_{i=1}^{K} p(x_i|y) \log \frac{c_2^i}{c_1^i} = \mathbf{a}^T \mathbf{p}(\mathbf{x}|y), \end{aligned} \tag{4.11}$$

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 $H(h)$ in the K-dimensional space parameterized by $h \in \mathbb{R}$ in the following equation:

$$\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) = h. \tag{4.12}$$

For a given h, the hyper-plane $H(h)$ 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

4.1, there is also a family of convex quantizers $\tilde{Q}$ for each h. Our goal is to show that there exists a hyper-plane $H(\tilde{h})$ associated with a convex quantizer $\tilde{Q}$ that separates the space of posterior distribution into two disjoint sets $\{\tilde{Z}_1, \tilde{Z}_2\}$ such that $p(Z_i) = p(\tilde{Z}_i)$, and $D(\tilde{Q}) \leq D(Q)$.

Proof of claim (1). Assume that Q produces two output sets Z_1 and Z_2 with the probability $p(Z_1)$ and $p(Z_2)$, $p(Z_1)+p(Z_2) = 1$. Our first claim is that one can always find a convex quantizer $\tilde{Q}$ corresponding to a hyper-plane $H(\tilde{h})$ that produces $\tilde{Z}_1$ and $\tilde{Z}_2$ such that $p(\tilde{Z}_1) = p(Z_1)$ and $p(\tilde{Z}_2) = p(Z_2)$.

Consider the following convex quantizer:

$$\tilde{Q}(y) = \begin{cases} \tilde{Z}_1 & \text{if } F(\mathbf{p}(\mathbf{x}|y)) \leq h, \\ \tilde{Z}_2 & \text{if } F(\mathbf{p}(\mathbf{x}|y)) > h. \end{cases} \tag{4.13}$$

By increasing value of h, $h \in (-\infty, +\infty)$, the set $\tilde{Z}_1$ must enlarge while $\tilde{Z}_2$ must reduce. Thus, by increasing/decreasing the value of h, one can always choose an appropriate value of $h = \tilde{h}$ such that $p(\tilde{Z}_1) = p(Z_1)$ and $p(\tilde{Z}_2) = p(Z_2)$. $\tilde{h}$ corresponds to the hyper-plane $H(\tilde{h})$ 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 $H(\tilde{h})$ in the proof of claim (1) which produces two discrete output sets $\tilde{Z}_1$ and $\tilde{Z}_2$. Let $A = \tilde{Z}_1 \cap Z_2$ and $B = \tilde{Z}_2 \cap Z_1$. Note that if A or B is empty set then Q can be readily shown to be a convex quantizer. Let $p(A) = P(y \in A)$ and $p(B) = P(y \in B)$. We first show that $p(A) = p(B)$ as follows.

$$p(Z_1) \overset{\tilde{Z}_1 \cap \tilde{Z}_2 = \emptyset}{=} p((Z_1 \cap \tilde{Z}_1) \cup (Z_1 \cap \tilde{Z}_2)) = p(Z_1 \cap \tilde{Z}_1) + p(Z_1 \cap \tilde{Z}_2) = p(Z_1 \cap \tilde{Z}_1) + p(B). \tag{4.14}$$

Similarly,

$$p(\tilde{Z}_1) \overset{Z_1 \cap Z_2 = \emptyset}{=} p((Z_1 \cap \tilde{Z}_1) \cup (\tilde{Z}_1 \cap Z_2)) = p(Z_1 \cap \tilde{Z}_1) + p(\tilde{Z}_1 \cap Z_2) = p(Z_1 \cap \tilde{Z}_1) + p(A). \quad (4.15)$$

Since $p(Z_1) = p(\tilde{Z}_1)$, from (4.14) and (4.15), we have $p(A) = p(B)$.

Next, let $\mu(y)$ be the density of Y. From $F(\mathbf{p}(\mathbf{x}|y_i)) \leq \tilde{h} < F(\mathbf{p}(\mathbf{x}|y_j))$, $\forall y_i \in \tilde{Z}_1$ and $\forall y_j \in \tilde{Z}_2$, together with $A = \tilde{Z}_1 \cap Z_2$ and $B = \tilde{Z}_2 \cap Z_1$, then $F(\mathbf{p}(\mathbf{x}|y_i)) \leq \tilde{h} < F(\mathbf{p}(\mathbf{x}|y_j))$, $\forall y_i \in A$ and $\forall y_j \in B$, (4.16) is established.

Next, by adding $\int_{y \in \{Z_1 \cap \tilde{Z}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y \in \{Z_2 \cap \tilde{Z}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy$ to both sides of (4.16), we obtain (4.17).

By moving $-\int_{y \in A} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy$ to the right hand side and $-\int_{y \in B} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy$ to the left hand side of (4.17), we obtain (4.18).

Now, since $Z_1 \cap Z_2 = \emptyset$, $A \cap \{Z_1 \cap \tilde{Z}_1\} = \{\tilde{Z}_1 \cap Z_2\} \cap \{Z_1 \cap \tilde{Z}_1\} = \emptyset$. Thus, the integral over A and $\{Z_1 \cap \tilde{Z}_1\}$ is equivalent to the integral over $A \cup \{Z_1 \cap \tilde{Z}_1\} = \tilde{Z}_1$. Similarly, using $B \cup \{Z_2 \cap \tilde{Z}_2\} = \tilde{Z}_2$, $B \cup \{Z_1 \cap \tilde{Z}_1\} = Z_1$ and $A \cup \{Z_2 \cap \tilde{Z}_2\} = Z_2$, (4.19) is obtained from (4.18).

Let $\tilde{\mathbf{c}}_1$ and $\tilde{\mathbf{c}}_2$ be the new centroids of $\tilde{Z}_1$ and $\tilde{Z}_2$. From Definition 4.3, (4.20) follows.

Finally, from (4.19) and (4.20), (4.21) is established. Combining (4.21) and Definition 4.4, $D(\tilde{Q}) \leq D(Q)$. 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. □

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

Theorem 4.2. *Let Q be an arbitrary quantizer having discrete output sets $\{Z_1, Z_2, \ldots, Z_N\}$ with N centroids $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N$, there exists a convex quantizer $\tilde{Q}$ with N output sets $\{\tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_N\}$*

$$\begin{aligned}\int_{y\in A} F(\mathbf{p}(\mathbf{x}|y))\mu(y)dy &= \int_{y\in A}[D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)]\mu(y)dy \\ &\leq \int_{y\in B}[D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1) - D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)]\mu(y)dy = \int_{y\in B} F(\mathbf{p}(\mathbf{x}|y)) \end{aligned} \tag{4.16}$$

$$\begin{aligned}&\int_{y\in A} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy - \int_{y\in A} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy + \int_{y\in\{Z_1\cap\tilde{Z}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_2\cap\tilde{Z}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy \\ \leq\ &\int_{y\in B} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy - \int_{y\in B} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy + \int_{y\in\{Z_1\cap\tilde{Z}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_2\cap\tilde{Z}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy \end{aligned} \tag{4.17}$$

$$\begin{aligned}&\Big(\int_{y\in A} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_1\cap\tilde{Z}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy\Big) + \Big(\int_{y\in\{Z_2\cap\tilde{Z}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy + \int_{y\in B} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy\Big) \\ \leq\ &\Big(\int_{y\in B} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in\{Z_1\cap\tilde{Z}_1\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy\Big) + \Big(\int_{y\in\{Z_2\cap\tilde{Z}_2\}} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy + \int_{y\in A} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy\Big) \end{aligned} \tag{4.18}$$

$$\int_{y\in\tilde{Z}_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in\tilde{Z}_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy \leq \int_{y\in Z_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in Z_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy \tag{4.19}$$

$$\int_{y\in\tilde{Z}_1} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_1)\mu(y)dy + \int_{y\in\tilde{Z}_2} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_2)\mu(y)dy \leq \int_{y\in\tilde{Z}_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in\tilde{Z}_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy \tag{4.20}$$

$$\int_{y\in\tilde{Z}_1} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_1)\mu(y)dy + \int_{y\in\tilde{Z}_2} D(\mathbf{p}(\mathbf{x}|y)||\tilde{\mathbf{c}}_2)\mu(y)dy \leq \int_{y\in Z_1} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_1)\mu(y)dy + \int_{y\in Z_2} D(\mathbf{p}(\mathbf{x}|y)||\mathbf{c}_2)\mu(y)dy \tag{4.21}$$

such that $\tilde{Z}_i$ and $\tilde{Z}_j$ are separated by a hyper-plane $H(h_{ij})$ in the space of posterior distribution $\forall i, j$, $p(Z_i) = p(\tilde{Z}_i)$ $\forall i$, and $D(\tilde{Q}) \leq D(Q)$.

Proof. Let Q be an arbitrary quantizer that produces N output sets $\{Z_1, Z_2, \ldots, Z_N\}$. Consider any two output sets Z_i and Z_j, $i \neq j$. Now, let $Y_{ij} = Z_i \cup Z_j$. Based on Theorem 4.1, there is a convex quantizer $\tilde{Q}$ corresponding to a hyper-plane $H(h_{ij})$ separates the K-dimensional points $\mathbf{p}(\mathbf{x}|y), \forall y \in Y_{ij}$ into two sets $\tilde{Z}_i$, and $\tilde{Z}_j$ with $p(Z_i) = p(\tilde{Z}_i)$, $p(Z_j) = p(\tilde{Z}_j)$ and $D(\tilde{Q}) \leq D(Q)$. Specifically, we have:

$$\tilde{Q}(y) = \begin{cases} \tilde{Z}_i & \text{if } y \in Y_{ij} \text{ and } F(\mathbf{p}(\mathbf{x}|y)) \leq h_{ij}, \\ \tilde{Z}_j & \text{if } y \in Y_{ij} \text{ and } F(\mathbf{p}(\mathbf{x}|y)) > h_{ij}, \end{cases} \tag{4.22}$$

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

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

Remark 4.1. *(Optimality) For a given quantizer Q, there exists a convex quantizer $\tilde{Q}$ having the same output probability $\mathbf{p}(z)$ with a lower distortion. This leads to the same cost function $C(\mathbf{p}(z))$ 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 $\beta I(X;Z) - C(\mathbf{p}(z))$ must belong to the set of convex quantizers.*

Remark 4.2. *(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 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 $O(M^{K-1})$ [51]. Typically, $M >> K$, thus searching over the set of convex quantizers is much faster.

Remark 4.3. *(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 [73] for the details.*

Remark 4.4. *(Locally optimal solution) While this chapter 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 [71], [67]. Indeed, using a similar approach in [71], [67], it is possible to show that a locally optimal quantizer Q^* must satisfy:*

$$Q(y) \to Z_i \longleftrightarrow d(y, Z_i) \leq d(y, Z_j), \forall j \neq i,$$

where the "distance" $d(y, Z_i)$ from y to Z_i is defined by:

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

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 [71], [67].

4.6 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 $\mathbf{R}_i \in \mathbb{R}$ 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 $\mathbf{R}_i$'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 separate the regions $\mathbf{R}_i$'s associated with an optimal quantizer.

As an example, if the output is binary, i.e., $Z = \{z_1, z_2\}$, then $t_1, t_2, \ldots, t_S$ divide $\mathbb{R}$ into $S + 1$ contiguous disjoint segments $\mathbf{R}_i = (t_{i-1}, t_i)$, with $t_0 = -\infty$ and $t_{S+1} = \infty$. Each y in $\mathbf{R}_i$ 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 [52], 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 [73], the author extended the single threshold condition in [52] 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 4.1 and Theorem 4.2 to construct an upper bound on the required number of thresholds t_i's for an optimal quantizer.

Theorem 4.1 and Theorem 4.2 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 4.3 formally states this result.

Theorem 4.3. *Let* $\mathbf{R}_l \cup \mathbf{R}_r = \mathbb{R}$ *and* $\mathbf{R}_l \cap \mathbf{R}_r = \emptyset$. *If* $\forall y_l \in \mathbf{R}_l$ *and* $\forall y_r \in \mathbf{R}_r$,

$$\mathbf{a}^T \mathbf{p}(\mathbf{x}|y_r) \geq h, \ \mathbf{a}^T \mathbf{p}(\mathbf{x}|y_l) < h \tag{4.23}$$

for given $h > 0$ *and* $\mathbf{a}$, *then* $\mathbf{R}_l$ *and* $\mathbf{R}_r$ *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:*

$$\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) = h. \tag{4.24}$$

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 $\mathbf{R}_l$ if $\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) < h$ or $\mathbf{R}_r$ if $\mathbf{a}^T \mathbf{p}(\mathbf{x}|y) \geq h$. □

Theorem 4.3 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 4.3, we construct an upper bound on the number of thresholds for additive white Gaussian noise (AWGN) channels.

Theorem 4.4. *For an additive white Gaussian noise (AWGN) channel, the input symbols satisfy* $x_{i+1}-x_i = \delta, i = 1, 2, \ldots N$, *where* δ *is a constant, and* K *quantization levels, the optimal quantizer requires no more than* $\frac{N(N-1)(K-1)}{2}$ *thresholds.*

Proof. Using (4.3), (4.24) can be rewritten by:

$$a_1 \frac{p_1\phi_1(y)}{\sum_{i=1}^{K} p_i\phi_i(y)} + a_2 \frac{p_2\phi_2(y)}{\sum_{i=1}^{K} p_i\phi_i(y)} + \cdots + a_K \frac{p_K\phi_K(y)}{\sum_{i=1}^{K} p_i\phi_i(y)} = h, \tag{4.25}$$

or,

$$\sum_{i=1}^{K} (a_i - h) p_i \phi_i(y) = 0, \tag{4.26}$$

where

$$\phi_i(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-x_i}{\sigma})^2}. \tag{4.27}$$

Since $x_{i+1}-x_i = \delta$, $x_i-x_1 = (i-1)\delta$. Substituting (4.27) into (4.26) and using $x_i-x_1 = (i-1)\delta$,

we have:

$$\sum_{i=1}^{K}(a_i-h)p_i\phi_i(y) \tag{4.28}$$

$$= \sum_{i=1}^{K}(a_i-h)p_i\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-x_i}{\sigma}\right)^2} \tag{4.29}$$

$$= \sum_{i=1}^{K}(a_i-h)p_i\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y^2-2yx_i+x_i^2-2yx_1+2yx_1}{\sigma^2}\right)} \tag{4.30}$$

$$= \sum_{i=1}^{K}(a_i-h)p_i\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y^2-2yx_1+x_i^2-2y(x_i-x_1)}{\sigma^2}\right)} \tag{4.31}$$

$$= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{y^2-2yx_1}{2\sigma^2}}\Big(\sum_{i=1}^{K}(a_i-h)p_ie^{-\frac{x_i^2}{2\sigma^2}}e^{\frac{y(x_i-x_1)}{\sigma^2}}\Big) \tag{4.32}$$

$$= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{y^2-2yx_1}{2\sigma^2}}\Big(\sum_{i=1}^{K}(a_i-h)p_ie^{-\frac{x_i^2}{2\sigma^2}}e^{\frac{y(i-1)\delta}{\sigma^2}}\Big). \tag{4.33}$$

Let $e^{\frac{y}{\sigma^2}}=w$, $\sum_{i=1}^{K}(a_i-h)p_ie^{-\frac{x_i^2}{2\sigma^2}}=b_i$, and since $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{y^2-2yx_1}{2\sigma^2}}\neq 0$, from (4.26) and (4.33), we have:

$$\sum_{i=1}^{K}b_i(w^\delta)^{i-1}=0. \tag{4.34}$$

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, (4.34) 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 $\frac{N(N-1)}{2}$ hyper-plane cuts, a quantizer with $\frac{N(N-1)(K-1)}{2}$ thresholds is sufficient to maximize the mutual information. □

Remark 4.5. *AWGN is one of the most common channels in telecommunication, and the as-*

sumption 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 4.4.

Remark 4.6. *As a consequence of Theorem 4.4, 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 [44], [73]. Furthermore, if the channel is AWGN binary-input binary-output ($N = K = 2$), then a single threshold quantizer is optimal.*

Remark 4.7. *Based on the proposed upper bound on the number of thresholds, a simple 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. Therefore, a simple exhaustive search algorithm using search resolution ϵ would have a time complexity of $O(M^{K-1})$ where $M = \frac{1}{\epsilon}$.*

Remark 4.8. *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.*

4.7 Numerical results

First, we want to refer the reader to the numerical results in [73] which can be considered as special cases for illustrating our Theorem 4.1 and Theorem 4.2. In this section, we only focus on providing some examples to verify the theoretical results in our proposed Theorem 4.4.

Example 4.1. *We consider a binary-input channel having $X = \{x_1 = -10, x_2 = 10\}$ and $\mathbf{p}(x) = [0.6, 0.4]$. X is corrupted by an additive white Gaussian noise having probability density*

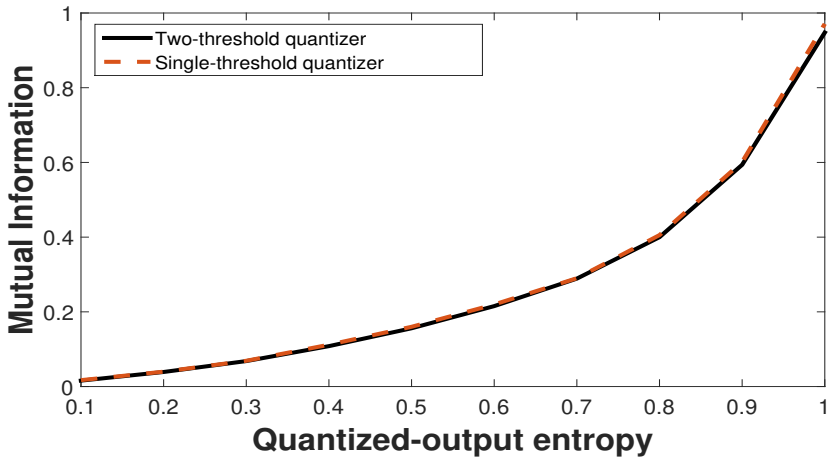


Figure 4.1: Maximum values of mutual information using single-threshold quantizers *vs.* two-threshold quantizers under output constraint $H(Z) \leq \gamma$ for various values of $\gamma = 0.1, 0.2, \ldots, 0.9, 1$.

function $N(\mu = 0, \sigma = 2)$ *with* $\phi_1(y) = N(-10, 2)$ *and* $\phi_2(y) = N(10, 2)$. *Next, we want to design an optimal quantizer* Q *that quantizes* $y \in \mathbb{R}$ *to a binary output* $Z = \{z_1, z_2\}$ *such that the mutual information* $I(X;Z)$ *is maximized while* $H(Z) \leq \gamma$ *for a given* γ.

Since $N = K = 2$, *Theorem 4.4 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. 4.1, 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 4.4.*

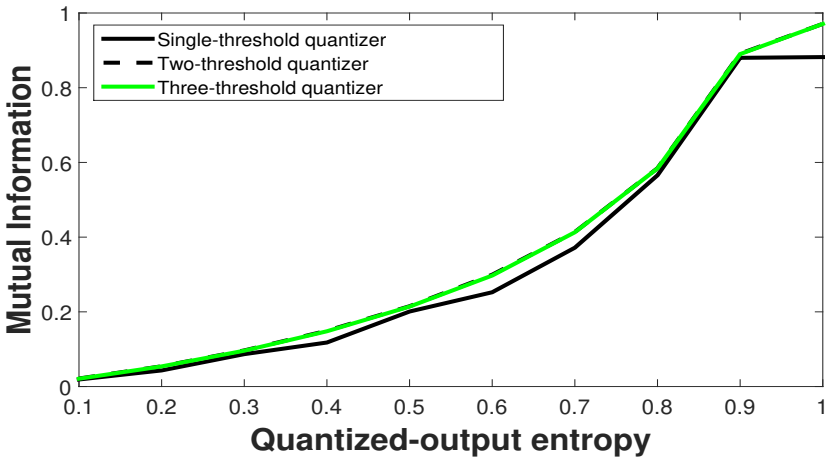


Figure 4.2: Maximum values of mutual information using single-threshold quantizers, two-threshold quantizers and three-threshold quantizers under output constraint $H(Z) \leq \gamma$ for various values of $\gamma = 0.1, 0.2, \ldots, 0.9, 1$.

Example 4.2. *We consider a channel having input* $X = \{x_1 = -10, x_2 = 0, x_3 = 10\}$ *and* $\mathbf{p}(x) = [0.3, 0.4, 0.3]$. *Similar to Example 4.1,* X *is corrupted by an additive white Gaussian noise having probability density function* $N(\mu = 0, \sigma = 1)$ *with* $\phi_1(y) = N(-10, 1)$, $\phi_2(y) = N(0, 1)$, *and* $\phi_3(y) = N(10, 1)$. *We want to design an optimal quantizer* Q *that quantizes* $y \in \mathbb{R}$ *to a binary quantized-output* $Z = \{z_1, z_2\}$ *to maximize* $I(X; Z)$ *while* $H(Z) \leq \gamma$, *for a given* γ.

Based on Theorem 4.4, 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. 4.2, respectively. As seen,* $I(X; Z)$ *provided by two-threshold quantizers are slightly larger than that of three-threshold quan-*

tizers. 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 4.4.

4.8 Conclusion

In this chapter, 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. For small numbers of inputs and outputs, using this upper bound, it is feasible to use an exhaustive search with polynomial time complexity to find an optimal solution.

Chapter 5: Capacity Achieving Quantizer Design for Binary Channels

5.1 Introduction

A primary goal of a communication system is to transmit the information reliably and fast over an error-prone channel. The fastest rate with a vanishing error for a given channel is equal to the maximum mutual information $I(X;Z)$ between two random variables X and Z used to model the input and the output of channel. For a given discrete memoryless channel (DMC) specified by a channel matrix $\mathbf{A}$, it is well-known that the mutual information is a concave function in the input probability mass function $\mathbf{p}_X$ [3]. Consequently, determining the capacity achieving optimal input distribution $\mathbf{p}_X^*$ that maximizes $I(X;Z)$ for a given $\mathbf{A}$ is not difficult using existing convex optimization algorithms or other iterative algorithms [4]. Furthermore, under some special conditions on $\mathbf{A}$, it is possible to obtain closed-form expressions for the capacities of many DMCs [3], [76], [9], [77].

On the other hand, rather than using a given channel matrix $\mathbf{A}$, one assumes a given input distribution $\mathbf{p}_X$. The goal is to design an optimal quantizer Q^*, which is equivalent to selecting an optimal channel matrix $\mathbf{A}^*$ subject to a certain structure that maximizes the mutual information between the input X and the quantized output Z [44, 65, 78–80]. We note that this is not the same as designing a quantizer that achieves the capacity since the input distribution $\mathbf{p}_X$ is given. Our goal is to determine the optimal quantizer Q^* together with the optimal input distribution $\mathbf{p}_X^*$ that achieves the channel capacity. To the best of our knowledge, this problem still remains a hard problem for a general setting [45, 57, 59, 81]. In [59], Singh et al. provided an algorithm for multilevel quantization, which gave near-optimal results. In [57], the author proposed a heuristic

near-optimal quantization algorithm which alternatively maximizes the mutual information for a given quantizer and minimizes the probability of error for a given input distribution. However, this algorithm only works well when the signal-to-noise ratio of the channel is high. For 2-level (1-bit) quantization of general additive channels, Mathar and Dorpinghaus proved that the optimal mutual information could be achieved by using an input distribution between two support points [45]. However, it is worth noting that the result in [45] is limited for single threshold quantizers and the truly optimal quantizer may contain more than one threshold [52]. In [81], the author gave a near-optimal algorithm to find the optimal value of mutual information for binary input and an arbitrary number of the quantized output, however, this algorithm may declare a failure outcome. There are also several recent works on finding the channel capacity for Gaussian channels with quantized output. In [82], Vu et al. investigated the problem of designing the optimal signaling schemes together with capacity-achieving input distribution for Gaussian channels under the assumption of a low-resolution output quantization. In [83], Ranjbar et al. constructed the capacity region and capacity-achieving signaling schemes for 1-bit quantization with two users communicating in Rayleigh-fading channels. These works focus on finding the optimal input distribution for a pre-specified channel (Gaussian and Rayleigh) and under a given quantization scheme. In contrast, the work in this chapter is more general as our focus is on obtaining both an optimal quantization scheme and optimal input distribution simultaneously. Furthermore, our results can be applied to any communication channel specified by an arbitrary conditional density of the received output given the transmitted input.

In this chapter, we consider a special case where the channel matrix $\mathbf{A}$ is a 2×2 matrix. In particular, we consider a communication channel with a binary input X being distorted by a given arbitrary continuous-valued noise which results in a continuous-valued signal Y at the receiver. A quantizer Q is used to quantize Y back to a binary output Z. Our goal is to determine the optimal quantizer Q^*, and therefore, an induced optimal $\mathbf{A}^*$ that achieves the capacity. Importantly, we

do not assume that $\mathbf{p}_X$ is given. Rather, after the optimal $\mathbf{A}^*$ is determined, the optimal $\mathbf{p}_X^*$ then can be obtained using any classic method. The main contributions of this chapter include the new lower bound and upper bound of the capacity in terms of the quantization parameters, together with the structure of the associated channel matrix. Based on these, we propose an efficient algorithm for finding Q^*.

5.2 Problem description

We consider the setting shown in Fig. 5.1. The binary input modeled as a random variable $X \in \{0, 1\}$, is transmitted over a channel that distorts X into a continuous valued signal modeled as a random variable Y at the receiver. The channel distortion is modeled by a conditional density of Y given X: $f_{Y|X}(y|x)$. To recover X, the receiver uses a quantizer Q that quantizes Y to a binary signal $Z \in \{0, 1\}$. Formally,

$$Q(y) = \begin{cases} z = 0 & \text{if } y \in \mathbb{H}, \\ z = 1 & \text{if } y \in \bar{\mathbb{H}}, \end{cases} \tag{5.1}$$

where $\mathbb{H} \cap \bar{\mathbb{H}} = \emptyset$ and $\mathbb{H} \cup \bar{\mathbb{H}} = \mathbb{R}$. For a given conditional density $f_{Y|X}(y|x)$, our goal is to design an optimal quantizer Q^* together with an optimal input distribution $\mathbf{p}_X^*$ such that the mutual information $I(X; Z)$ between X and Z is maximized:

$$Q^*, \mathbf{p}_X^* = \arg\max_{Q, \mathbf{p}_X} I(X; Z). \tag{5.2}$$

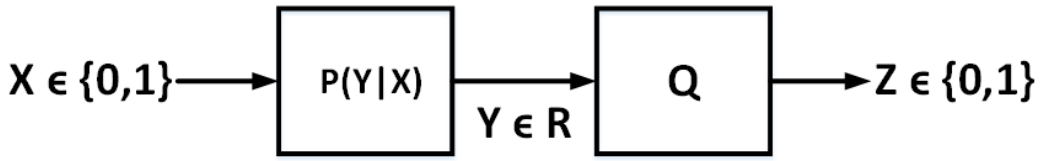


Figure 5.1: A binary input $X = \{0, 1\}$ is transmitted over a noisy channel which results in a continuous-valued $y \in Y$ at the receiver. The receiver attempts to recover X by quantizing Y to a discrete binary signal $Z = \{0, 1\}$.

5.3 Preliminaries

We consider the setting in Fig. 5.1. Let $\mathbf{p}_X = (p_0, p_1)$ is the input probability mass distribution and $f_{Y|X}(y|x)$ is the conditional density function of Y given X. For given $f_{Y|X}(y|x)$, let $\phi_0(y) = f_{Y|X}(y|x = 0)$ and $\phi_1(y) = f_{Y|X}(y|x = 1)$ and define:

$$u(y) = \frac{p_1\phi_1(y)}{p_0\phi_0(y) + p_1\phi_1(y)}. \tag{5.3}$$

Definition 5.1. *A binary quantizer Q_u is called a convex quantizer if it has the following structure:*

$$Q_u(y) = \begin{cases} z = 0 & \text{if } u(y) \leq u, \\ z = 1 & \text{if } u(y) > u, \end{cases} \tag{5.4}$$

where $0 < u < 1$.

Burshtein et al. [51] and Kurkoski and Yagi [52] showed that the optimal binary quantizer is indeed a convex quantizer as stated in Theorem 5.1 below.

Theorem 5.1. *[51], [52] For a given p_0 and p_1, the optimal binary quantizer that maximizes the mutual information $I(X; Z)$ is a convex quantizer Q_{u^*} for some optimal threshold u^*.*

We should make an important remark about Theorem 5.1.

Remark 5.1. *Q_{u^*} is not a capacity achieving quantizer even though it maximizes $I(X;Z)$. This is because Q_{u^*} assumes a given input distribution $\mathbf{p}_X$. On the other hand, our goal is to find the capacity achieving quantizer Q^* which maximizes $I(X;Z)$ over all the possible $\mathbf{p}_X$. A straightforward way of applying Theorem 5.1 to find Q^* is to search over all possible values of p_0, p_1, and u^* that maximizes $I(X;Z)$. This is however still a 2-dimensional search.*

Next, instead of given $\mathbf{p}_X$, suppose a quantizer Q is given, we want to determine the capacity $C = \max_{\mathbf{p}_X} I(X;Z)$. The given Q induces a channel matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where $a_{12} = 1 - a_{11}$ and $a_{21} = 1 - a_{22}$. We will show how $\mathbf{A}$ is related to Q shortly. The capacity of this binary DMC is given in Theorem 5.2 below [76].

Theorem 5.2. *[76] The capacity of a binary DMC for a given channel matrix $\mathbf{A}$ is:*

$$\begin{aligned} C &= \log_2 \Big(2^{-\frac{a_{22}H(a_{11}) + (a_{11}-1)H(a_{22})}{a_{11}+a_{22}-1}} \\ &+ 2^{-\frac{(a_{22}-1)H(a_{11}) + a_{11}H(a_{22})}{a_{11}+a_{22}-1}} \Big), \end{aligned} \tag{5.5}$$

where $H(w) = -w\log_2(w) - (1-w)\log_2(1-w)$.

We will use Theorems 5.1 and 5.2 to describe a more efficient procedure for finding the capacity achieving Q^*.

5.4 Design of Capacity Achieving Quantizer

Theorem 5.3 below is a variant of Theorem 5.1 which will be used to design a capacity achieving quantizer.

Theorem 5.3. *(Structure of optimal quantizer)*

For given $\phi_0(y)$ and $\phi_1(y)$, define:

$$r(y) = \frac{\phi_0(y)}{\phi_1(y)}. \tag{5.6}$$

Let Q_r be a convex quantizer with the following structure:

$$Q_r(y) = \begin{cases} 0 & \text{if } r(y) \geq r, \\ 1 & \text{if } r(y) < r, \end{cases} \tag{5.7}$$

for some $0 < r < \infty$, then there exists a capacity achieving convex quantizer Q_{r^*} for some optimal threshold r^*.

Proof. For any $\mathbf{p}_X$, we have:

$$u(y) = \frac{p_1\phi_1(y)}{p_0\phi_0(y)+p_1\phi_1(y)} = \frac{1}{\frac{p_0}{p_1}\frac{\phi_0(y)}{\phi_1(y)}+1} = \frac{1}{\frac{p_0}{p_1}r(y)+1}. \tag{5.8}$$

Thus,

$$r(y) = \frac{p_1(1-u(y))}{p_0u(y)}. \tag{5.9}$$

Now using Theorem 5.1, and writing $u(y)$ in terms of $r(y)$, we obtain:

$$r^* = \frac{p_1(1-u^*)}{p_0 u^*}. \tag{5.10}$$

Furthermore, for any valid $p_0 > 0$, it is straightforward to show that $0 < r^* < \infty$. It is important to note that p_0 and p_1 need not to be given, even though they are related to $r(y)$ through (5.9) for some p_0 and p_1. Instead, $r(y)$ is defined as $r(y) = \dfrac{\phi_0(y)}{\phi_1(y)}$. If there is a method to find the optimal r^* directly, then the corresponding p_0^* and p_1^* can be found based on r^*. Importantly, since Q_{r^*} maximizes $I(X;Z)$ (by Theorem 5.1) without given p_0 and p_1, Q_{r^*} is a capacity achieving quantizer. □

Remark 5.2. *The use of $r(y)$ in Theorem 5.3 rather than $u(y)$ in Theorem 5.1 is an important step in designing a capacity achieving quantizer. $r(y)$ as defined in (5.6), does not depend on p_0 and p_1. Therefore, to find a capacity achieving quantizer, one can employ an exhaustive search to find the optimal threshold r^*. Specifically, for each value of the threshold r, a quantizer Q can be constructed based on Theorem 5.3 in which one compares $r(y)$ with r. This comparison does not need p_0 and p_1. On the other hand, using $u(y)$ and search for the optimal u^* in Theorem 5.1, one is required to know p_0 and p_1 since $u(y)$ is defined in terms of p_0 and p_1.*

We now derive the channel matrix $\mathbf{A}$ for a given quantizer Q_r in Theorem 5.3. Define:

$$\mathbb{H}_r = \{y : r(y) = \frac{\phi_0(y)}{\phi_1(y)} \geq r\}, \tag{5.11}$$

$$\bar{\mathbb{H}}_r = \{y : r(y) = \frac{\phi_0(y)}{\phi_1(y)} < r\}. \tag{5.12}$$

Thus,

$$Q_r(y) = \begin{cases} z = 0 & \text{if } y \in \mathbb{H}_r, \\ z = 1 & \text{if } y \in \bar{\mathbb{H}}_r. \end{cases} \tag{5.13}$$

The channel matrix $\mathbf{A}$ that corresponds to quantizer Q_r is:

$$\mathbf{A} = \begin{bmatrix} a_{11}(r) & a_{12}(r) \\ a_{21}(r) & a_{22}(r) \end{bmatrix},$$

where

$$a_{11}(r) = \int_{y \in \mathbb{H}_r} \phi_0(y) dy, \tag{5.14}$$

$$a_{22}(r) = \int_{y \in \bar{\mathbb{H}}_r} \phi_1(y) dy, \tag{5.15}$$

and $a_{12}(r) = 1 - a_{11}(r)$, $a_{21}(r) = 1 - a_{22}(r)$.

Using Theorem 5.2 and Theorem 5.3, the capacity in (5.5) is a function of r:

$$\begin{aligned} C(r) &= \log_2 \Big(2^{-\frac{a_{22}(r)H(a_{11}(r)) + (a_{11}(r)-1)H(a_{22}(r))}{a_{11}(r) + a_{22}(r) - 1}} \\ &+ 2^{-\frac{(a_{22}(r)-1)H(a_{11}(r)) + a_{11}(r)H(a_{22}(r))}{a_{11}(r) + a_{22}(r) - 1}} \Big). \end{aligned} \tag{5.16}$$

We note that each value of r corresponds to a different channel matrix $\mathbf{A}$ associated with a different Q_r. Therefore, based on (5.16), an exhaustive search can be used to find r^* that maximizes $C(r)$. This is a one-dimensional search on r which is more efficient than searching for u, p_0, and p_1 as discussed earlier. Furthermore, we will derive an upper and lower bound on r to increase the search efficiency. Lemma 5.1 below describes the structure of the channel matrix that corresponds to a convex quantizer Q_r.

Lemma 5.1. *(Structure of the channel matrix induced by Q_r)*

For $\forall\ r \in (0, +\infty)$,

(1) $a_{11}(r) \in (0, 1)$ and is a monotonic decreasing function.

(2) $a_{22}(r) \in (0, 1)$ and is a monotonic increasing function.

(3) 1 ¡ $a_{11}(r) + a_{22}(r) \leq a_{11}(1) + a_{22}(1)$.

Proof. Please see the proof in Appendix 5.7.1. □

Theorem 5.4. *(Capacity bounds)*

Define $\delta = a_{11}(1)+a_{22}(1)$, then the maximum capacity $C(r^*)$ over all possible channel matrices induced by all convex quantizers Q_r is bounded by:

$$1 - H(\frac{2-\delta}{2}) \leq C(r^*) \leq \log_2 \delta. \tag{5.17}$$

Proof. From Lemma 5.1, $\forall r$, we have:

$$a_{11}(r) + a_{22}(r) > 1 = a_{11}(r) + a_{12}(r), \tag{5.18}$$

$$a_{11}(r) + a_{22}(r) > 1 = a_{21}(r) + a_{22}(r). \tag{5.19}$$

Thus,

$$a_{22}(r) > a_{12}(r), \tag{5.20}$$

$$a_{11}(r) > a_{21}(r). \tag{5.21}$$

Upper bound: The Boyd-Chiang's upper bound [13] of the channel capacity associated with

a given channel matrix $\mathbf{A}$ is:

$$C_{\mathbf{A}} \leq \log_2(\sum_j \max_i a_{ij}). \tag{5.22}$$

For a binary channel associated with a convex quantizer Q_r, using (5.22), we have:

$$\begin{aligned} C(r) &\leq \log_2\big(\sum_{j=1} \max_i a_{ij}(r)\big) = \log_2\big(a_{11}(r)+a_{22}(r)\big) && (5.23)\\ &\leq \log_2\big(a_{11}(1)+a_{22}(1)\big) = \log_2\delta, && (5.24) \end{aligned}$$

where (5.23) is due to (5.20) and (5.21), (5.24) is due to (3) in Lemma 5.1. Since the upper bound in (5.24) holds for every r, it must hold for r^*.

Lower bound: Recall that the Fano's inequality [3] with alphabet size of $|X| = 2$ is:

$$H(X|Z) \leq H(p_e) + p_e \log(|X| - 1) = H(p_e), \tag{5.25}$$

where p_e is the probability of error when transmitting a signal over the channel and using a quantizer Q_r for recovering the signal. Next, using the uniform input distribution $\mathbf{p}_X$ i.e., $p_0 = p_1 = 1/2$ and the convex quantizer Q_1 ($r = 1$), we have:

$$\begin{aligned} p_e &= p_0 a_{12}(1) + p_1 a_{21}(1) = \frac{1}{2}(a_{12}(1) + a_{21}(1)) && (5.26)\\ &= \frac{1}{2}(2 - a_{11}(1) - a_{22}(1)) = \frac{2-\delta}{2}, && (5.27) \end{aligned}$$

and $H(X) = 1$.

Now, since the maximum capacity $C(r^*)$ is at least as large as the mutual information using

$p_0 = p_1 = 1/2$, and $r = 1$, from (5.25) and (5.27), we have:

$$\begin{aligned} C(r^*) &\geq I(X;Z) = H(X) - H(X|Z) && (5.28) \\ &\geq H(X) - H(p_e) = 1 - H(\frac{2-\delta}{2}). && (5.29) \end{aligned}$$

□

Remark 5.3. *We note that the lower bound reaches to the upper bound when $\delta \to 2$ or $\delta \to 1$ as illustrated in Fig. 5.2. Moreover, a larger value of δ implies a smaller overlapped area between the noise density $\phi_0(y)$ and $\phi_1(y)$ or a higher probability of correct decoding. In an additive channel with an identical noise for transmitting symbols 0 and 1, $\phi_0(y)$ and $\phi_1(y)$ are shifted versions of each other. The larger shift results in a larger value of δ and a higher probability of correct decoding. Thus, our bounds are tighter for low noise regimes. The upper and lower bounds as functions of δ are visualized in Fig. 5.2.*

Also, as an extension, if the input size is more than two, then using the identical proof, the δ in the upper bound in Theorem 5.4 is:

$$\delta = \int_{y \in Y} \max_i \phi_i(y) dy, \tag{5.30}$$

where $\phi_i(y) = f_{Y|X}(y|x_i), i = 1, 2, \ldots, N$ with N being the size of the input alphabet.

Remark 5.4. *From (5.20) and (5.21), $a_{22}(r) > a_{12}(r)$ and $a_{11}(r) > a_{21}(r)$. Thus, using the equations (1) and (2) in [84], it is possible to show that the optimal input distribution has to satisfy: $1/e = 0.3679 \leq p_0^* \leq \frac{1}{2} \leq p_1^* \leq 1 - 1/e = 0.6321$. This fact can help to reduce the running time of a 2-dimensional exhaustive search algorithm over both p_0 and $u(y)$ variables in order to determine the channel capacity. In addition, from Theorem 2 in [84], the mutual information*

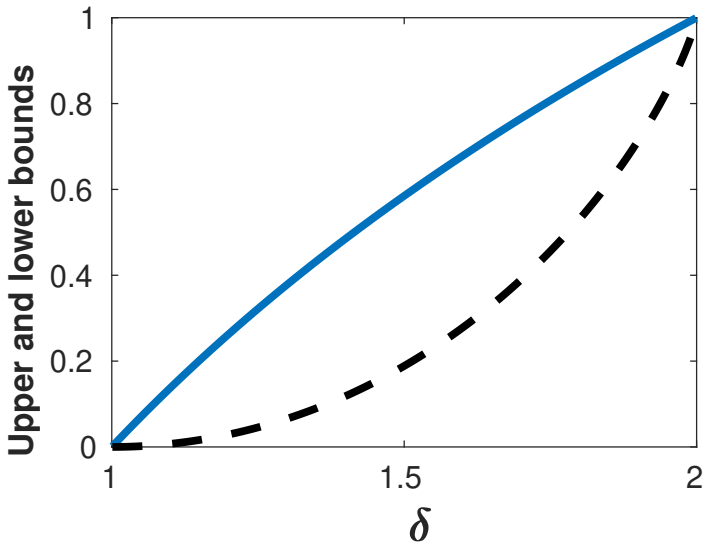


Figure 5.2: Upper bound and lower bound of channel capacity as functions of δ.

induced by using a uniform input distribution is at least 94.21% *of the channel capacity for any given channel matrix* $\mathbf{A}$. *Therefore, by using a uniform input distribution together with performing a 1-dimensional exhaustive search over* $u(y)$, *it is possible to achieve at least* 94.21% *of the truly channel capacity.*

Next, we will use Theorem 5.4 to narrow down the range to search for the optimal r^*. We have the following theorem.

Theorem 5.5. *(Bound on optimal* r^**)*

Let $0 < v \leq \frac{1}{2}$ *be a positive number such that:*

$$H(v) = H(1-v) = 1 - H(\frac{2-\delta}{2}). \tag{5.31}$$

If Q_{r^*} is optimal, then:

$$a_{11}(r^*) \geq v, \tag{5.32}$$

$$a_{22}(r^*) \geq v. \tag{5.33}$$

Furthermore, $r^* \in [r_2, r_1]$ where $a_{11}(r_1) = a_{22}(r_2) = v$.

Proof. Suppose that a quantizer Q_r produces $H(Z)$, and

$$H(Z) \leq 1 - H(\frac{2-\delta}{2}) = H(v) = H(1-v) \tag{5.34}$$

for some $v \in (0, 0.5]$. Since $1 - H(\frac{2-\delta}{2}) \geq H(Z) \geq I(X;Z)$, based on the lower bound of Theorem 5.4, Q_r cannot be an optimal quantizer.

We will show that if:

$$a_{11}(r) < v, \tag{5.35}$$

$$a_{22}(r) < v, \tag{5.36}$$

then Q_r is suboptimal.

Since the binary entropy is symmetric i.e., $H(v) = H(1-v)$ and $v \leq 1/2$, then $v \leq 1/2 \leq 1-v$. From (5.21),

$$\begin{aligned} p(Z=0) &= p_0 a_{11}(r) + p_1 a_{21}(r) \geq p_0 a_{21}(r) + p_1 a_{21}(r) \\ &= a_{21}(r) = 1 - a_{22}(r). \end{aligned} \tag{5.37}$$

Since the binary entropy is monotonically increased over $[0, 0.5]$ and monotonically decreased

over $[0.5, 1]$, if $1 - a_{22}(r) > 1 - v$ or $a_{22}(r) < v$, then:

$$H(Z) = H(p(Z = 0)) < H(1 - v) = 1 - H(\frac{2 - \delta}{2}). \tag{5.38}$$

From (5.34) and (5.38), a quantizer that produces $a_{22}(r) < v$ is not the optimal one. Therefore, $a_{22}(r^*) \geq v$. A similar proof can be constructed to show that $a_{11}(r^*) \geq v$.

Next, due to $\delta > 1$ (Lemma 5.1-(3)), we have $0 < 1 - H(\frac{2 - \delta}{2}) \leq 1$. Therefore, there exists $v \in (0, 1/2]$ that satisfies (5.31). From Lemma 5.1, there exists two positive numbers r_1 and r_2 such that $a_{11}(r_1) = v$ and $a_{22}(r_2) = v$. Moreover, $a_{11}(r)$ and $a_{22}(r)$ are monotonic decreasing and increasing functions, respectively (Lemma 5.1), thus $r^* \in [r_2, r_1]$. □

Exhaustive search. The proposed algorithm is to search for r in the range of $[r_2, r_1]$. Since $a_{11}(r)$ and $a_{22}(r)$ are monotonic decreasing and increasing functions, finding r_1 and r_2 such that $a_{11}(r_1) = v$ and $a_{22}(r_2) = v$ can be performed efficiently using existing root-finding algorithms, for example, the bisection method. For each value of r in the range $[r_2, r_1]$, we determine the channel matrix $\mathbf{A}$ then use (5.16) to compute the corresponding capacity. The algorithm outputs the largest mutual information in this range together with r^*. From r^*, Q_{r^*} can be found. Next, based on [76], p_0^* can be obtained as:

$$\begin{aligned} p_0^* &= a_{21}(r^*)[a_{21}(r^*) - a_{11}(r^*)]^{-1} \\ &\quad - [a_{21}(r^*) - a_{11}(r^*)]^{-1}\Big[1 + 2^{\dfrac{H(a_{21}(r^*)) - H(a_{11}(r^*))}{a_{21}(r^*) - a_{11}(r^*)}}\Big]^{-1}, \end{aligned}$$

where $H(x) = -[x \log_2(x) + (1 - x) \log_2(1 - x)]$ is the binary entropy function.

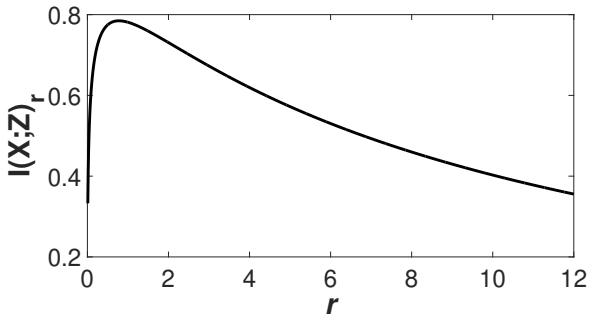


Figure 5.3: Mutual information $I(X;Z)_r$ as a function of r.

5.5 Numerical Results

Consider a channel having $\phi_0(y) = N(\mu_0 = -1, \sigma_0 = 0.5)$ and $\phi_1(y) = N(\mu_1 = 1, \sigma_1 = 0.6)$. One wants to find the optimal quantizer together with the input distribution such that the mutual information is maximized.

Now, for $r = 1$, $\delta = 1.9299$, $v = 0.2316$, and $r_2 = 0.11$, $r_1 = 11.08$. By performing an exhaustive search with the resolution $\epsilon = 0.01$ over $r \in [r_2, r_1]$, the optimal of mutual information is $I(X;Z)^* = 0.7847$ at $r^* = 0.78$. The corresponding channel capacity upper and lower bounds using Theorem 5.4 are 0.9479 and 0.7787, respectively. Fig. 5.3 illustrates $I(X;Z)_r$ as a function of r.

5.6 Conclusion

We presented both a new lower bound and a new upper bound on the capacity in terms of quantization parameters and the structure of the associated channel matrix for binary quantization channel. Based on these theoretical results, we propose an efficient algorithm for finding the optimal quantizer.

5.7 Appendix

5.7.1 Proof of Lemma 5.1

Proof for (1) and (2). From the definition in (5.14), $a_{11}(r)$ represents the quantized bit "0" which is the integral of $\phi_0(y)$ over the set of y such that $r(y) \geq r$. Thus, if r increases, the set of y reduces. Since $\phi_0(y) \geq 0$ and the set of y reduces, $a_{11}(r)$ must decrease.

Moreover, if $r \to 0$, $r(y) \geq r \ \forall y$ then $a_{11}(r) \to 1$. On the other hand, if $r \to +\infty$, then $r(y) < r \ \forall y$ and $a_{11}(r) \to 0$. Thus, $a_{11}(r) \in (0, 1)$.

A similar proof can be constructed for $a_{22}(r)$.

Proof for (3). From the definition of $r(y)$, $\mathbb{H}_r$ and $\bar{\mathbb{H}}_r$, we have $\dfrac{\phi_0(y)}{\phi_1(y)} \geq r, \forall y \in \mathbb{H}_r$ and $\dfrac{\phi_0(y)}{\phi_1(y)} < r, \forall y \in \bar{\mathbb{H}}_r$. Next, we consider two possible cases: $r > 1$ and $r \leq 1$. In both cases, we show that $a_{11}(r) + a_{22}(r) > 1$.

• If $r > 1$, $\phi_0(y) > \phi_1(y)$ for $\forall\ y \in \mathbb{H}_r$. Therefore,

$$\begin{aligned} a_{11}(r) + a_{22}(r) &= \int_{y \in \mathbb{H}_r} \phi_0(y)dy + \int_{y \in \bar{\mathbb{H}}_r} \phi_1(y)dy \\ &> \int_{y \in \mathbb{H}_r} \phi_1(y)dy + \int_{y \in \bar{\mathbb{H}}_r} \phi_1(y)dy \\ &= 1. \end{aligned} \tag{5.39}$$

• If $r \leq 1$, $\phi_1(y) > \phi_0(y)$ for $\forall\ y \in \bar{\mathbb{H}}_r$. Therefore,

$$\begin{aligned} a_{11}(r) + a_{22}(r) &= \int_{y \in \mathbb{H}_r} \phi_0(y)dy + \int_{y \in \bar{\mathbb{H}}_r} \phi_1(y)dy \\ &> \int_{y \in \mathbb{H}_r} \phi_0(y)dy + \int_{y \in \bar{\mathbb{H}}_r} \phi_0(y)dy \\ &= 1. \end{aligned} \tag{5.40}$$

Combining (5.39) and (5.40), $a_{11}(r) + a_{22}(r) > 1$, $\forall r$.

Next, we show that $a_{11}(1) + a_{22}(1) \geq a_{11}(r) + a_{22}(r)$, $\forall r$. Indeed, from the definition of $\mathbb{H}_1$ and $\bar{\mathbb{H}}_1$, $\dfrac{\phi_0(y)}{\phi_1(y)} \geq 1, \forall y \in \mathbb{H}_1$ and $\dfrac{\phi_0(y)}{\phi_1(y)} < 1, \forall y \in \bar{\mathbb{H}}_1$. Thus,

$$\phi_0(y) \geq \phi_1(y), \forall y \in \mathbb{H}_1, \tag{5.41}$$

$$\phi_0(y) < \phi_1(y), \forall y \in \bar{\mathbb{H}}_1. \tag{5.42}$$

From the definition of $a_{11}(r)$ and $a_{22}(r)$ in (5.14) and (5.15),

$$\begin{aligned}
a_{11}(r) + a_{22}(r) &= \int_{y \in \mathbb{H}_r} \phi_0(y)dy + \int_{y \in \bar{\mathbb{H}}_r} \phi_1(y)dy \\
&\leq \int_{y \in \mathbb{H}_r} \max\big(\phi_0(y), \phi_1(y)\big)dy \\
&\quad + \int_{y \in \bar{\mathbb{H}}_r} \max\big(\phi_0(y), \phi_1(y)\big)dy \\
&= \int_{y \in \mathbb{H}_r \cup \bar{\mathbb{H}}_r = \mathbb{R}} \max\big(\phi_0(y), \phi_1(y)\big)dy \\
&= \int_{y \in \mathbb{H}_1} \max\big(\phi_0(y), \phi_1(y)\big)dy \\
&\quad + \int_{y \in \bar{\mathbb{H}}_1} \max\big(\phi_0(y), \phi_1(y)\big)dy \\
&= \int_{y \in \mathbb{H}_1} \phi_0(y)dy + \int_{y \in \bar{\mathbb{H}}_1} \phi_1(y)dy \qquad (5.43) \\
&= a_{11}(1) + a_{22}(1) \qquad (5.44) \\
&= \delta, \qquad (5.45)
\end{aligned}$$

where (5.43) due to (5.41) and (5.42), (5.44) and (5.45) due to the definitions of $a_{11}(r)$, $a_{22}(r)$ and δ. The equality happens if $r = 1$.

Chapter 6: Bounded Guaranteed Algorithm for Concave Impurity Minimization Via Maximum Likelihood

6.1 Introduction

Partitioning plays a key role in many scientific and engineering disciplines. It is a key building block in many popular algorithms such as clustering and classification in machine learning. In signal processing and communications, partitioning algorithms, which are usually called quantization, aim to minimize the distortion or maximize the mutual information between the original signal and the quantized signals. A partitioning algorithm divides a set of M N-dimensional elements into K disjoint subsets or partitions. Often, the quality of the resulted partitions is measured by the amount of impurity in each partition, the smaller impurity the higher quality of the partitions. Typically, the amount of impurity is measured by a real-valued function over the resulted partitions.

When the elements can be modeled as the outcomes of an underlying probabilistic model, it makes sense to consider some statistical measures such as the average or the variance of the impurity. Naturally, a partitioning algorithm in this scenario might classify the elements into different partitions using probability distributions, rather than the values of the elements. For example, let us consider a popular impurity function using the Shannon entropy [85], [53], [74]. Consider a set whose elements are outcomes of a random variable W. A large entropy of a random variable implies that the elements are likely to be different, i.e., the set has a high level of non-homogeneity or "impurity". For a given K, a K-optimal partition algorithm divides the

original set into K subsets such that the weighted sum of entropies in each subset is minimal. Since entropy is a concave function of the probability mass vector/function (pmf), not the values of a random variable, the partition algorithms, in this case, work directly with the N-dimensional probability mass vector. In contrast, the popular k-means algorithms do not assume an underlying probabilistic model for how the elements come about. Thus, the elements are clustered using a distance measure (typically Euclidean) which is a function of the actual values of the elements.

In general, for a given impurity measure specified by a function over the partitions, finding the minimum impurity partitions is an NP-hard problem. Since the number of possible partitions is K^M, an exhaustive search over all the possible partitions to find a minimum partition has the complexity of $O(K^M)$ which quickly becomes impractical for many applications with modest values of K and M. To that end, many approximate algorithms with polynomial time complexity have been proposed, but few provide bounded guarantee [85], [53], [51], [86]. Many of these algorithms exploit the concavity of the impurity function to speed up the running time. For example, in [86], an algorithm is proposed to find the optimal partition using a concave impurity function with the computational complexity of $O(M \log M)$ for binary classification tasks ($K = 2$). Burshtein et al. [51] and Coppersmith et al. [53] provided algorithms and theoretical analysis for the partitioning problem for a general concave impurity function called "frequency-weighted impurity". These "frequency-weighted impurity" are concave functions over its second argument. Two popular impurity functions the Gini index [86] and Shannon entropy [85] belong to this class of frequency-weighted impurity. Burshtein et al. and Coppersmith et al. showed that an optimal frequency-weighted impurity partition is separated by hyperplane cuts in the space of probability distributions. Based on this insight, they also proposed polynomial time algorithms to determine the optimal partitions [53], [51], [68]. Based on the work of Burshtein et al., Kurkoski and Yagi proposed an algorithm to find the globally optimal partition that minimizes entropy impurity in $O(M^3)$ when $N = 2$ [44].

Although many heuristic algorithms have been proposed, there are few results in finding algorithms that provides a bounded guarantee on the performance. To fill this gap, recently Laber et al. [56] constructed a 2-approximation algorithm with the computational complexity of $O(2^N M \log M)$ for binary partition ($K = 2$). In other words, Laber et al. showed that the impurity achieved by their algorithm is at most a factor of 2 away from the true optimal impurity. The complexity can be further reduced to $O(MN + M \log M)$ at the expense of increasing the approximation factor from 2 to $3+\sqrt{3}$. We also note that the algorithm in [56] is closely related to the well established Twoing method that was already proposed in [86]. Moreover, the application of the algorithm in [56] is somewhat limited due to the requirement of $K = 2$. As the extension of the work in [56], Cicalese et al. [1] proposed a heuristic algorithm for $K > 2$. The algorithm can achieve $\log^2(\min\{N, K\})$-approximation for the entropy impurity, 3-approximation for the Gini index impurity if $K < N$ and 2-approximation for the Gini index impurity if $K \geq N$. It is the first constant factor algorithm for clustering based on minimizing entropy impurity that does not rely on assumptions about the input data. Our analysis in Appendix 6.8.1 shows that the complexity of the algorithm in [1] is polynomial due to its most time consuming step is based on dynamic programming technique in [44] with the time complexity of $O(M^3)$. Using the SMAWK algorithm [87], the dynamic programming step of the algorithm in [1] can be further reduced to $O(M \log M)$. The analysis of the algorithm in [1] together with a suggested method for reducing the computational complexity is shown in Appendix 6.8.1.

The partitioning algorithm also tightly relates to clustering algorithms which group M probability distributions into K clusters in such a way to minimize a certain distance. For example, minimizing entropy impurity partition is equivalent to finding the optimal clusters that minimizes the Kullback-Leibler (KL) distance [54]. Generally, the local optimal solution minimizing impurity partition can be found based on the famous k-means algorithm with a suitable distance [53], [68], [54]. The detail of these algorithms finding the local optimal solution for entropy

and Gini index impurity are discussed at Appendix 6.8.10. Thus, the results about approximation for clustering with KL-divergence in [88], [89] can be applied to find a good partition that minimizes the entropy impurity. For example, in [88], Sra et. al. showed that a k-means algorithm using the KL-divergence distance metric with an exponential time worst-case complexity (see [90]) can obtain $O(\log K)$-approximation of the optimal clustering. The algorithm of Chaudhuri and McGregor [89] can provide $\log(M)$-approximation for finding a good clustering in polynomial time complexity. On the other hand, the quality of the approximation algorithm in [89] is dependent on the size of the dataset. In many settings where M (number of data points) tends to be large while N (data dimension) and K (number of clusters) tend to be smaller, thus the algorithm proposed by Cicalese et al. [1] is useful due to a smaller constant factor approximation of $\log^2(\min\{N, K\})$ as compared to $\log(M)$ in [89].

The contributions of this chapter are fivefold:

- We describe an approximate algorithm based on the maximum likelihood principle for a wide class of impurity functions including both Gini index and entropy. The proposed algorithm is called maximum likelihood algorithm (Algorithm 1) which provides a comparable approximation factor than that of the state-of-art method in [1] for both Gini index impurity and entropy impurity. Particularly, our theoretical bound is 2-approximation for Gini index and $\log^2 N$-approximation for entropy impurity. In addition, the running time of the proposed algorithm is $O(NM)$ and $O(2^{N/2}NM)$ for the case of $K \geq N$ and $K < N$, respectively. That said, the proposed maximum likelihood algorithm (Algorithm 1) provide a better approximation together with a lower running time in comparison to the proposed algorithm in [1] when $K \geq N$.

- Based on the Algorithm 1, we propose a so-called greedy-splitting algorithm (Algorithm 2) to achieve a better splitting quality when $K > N$. Greedy-splitting algorithm still

runs in $O(KNM)$ and achieves the bound at least equal or better compared to the bound of the original maximum likelihood algorithm e.g., 2-approximation for Gini index and $\log^2 N$-approximation for entropy impurity. When $K < N$, the proposed Algorithm 1 runs in $O(2^{N/2}NM)$ which is exponential in term of N. To reduce the running time, we proposed a so-called greedy-merge algorithm (Algorithm 3) having the time complexity of $O((N-K)N^2 + NM)$ which is linear in the size of the data set M. Although the greedy-merge algorithm does not provide a guarantee on splitting quality, it shows a comparable performance to the results provided by the proposed algorithm in [1].

- To keep the generality of the impurity functions, instead of the providing a constant factor approximation, we provide both the upper bound and the lower bound differently for different impurity functions.

- We suggest a method that can improve the complexity of the algorithm in [1] from $O(NM + M^3)$ down to $O(NM + M \log M)$ based on the matrix searching SMAWK algorithm [87].

- Our technique on bounding the solution confirms and generalizes well-established results in signal processing and information theory, specifically the Fano's inequality [3] and Boyd-Chiang upper bound [13] for channel capacity.

From the signal processing, communication, and information theory's perspective, our work is related to the optimal quantization design for constructing polar code [41] and low density parity code (LDPC) decoder [39]. In these settings, an optimal quantizer maximizes the mutual information between input and quantized output [44], [54], [45, 52, 57, 91–93]. However, Kurkoski et al. [44] showed that for a given input distribution, finding an optimal quantizer that maximizes the mutual information is equivalent to finding an optimal partition that minimizes the entropy impurity. Thus, our algorithm can be applied to find a good quantizer that maximizes

mutual information. In addition, the problem of minimizing impurity partitions also relates to the well-known Information Bottleneck Method (IBM) [94] and Deterministic Information Bottleneck (DIB) [67]. Particularly, both IBM and DIB can be viewed as the problems of minimizing entropy impurity partition under constraints for a given input distribution. From the relationship between minimizing impurity partition problem with IBM and DIB, we strongly believe that there are some advantages to use the technical results in approximation impurity partition to designing the good approximation algorithms for IBM and DIB.

The outline of this chapter is as follows. In Section 6.2, we describe the problem formulation. In Section 6.3, an upper bound of impurity partition is constructed together with an algorithm that provably achieves near-optimal impurity. The proof of near-optimal partition together with a lower bound of impurity function is characterized in Section 6.4. To reduce the running time of the proposed algorithm in Section 6.3, we propose two linear time complexity greedy algorithms in Section 6.5. The numerical results are provided in Section 6.6. Finally, we provide a few concluding remarks in Section 6.7.

6.2 Problem Formulation

6.2.1 Problem formulation

We assume that the set $\mathbb{Y}$ to be partitioned consists of M discrete data points generated from an underlying probabilistic model. Specifically, let X be a discrete random variable taking on the values $x_1, x_2, \ldots, x_N$ with a given probability mass vector $\mathbf{p_x} = (p(x_1), p(x_2), \ldots, p(x_N))$. Let Y be another discrete random variable taking on values $y_1, y_2, \ldots, y_M$ which follows a given conditional probability $p(y_j|x_i)$. The goal is to partition $\mathbb{Y}$ into K partitions to minimize a given impurity function over the resulted partitions. For convenience of analysis and notations, we

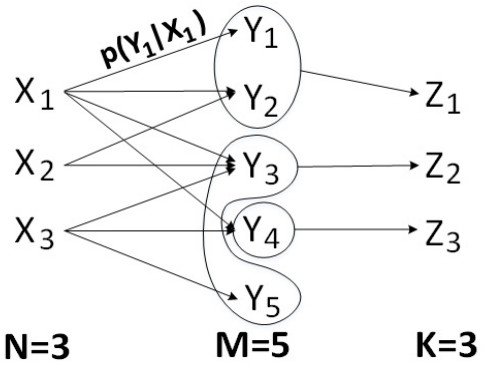


Figure 6.1: Finding an optimal quantizer $Q^*(Y) \rightarrow Z$ such that I_{Q^*} is minimized.

assume x_i and y_j are scalar values. If x_i and y_j are discrete multi-dimensional vectors, they can be mapped to scalar values appropriately. The result is N and M change, but the overall analysis does not change. For example, if x_i is a 2-dimensional vector and y_j is a 3-dimensional vector, then the maximum number of x_i is $N' = N^2$ and the maximum number of y_j is $M' = M^3$. Fig. 6.1 shows a generative model for Y. Y is then quantized to Z using a partition scheme/quantizer Q.

$$Q(Y) \rightarrow Z.$$

Z is modeled as a discrete random variable Z taking on values $z_1, z_2, \ldots, z_K$. In this setting, for given $p(x_i)$ and $p(y_j|x_i)$, $p(x_i, y_j)$ are assumed to be given $\forall i, j$. Thus, each data point y_j is represented by a joint distribution vector $\mathbf{p}_{\mathbf{x},y_j} = (p(x_1, y_j), p(x_2, y_j), \ldots, p(x_N, y_j))$. Each quantizer Q induces a joint distribution vector $\mathbf{p}_{\mathbf{x},z_k} = (p(x_1, z_k), p(x_2, z_k), \ldots, p(x_N, z_k))$ between X and $Z = z_k$. The conditional distribution $p(x_i|z_k)$ of X given Z and the marginal probability mass function $p(z_k)$ of Z can be determined from $p(x_i, z_k)$. We want to find an optimal quantizer Q^* that minimizes the impurity function I_Q that satisfies two following conditions:

- (**Required**) I_Q has the following form:

$$I_Q = \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_k) f(p(x_i|z_k)), \tag{6.1}$$

where $f(.) : \mathbb{R} \to \mathbb{R}^+$ is a non-negative concave function. $f(x)$ is concave over a continuous interval $\mathbb{S}$ if for any $a, b \in \mathbb{S}$,

$$f(\lambda a + (1-\lambda)b) \geq \lambda f(a) + (1-\lambda) f(b), \forall \lambda \in (0,1). \tag{6.2}$$

We note that $\sum_{i=1}^{N} f(p(x_i|z_k))$ in (6.1) is the impurity contribution from partition z_k. Therefore, I_Q is viewed as the weighted average impurity over all the partitions. Many well-known impurity functions such as Gini index [85] and entropy [86] have concave $f(.)$.

- (**Optional**) $f(x) = xl(x)$ where $l(x) : \mathbb{R} \to \mathbb{R}$ is a convex function. $l(x)$ is convex over a continuous interval $\mathbb{S}$ if for any $a, b \in \mathbb{S}$,

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b), \forall \lambda \in (0,1). \tag{6.3}$$

This second condition is optional in the sense that we use it in the analysis of the constant factor approximation for the proposed algorithm. The algorithm itself does not make use of this condition. Furthermore, many popular impurity functions indeed satisfy this second condition.

Examples of popular impurity functions:

- Entropy function: Let $f(x) = -x \log x$ which can be shown to be a concave function.

Replacing $f(x) = -x \log x$ with $x = p(x_i|z_k)$ into (6.1), we have:

$$I_Q = \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_k)[-p(x_i|z_k) \log (p(x_i|z_k))], \tag{6.4}$$

which is the weighted conditional entropy of X given Z. Also let $l(x) = -\log x$, $l(x)$ is a convex function. Thus $f(x) = -x \log x$ satisfies the second optional condition.

- Gini index function: Given a set A of elements with random N labels according to the distribution of the labels $\mathbf{p_A} = (p(A_1), p(A_2), \ldots, p(A_N))$. The Gini impurity is a measure of how often a randomly chosen element based on the label distribution would be mislabeled. Specifically, since the probability of picking an element with the label i is $p(A_i)$ then the probability of mislabeling that element is $\sum_{l \neq i} p(A_i) = 1 - p(A_i)$. Summing all i, the probability of mislabeling an element is:

$$\sum_{i=1}^{N} p(A_i)(1 - p(A_i)).$$

Let $f(x) = x(1-x)$ which can be shown to be a concave function. Replacing $f(x) = x(1-x)$ using $x = p(x_i|z_k)$ into (6.1), the Gini index impurity [53] has the following form:

$$I_Q = \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_k)[p(x_i|z_k)(1 - p(x_i|z_k))]. \tag{6.5}$$

Additionally, let $l(x) = 1 - x$, $l(x)$ is a linear function, therefore, $l(x)$ is a convex function. Thus the Gini index impurity satisfies the second optional condition.

We note that in [1] and [56], to guarantee the constant factor approximation, the authors considered a class of impurity concave functions $f(.)$ with an additional condition on $xf''(x)$ being a non-increasing function.

6.3 Impurity Minimization Algorithm

In this section, we first construct both upper and lower bounds for impurity functions of the form in (6.1). Using these bounds, we show that the proposed maximum likelihood algorithm achieves a constant factor approximation. In other words, the resulted solution is guaranteed to be away from the true solution by at most a constant factor that does not depend on the number of data points M.

We define three important quantities below.

$$k^* = \arg\max_{1 \le i \le N} p(x_i|z_k), \tag{6.6}$$

$$e_Q = \sum_{k=1}^{K} p(z_k) p(x_{k^*}|z_k), \tag{6.7}$$

and

$$e^{\max} = \max_{Q} e_Q. \tag{6.8}$$

For a given k, z_k is most likely be produced by x_{k^*}. Therefore, e_Q is the weighted sum of the maximum likelihood of each x_{k^*} for each z_k. We note that each partition scheme/quantizer Q induces a $p(x_i, z_k)$ and thus $p(x_i|z_k)$. So k^* and e_Q are different for different Q. Our approach to find the minimum impurity is to find two functions: $u(e_Q)$ and $l(e_Q)$ such that $l(e_Q) \le I_Q \le u(e_Q)$. Furthermore, we show that $u(e_Q)$ and $l(e_Q)$ are decreasing functions for many impurities. Therefore, by minimizing $u(e_Q)$, i.e., maximizing e_Q, we can bound the minimum value of I_Q between $u(e_Q)$ and $l(e_Q)$ for some e_Q.

6.3.1 Upper Bound of The Impurity Function

We have the following theorem for the upper bound of an impurity function I_Q.

Theorem 6.1. *(***Upper bound***) For any given quantizer Q that induces the corresponding $p(x_i|z_k)$ and*

$$e_Q = \sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k), \tag{6.9}$$

let

$$u(e_Q) = f(e_Q) + (N-1)f(\frac{1-e_Q}{N-1}), \tag{6.10}$$

then $\forall e_Q$, we have:

$$u(e_Q) \geq I_Q. \tag{6.11}$$

Proof. From the definition of the impurity function, we have

$$\begin{aligned}
I_Q &= \sum_{k=1}^{K}\sum_{i=1}^{N} p(z_k)f(p(x_i|z_k)) \\
&= \sum_{k=1}^{K} p(z_k)f(p(x_{k^*}|z_k)) + \sum_{k=1}^{K}\sum_{i\neq k^*,i=1}^{N} p(z_k)f(p(x_i|z_k)) \\
&\leq f(\sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k)) + \sum_{k=1}^{K}\sum_{i\neq k^*,i=1}^{N} p(z_k)f(p(x_i|z_k)) && (6.12) \\
&\leq f(\sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k)) + \sum_{k=1}^{K} p(z_k)[(N-1)f(\frac{\sum_{i=1,i\neq k^*} p(x_i|z_k)}{N-1})] && (6.13) \\
&= f(e_Q) + (N-1)\sum_{k=1}^{K} p(z_k)f(\frac{1-p(x_{k^*}|z_k)}{N-1}) && (6.14) \\
&\leq f(e_Q) + (N-1)f(\frac{\sum_{k=1}^{K} p(z_k)(1-p(x_{k^*}|z_k))}{N-1}) && (6.15) \\
&= f(e_Q) + (N-1)f(\frac{1-e_Q}{N-1}), && (6.16)
\end{aligned}$$

where (6.12) is due to concavity of $f(.)$ and $\sum_{k=1}^{K} p(z_k) = 1$, (6.13) is due to Jensen inequality for concave function (please see the Appendix 6.8.2), (6.14) is due to the definition of e_Q and $\sum_{i=1, i\neq k^*}^{N} p(x_i|z_k) + p(x_{k^*}|z_k) = 1$, (6.15) is due to concavity of $f(.)$ and $\sum_{k=1}^{K} p(z_k) = 1$, (6.16) is due to $\sum_{k=1}^{K} p(z_k) = 1$. □

Remark 6.1. *(***Fano's inequality.***) There is an interesting connection between $u(e_Q)$ and the well-known Fano's inequality from the information theory. Specifically, if $f(x)$ is the entropy function, then the upper bound in Theorem 6.1 is identical to the Fano's inequality. Please see the details of the derivations in the Appendix 6.8.3.*

Remark 6.2. *(***Maximum likelihood decoding.***) Consider a communication setting with X and Z being the two random variables that represent the transmitted symbols and the received symbols, a respectively. The goal for a receiver is to recover X based on Z. A maximum likelihood decoder maximizes the posterior probability of X given Z. Specifically, if a symbol z_k is received, then the transmitted symbol is x_{k^*} where $k^* = \arg\max_{1\leq i\leq N} p(x_i|z_k)$. Consequently, $e_Q = \sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k)$ is the probability of decoding a transmitted symbol correctly using the mapping Q, and $P_e = 1 - e_Q$ is the probability of decoding a transmitted symbol incorrectly.*

Theorem 6.2. *$u(e_Q)$ is a monotonic decreasing function. Moreover, $I_Q = u(e_Q)$ when $e_Q = \frac{1}{N}$ or $e_Q = 1$.*

Proof. Please see the Appendix 6.8.5. □

Based on Theorem 6.2, let $e^{\max}$ be the maximum value over all e_Q i.e., $e^{\max} = \max_Q e_Q$, then $u(e^{\max})$ has the minimum value. Since $u(e_Q)$ is an upper bound of I_Q, $u(e^{\max})$ provides a good upper bound for I_{Q^*}. We now state an important result for a special case where the sample space of Z is identical to that of X. In other words, $K = N$ and $z_k = x_i$ for some k and i.

Theorem 6.3. *(***Structure of the e$^{\max}$ quantizer***) Let $\mathcal{Z}$ and $\mathcal{X}$ be the sample spaces of Z and X, respectively. Let $j^* = \arg\max_i p(x_i, y_j)$ and define quantizer $Q_{e^{\max}}$ with the following structure:*

$$Q_{e^{\max}}(y_j) = z_{j^*}. \tag{6.17}$$

(a) If $|\mathcal{Z}| = |\mathcal{X}|$, then $Q_{e^{\max}}$ produces $e^{\max} = \max_Q e_Q$. Conversely, for any Q that produces $e^{\max}$, Q must have the structure of $Q_{e^{\max}}$.

(b) If $|\mathcal{Z}| > |\mathcal{X}|$, then $Q_{e^{\max}}$ still produces $e^{\max} = \max_Q e_Q$. However, it is not necessary that for any Q that produces $e^{\max}$, Q must have the structure of $Q_{e^{\max}}$.

Proof. Please see the Appendix 6.8.6. □

We note that j^* takes on values $1, 2, \ldots, N$, and z_{j^*}'s represent the $K = N$ partitions. In other words, when $K > N$ then existing an optimal quantizer $Q_{e^{\max}}$ that produces exactly N-partition rather than K-partition. Interestingly, for $K > N$, the mapping using only N-partition in Theorem 6.3-(b) is still optimal i.e., it produces the partitions achieving $e^{\max}$. However, Theorem 6.3-(b) does not guarantee any Q that produces $e^{\max}$ must have the structure of $Q_{e^{\max}}$. Indeed, there might exist other quantizers that achieve $e^{\max}$. On the other hand, Theorem 6.3-(a) states that if $K = N$, then any quantizer producing $e^{\max}$ must have the structure of $Q_{e^{\max}}$ in (6.17). This necessary condition helps to find $Q_{e^{\max}}$ when $K < N$ as to be shown later. The detail of proof is presented in Appendix 6.8.6.

6.3.2 Algorithm

Based on the upper bound in Theorem 6.1, to minimize the impurity function, one wants to minimize the impurity's upper bound $u(e_Q)$. Based on Theorem 6.2, to minimize $u(e_Q)$, one wants to maximize e_Q. To maximize e_Q, we propose the algorithm below which utilizes the result

of Theorem 6.3.

Let $\mathcal{V}_K$ be the set of binary N-dimensional vectors $\mathbf{v}$'s, each contains exactly K entries 1 and $N-K$ entries 0. Thus, the size of $\mathcal{V}_K$ is $\binom{N}{K}$. For each $\mathbf{v} = (v_1, v_2, \ldots, v_N)$, define the N-dimensional vector:

$$\mathbf{p}'_{\mathbf{x},y_j} = (v_1 p(x_1, y_j), v_2 p(x_2, y_j), \ldots, v_N p(x_N, y_j)) = (p'(x_1, y_j), p'(x_2, y_j), \ldots, p'(x_N, y_j)),$$

then $\mathbf{p}'_{\mathbf{x},y_j}$ has exactly K non-zero entries. Next, we consider the following possible cases.

- $K = N$: When $K = N$, $\mathcal{V}_K = \mathcal{V}_N$ contains exactly one $\mathbf{v}$ which is $\mathbf{v} = (1, 1, \ldots, 1)$. In this case, $p'(x_i, y_j) = p(x_i, y_j)$. Thus, using Theorem 6.3-(a) with $p(x_i, y_j)$ replaced by $p'(x_i, y_j)$ will produce $e^{\max}$.

- $K < N$: When $K < N$, there are $\binom{N}{K}$ quantizers Q that partition K-dimension vectors $\mathbf{p}'_{\mathbf{x},y_j}$ to K partitions. Moreover, from the necessary condition in Theorem 6.3-(a), at least one of quantizer in this $\binom{N}{K}$ quantizers must achieve $e^{\max}$.

- $K > N$: From Theorem 6.3-(b), the partition which achieves $e^{\max}$ is exactly the same with the partition when $K = N$. In other words, the partition can be achieved using the maximum likelihood principle using $\mathbf{v} = (1, 1, \ldots, 1)$, and the optimal partitions which produces $e^{\max}$ has N nonempty partitions together with $K - N$ empty partitions.

Based on three possible cases above, the algorithm follows. The detail of the proof is shown in Appendix 6.8.6.

Running time of Algorithm 1: To find the partition that generates $e^{\max}$, we need to search over all the possible mappings $\mathbf{v} \in \mathcal{V}_K$. For each $\mathbf{v}$, Algorithm 1 has complexity of $O(NM)$. Since there are $\binom{N}{K}$ possible $\mathbf{v}$ if $K < N$, Algorithm 1 has the complexity of $O(\binom{N}{K}NM)$. In the worst case when $K = N/2$, we have $\binom{N}{N/2} = 2^{N/2}$ and the complexity of Algorithm 1 is $O(2^{N/2}NM)$.

Algorithm 1 Finding $e^{\max}$ Algorithm.

Input: Dataset $Y = \{y_1, \ldots, y_M\}$ and $p(x_i, y_j)$, K and N.
Output: Partition $Z = \{z_1, z_2, \ldots, z_K\}$.
If $K < N$**:** $\mathcal{V} = \mathcal{V}_K$
If $K \geq N$**:** $\mathcal{V} = \mathcal{V}_N$
    For $\mathbf{v} \in \mathcal{V}$
        For $1 \leq j \leq M$**,** $1 \leq i \leq N$
            Step 1: Projection.

$$p'(x_i, y_j) = v_i p(x_i, y_j). \tag{6.18}$$

            Step 2: Finding the maximum likelihood.

$$j^* = \arg\max_{1 \leq i \leq N} \{p'(x_i, y_j)\}. \tag{6.19}$$

            Step 3: Partition assignment.

$$Q(y_j) \to z_{j^*}. \tag{6.20}$$

        End For
        Computing e_Q**:** Using the resulted partitions to compute e_Q.
    End For
Return: Returning the partition that produces $e^{\max} = \max_Q e_Q$.

However, if $K \geq N$, there is only one mapping $\mathbf{v}$ and the running time of algorithm is truly in linear of $O(NM)$.

6.4 Constant Factor Approximation Analysis for Entropy and Gini Index

In this section, we state a few results for establishing the constant approximation property of Algorithm 1. First, the following theorem establishes a lower bound for I_Q. This lower bound predicates on the second condition $f(x) = xl(x)$ where $l(x)$ is a convex function. It is not used explicitly in the algorithm but is used in the analysis to establish the constant factor approximation property of the algorithm.

Theorem 6.4. *(***Lower bound***) For any given quantizer Q that induces the corresponding $p(x_i|z_k)$ and*

$$e_Q = \sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k), \tag{6.21}$$

then $\forall e_Q$, we have:

$$I_Q \geq l(e_Q). \tag{6.22}$$

Proof. Using the concavity definition of $f(x)$ in (6.3), and let t and q be the positive scalars such that $0 \leq t \leq q$, we have:

$$f(t) \geq (1 - \frac{t}{q})f(0) + \frac{t}{q}f(q) = \frac{t}{q}f(q). \tag{6.23}$$

From the definition of the impurity function, we have

$$
\begin{aligned}
I_Q &= \sum_{k=1}^{K}\sum_{i=1}^{N} p(z_k) f(p(x_i|z_j)) \\
&= \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k)) + \sum_{k=1}^{K} p(z_k)\Big(\sum_{i\neq k^*, i=1}^{N} f(p(x_{k^*}|z_k))\Big) \\
&\geq \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k)) + \sum_{k=1}^{K} p(z_k)\Big(\sum_{i\neq j^*, i=1}^{N} \frac{p(x_i|z_k)}{p(x_{k^*}|z_k)} f(p(x_{k^*}|z_k))\Big) && (6.24) \\
&= \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k)) + \sum_{k=1}^{K} p(z_k) \frac{\sum_{i\neq k^*, i=1}^{N} p(x_i|z_k)}{p(x_{k^*}|z_k)} f(p(x_{k^*}|z_k)) && (6.25) \\
&= \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k)) + \sum_{k=1}^{K} p(z_k)\Big(\frac{1-p(x_{k^*}|z_k)}{p(x_{k^*}|z_k)}\Big) f(p(x_{k^*}|z_k)) && (6.26) \\
&= \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k))\Big(1 + \frac{1-p(x_{k^*}|z_k)}{p(x_{k^*}|z_k)}\Big) && (6.27) \\
&= \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k)) \frac{1}{p(x_{k^*}|z_k)} && (6.28) \\
&= \sum_{k=1}^{K} p(z_k) p(x_{k^*}|z_k) l(p(x_{k^*}|z_k)) \frac{1}{p(x_{k^*}|z_k)} && (6.29) \\
&= \sum_{k=1}^{K} p(z_k) l(p(x_{k^*}|z_k)) && (6.30) \\
&\geq l(\sum_{k=1}^{K} p(z_k) p(x_{k^*}|z_k)) && (6.31) \\
&= l(e_Q), && (6.32)
\end{aligned}
$$

with (6.24) is due to (6.23) using $t = p(x_i|z_k)$ and $q = p(x_{k^*}|z_k)$ and noting that $p(x_{k^*}|z_k) \geq p(x_i|z_k)\ \forall i$, (6.29) is due to $f(x) = xl(x)$, and (6.31) due to the Jensen inequality for the convex function $l(x)$. The lower bound is tight i.e., $I_Q = l(e_Q)$ if $e_Q = \frac{1}{N}$ or $e_Q = 1$. □

Remark 6.3. *There is a connection between the lower bound above and the well-known Boy-Chiang upper bound of channel capacity. Specifically, for a uniform input distribution, if $f(x)$ is the entropy function, then the upper bound in Theorem 6.4 implies the Boy-Chiang upper bound of channel capacity [13]. More details are in the Appendix 6.8.4.*

Theorem 6.5. *(***R**($\mathbf{e}^{\max}$)**-approximation**) *Algorithm 1 provides $R(e^{\max})$-approximation for both entropy and Gini index impurities where:*

$$R(e^{\max}) = \frac{u(e^{\max})}{l(e^{\max})}. \tag{6.33}$$

Proof. Let I_{Q^*} be the minimum impurity and $I_{Q_{e^{\max}}}$ be the impurity of the partition produced by running Algorithm 1. Now, assume that Q^* produces e_{Q^*}. From the definition of $e^{\max}$, $e_{Q^*} \leq e^{\max}$. Moreover, it is straightforward to show that $l(e_Q)$ for both entropy and Gini index impurities are decreasing functions. Thus, $I_{Q^*} \geq l(e_{Q^*}) \geq l(e^{\max})$. Therefore,

$$\frac{I_{Q_{e^{\max}}}}{I_{Q^*}} \leq \frac{u(e^{\max})}{\min_{e_Q} l(e_Q)} = \frac{u(e^{\max})}{l(e^{\max})} = R(e^{\max}). \tag{6.34}$$

Thus, the impurity produced by Algorithm 1 is guaranteed to be away from the true solution by at most a factor of $R(e^{\max})$. Fig. 6.2 shows $u(e_Q)$ and $l(e_Q)$ vs. $e_Q \in (0.01, 0.99)$ using $N = 100$ for both the entropy impurity and the Gini index impurity. As seen, $u(e_Q)$ and $l(e_Q)$ are monotonic decreasing functions for both entropy and Gini index impurities. Moreover, the upper bound and the lower bound are tight and equal when $e_Q = \frac{1}{N}$ or $e_Q = 1$.

□

The result in Theorem 6.5 can be applied for any concave impurity function $f(x) = xl(x)$ with $l(x)$ being a non-increasing function. Next, we show that $R(e^{\max})$-approximation is better than

the approximation in [1] for both the entropy impurity and the Gini index impurity.

Theorem 6.6.

- *For Gini index impurity,*

$$R(e^{\max}) = 1 + e^{\max}. \tag{6.35}$$

- *For Entropy impurity,*

$$R(e^{\max}) = \frac{H(e^{\max}) + (1 - e^{\max}) \log(N-1)}{-\log(e^{\max})}. \tag{6.36}$$

Proof. The proof follows (6.33) by using the upper bound and the lower bound in Theorem 6.1 and Theorem 6.4, respectively. The detail of proof can be viewed in Appendix 6.8.7 and 6.8.8. □

Theorem 6.7. *Algorithm 1 provides a 2-approximation for Gini index impurity.*

Proof. Please see Appendix 6.8.7. □

Remark 6.4. *Algorithm 1 always provides a 2-approximation for Gini index impurity while the algorithm in [1] provides a 3-approximation in the worst case.*

Theorem 6.8. *The entropy impurity approximation provided by Algorithm 1 is better than the approximation in [1] in case of $K \geq N$, i.e., $R(e^{\max}) < \log^2(\min\{K, N\}) = \log^2 N$ if*

$$N \geq N^{\min} = 2^{S(e^{\max})}, \tag{6.37}$$

where

$$S(e^{\max}) = \frac{1 - e^{\max}}{-2\log(e^{\max})} + \frac{\sqrt{4H(e^{\max})(-\log(e^{\max})) + (1 - e^{\max})^2}}{-2\log(e^{\max})}, \tag{6.38}$$

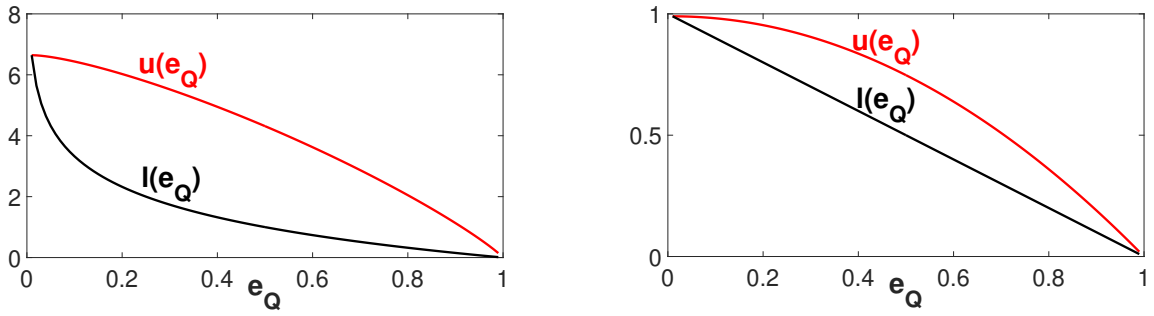


Figure 6.2: The monotonic decreasing of $u(e_Q)$ and $l(e_Q)$ for (a) entropy impurity and (b) Gini index impurity using $e_Q \in (0.01, 0.99)$ and $N = 100$.

and $H(x) = -(x \log x + (1-x)\log(1-x))$ *is the binary entropy of* x.

Proof. Theorem 6.8 provides a sufficient condition where the approximation produced by Algorithm 1 is better than that produced by the algorithm in [1]. In reality, $R(e^{\max})$ is smaller than $\log^2 N$ for a wider range of N. Please see the Appendix 6.8.8 for the details of proof. □

Fig. 6.3 shows the performance bound of the proposed algorithm vs. the state of the art in [1]. $R(e^{\max})$ vs. $e^{\max} \in (0.01, 0.99)$ for $N = 10$ and $N = 20$ are plotted in red while the approximations of [1] ($\log^2 N$) are plotted in blue. As seen, the red curves are always below the blue curves. Moreover, the gaps between our approximation and that of [1] are proportional to the size of N. That said, for large values of N, our approximation is progressively better than that of [1].

We also note that $S(e^{\max})$ is a monotonic increasing function as shown in Fig. 6.4-(a). Thus, if $e^{\max}$ increases, then $N^{\min}$ increases. For example, if $e^{\max} = 0.5$ then (6.37) holds for $N^{\min} = 2.42$, if $e^{\max} = 0.8$, (6.37) holds for $N^{\min} = 3.58$, if $e^{\max} = 0.9$, (6.37) holds for $N^{\min} = 4.34$, if $e^{\max} = 0.999$, (6.37) holds for $N^{\min} = 9.06$. As seen, when $N^{\min}$ increases to infinity, $e^{\max}$ increases to 1. Fig. 6.4-(b) illustrates the relationship between $e^{\max}$ and $N^{\min}$.

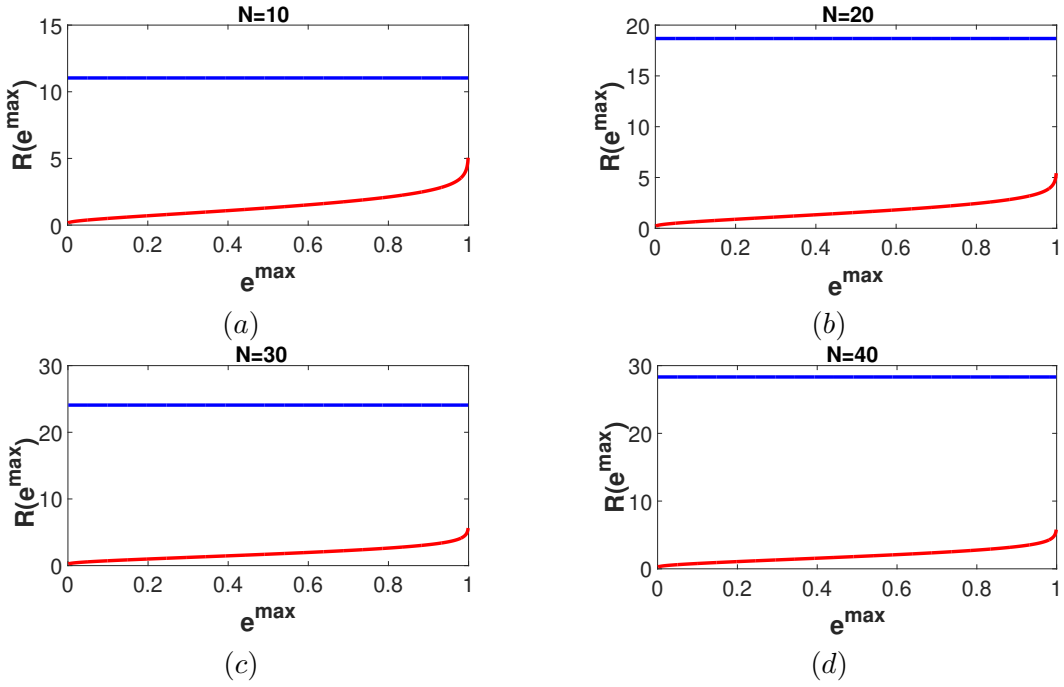


Figure 6.3: $R(e^{\max})$-approximation for entropy impurity using (a) $N = 10$; (b) $N = 20$; (c) $N = 30$; (d) $N = 40$. Our approximation ($R(e^{\max})$) are the red curves while the approximations of the algorithm in [1] are the blue curves.

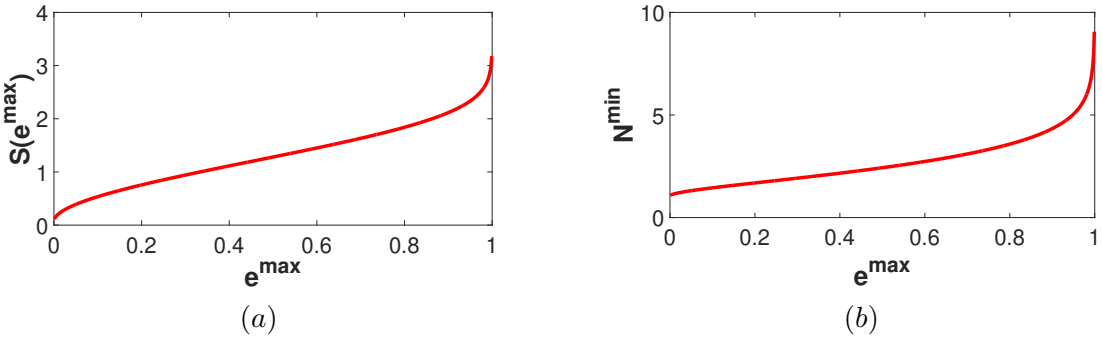


Figure 6.4: (a) $S(e^{\max})$ as a function of $e^{\max}$; (b) $N^{\min} = 2^{S(e^{\max})}$ as a function of $e^{\max}$.

Remark 6.5. *For $K \geq N$, it is obviously that the approximation guaranteed by Algorithm 1 for entropy impurity is better than that of the state-of-art approximation in [1]. On the other hand, when $K < N$, it is possible that the approximation in [1] provides a better bound than our approximation i.e.,* $\log^2(\min\{K, N\}) < R(e^{\max})$ *due to* $\min\{K, N\} = K$ *completely depends on K while $R(e^{\max})$ in (6.36) depends on both $e^{\max}$ and N. Therefore, a smaller value of K, a higher chance that the algorithm in [1] provides a better approximation. For example, consider the 20NEWS dataset having $e^{\max} = 0.2420$, using $N = 20$, the approximation in [1] is better than our approximation if $K \leq 2.64$. Similarly, consider the RCV1 dataset having $e^{\max} = 0.2185$, using $N = 103$, the approximation in [1] is better than our approximation if $K \leq 5.97$. To that end, our algorithm still provides a better approximation for a wide range of K even if $K < N$. For example, our bound is better $\forall K \geq 2.64$ using 20NEWS dataset, and $\forall K \geq 5.97$ using RCV1 dataset regardless of N. Please see detail of these datasets in Sec. 6.6.*

6.5 Practical algorithms

In the previous section, we show that the proposed Algorithm 1 is near-optimal and achieves 2-approximation for Gini index impurity and $\log^2 N$-approximation for entropy impurity with arbitrary values of K and N. However, there are some main drawbacks that limit the applications of Algorithm 1. Particularly,

- When $K > N$, Algorithm 1 produces the optimal partitions containing exactly N partitions due to $\mathcal{V} = \mathcal{V}_N$. Therefore, even though that Algorithm 1 still achieves the theoretical bounds, it produces $(K - N)$ empty partitions.
- When $K < N$, the running time of Algorithm 1 in the worst case is exponential in N.
- The partitions produced by Algorithm 1 might not be local optimal partitions. For the

convenience of the reader, the necessary condition for optimal partitions can be viewed in Theorem 6.10, Appendix 6.8.9.

To resolve these problems, we propose several modifications of Algorithm 1 which results in two linear-time complexity algorithms. These algorithms are build up based on greedy-splitting (Algorithm 2) or greedy-merge (Algorithm 3) the partitions produced by Algorithm 1.

6.5.1 Handling the case $K > N$: greedy-splitting algorithm

To resolving the problem of $(K-N)$ empty partitions when $K > N$, we propose a so-called greedy-splitting algorithm (Algorithm 2). The first step of greedy-splitting algorithm is using Algorithm 1 to generate N non-empty partitions. Next, by greedy splitting, one can generate more $(K-N)$ no-nempty partitions to achieve total K partitions. As will be shown later, Algorithm 2 runs in $O(KNM)$ and achieves all the theoretical bounds of Algorithm 1.

Algorithm: The first step of greedy-splitting algorithm is finding the partition that has the largest impurity (line 7, Algorithm 2). Next, this partition is separated based on the largest attribution. For example, if the largest impurity partition is z_{i^*}, and recall that:

$$j^* = \max_{1 \leq j \leq N} p(x_j|z_{i^*}).$$

Then $\forall y_q \in z_{i^*}$, $p(x_{j^*}|y_q)$ is the largest attribution of y_q. Using $p(x_{j^*}|z_{i^*})$ as a threshold, by comparing $p(x_{j^*}|y_q)$ to $p(x_{j^*}|z_{i^*})$, $\forall y_q \in z_{i^*}$, y_q is assigned into two new partitions (line 10 and 11, Algorithm 2). The process repeats until K partitions are generated. Although the splitting based on $p(x_{j^*}|z_{i^*})$ as a threshold is a heuristic method, it guarantees a better impurity than that provided by the original Algorithm 1 as will be shown later. The pseudo-code of our splitting procedure is presented in Algorithm 2.

Algorithm 2 Greedy-splitting algorithm for $K > N$.

Input: Dataset $Y = \{y_1, \ldots, y_M\}$ and $p(x_i, y_j)$.
Output: Partition $Z = \{z_1, z_2, \ldots, z_K\}$.
    Step 1: Running Algorithm 1 to achieve N-partition $z_1, z_2, \ldots, z_N$.
    Step 2: Greedy splitting.
        $t = 1$
        While: $t \leq K - N$
            Finding the largest impurity.

$$i^* = \max_i F(\mathbf{p}_{\mathbf{x},z_i}). \tag{6.39}$$

            Splitting based on the largest attribution.

$$j^* = \max_j p(x_j|z_{i^*}). \tag{6.40}$$

            For $y_q \in z_{i^*}$.
                If $p(x_{j^*}|y_q) \leq p(x_{j^*}|z_{i^*})$.

$$Q(y_q) \to z_{i^*}.$$

                Else $p(x_{j^*}|y_q) > p(x_{j^*}|z_{i^*})$.

$$Q(y_q) \to z_{N+t}.$$

            $t = t + 1$.
        End While.
    Step 3: Return K partitions.
Return: Return K partitions.

Proof of a better approximation: From Theorem 6.9 in Appendix 6.8.9, if $z_k = z_k^1 \cup z_k^2$ and $z_k^1 \cap z_k^2 = \emptyset$, then the impurity in z_k is at least large as the total impurity in z_k^1 and z_k^2. Therefore, by greedy splitting, the greedy-splitting algorithm produces a new partition having the impurity is monotonic decreasing over each splitting step. Finally, the impurity of K partitions produced by greedy-splitting algorithm is less than or at least equal the impurity provided by Algorithm 1. Thus, the partitions induced by greedy-splitting algorithm must satisfy our theoretical bounds in Sec. 6.4.

Running time of Algorithm 2: The time complexity of Step 1 and Step 2 of Algorithm 2 are NM and $(K-N)NM$, respectively. Therefore, the running time of Algorithm 2 is $O(KNM)$.

6.5.2 Handing the case $K < N$: greedy-merge algorithm

To resolving the high time complexity of Algorithm 1 when $K < N$, we propose a so-called greedy-merge algorithm (Algorithm 3) to reduce the time complexity. Particularly, we first use the Algorithm 1 for $K = N$ to achieve N-partition. Next, we perform $N - K$ times of the greedy-merge, each time the algorithm merges two partitions into one single partition that minimizes the impurity loss until achieving exactly K-partition ($K < N$). As will be shown later, the running time of this greedy-merge algorithm is $O((N-K)N^2+NM)$ that is linear in M and polynomial in N. Although the greedy-merge algorithm does not satisfy the theoretical bounds, it performance is comparable to the results provided by the proposed algorithm in [1] (please see the numerical results in Sec. 6.6).

Algorithm: As discussion earlier, greedy-merge algorithm first use Algorithm 1 to generate N-partition (line 3, Algorithm 3). Next, greedy-merge algorithm performs $(N-K)$ times of greedy merge, each time the algorithm merges two partitions into one single partition that minimizes the impurity loss Δ (line 10 and 13, Algorithm 3) until achieving exactly K-partition. The

Algorithm 3 Greedy-merge algorithm for $K < N$.

Input: Dataset $Y = \{y_1, \ldots, y_M\}$ and $p(x_i, y_j)$.
Output: Partition $Z = \{z_1, z_2, \ldots, z_K\}$.
    Step 1: Running Algorithm 1 to achieve N-partition $z_1, z_2, \ldots, z_N$.
    Step 2: Greedy merge.
        $t = 0$
        While: $t \leq N - K$
            For: $i = 1, 2, \ldots, K - t - 1$
                For: $j = i + 1, i + 2, \ldots, K - t$
                Merge z_i, z_j to z_{ij} and compute:

$$\mathbf{p}_{\mathbf{x},z_{ij}} = \mathbf{p}_{\mathbf{x},z_i} + \mathbf{p}_{\mathbf{x},z_j}, \tag{6.41}$$

$$F(\mathbf{p}_{\mathbf{x},z_{ij}}) = \sum_{k=1}^{N} p(x_k, z_{ij}) \sum_{k=1}^{N} f\Big(\frac{p(x_k, z_{ij})}{\sum_{i=1}^{N} p(x_k, z_{ij})}\Big), \tag{6.42}$$

$$\Delta_{ij} = F(\mathbf{p}_{\mathbf{x},z_{ij}}) - F(\mathbf{p}_{\mathbf{x},z_i}) - F(\mathbf{p}_{\mathbf{x},z_j}). \tag{6.43}$$

                End For
            End For
            Return the best merge that minimizes the impurity loss Δ_{ij}:

$$i^*, j^* = \min_{i,j} \Delta_{ij}. \tag{6.44}$$

        End While
    Step 3: Return K partitions.
Return: Return K partitions.

pseudo-code of our greedy-merge algorithm is provided in Algorithm 3.

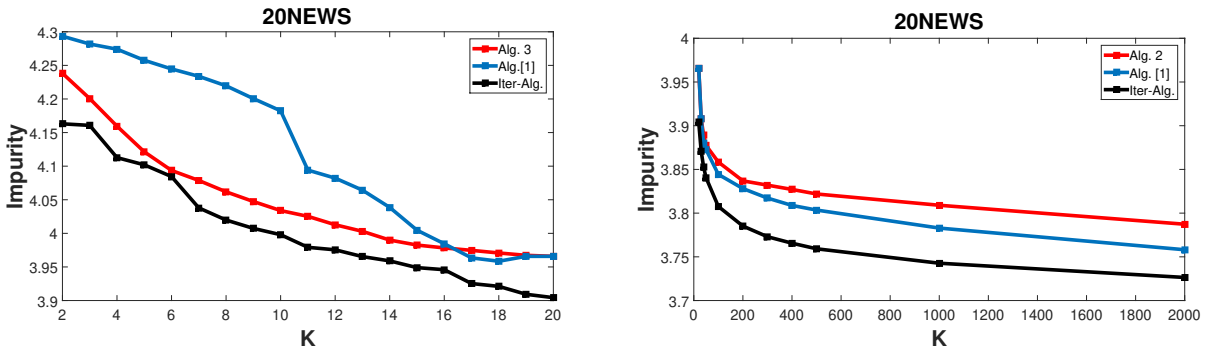
Running time of Algorithm 3: The running time of Step 1 and 2 in Algorithm 3 are NM and $(N-K)N^2$, respectively. Thus, the running time of Algorithm 3 is $O(NM + (N-K)N^2)$.

6.5.3 Reaching to the local optimal solutions

In [53], the authors proposed a necessary condition for which the partition is optimal (local or global). As the result, an iterative based k-means algorithm with a suitable distance can be used to find the local optimal partitions (please see Appendix 6.8.9 for the optimality conditions and Appendix 6.8.10 for the iterative algorithms). On the other hand, although our proposed algorithms can achieve a near-global optimal solution, there is no guarantee that the produced partitions are optimal i.e., the produced partitions might not satisfy the optimality condition in Theorem 6.10, Appendix 6.8.9 (Theorem 1 in [53]). Therefore, one can always perform the iterative algorithms in Appendix 6.8.10 over the partitions produced by Algorithm 1, Algorithm 2 and Algorithm 3 to achieve a local optimal solution. This optional step will improve the quality of our proposed algorithms at the expense of an additional time complexity $O(TKNM)$ where T is the number of iterations.

6.6 Numerical results

To evaluate the performance of the proposed algorithm, we used two datasets: 20NEWS and RCV1. These are widely used for evaluating text classification methods [1]. Existing algorithms [1], [53], [54] can only find locally optimal solutions. To approximate a globally optimal solution, many iterative algorithms use multiple random starting points and select the best solution. To that end, we compare the impurity provided by Algorithm 3 and Algorithm 2 when $K < N$

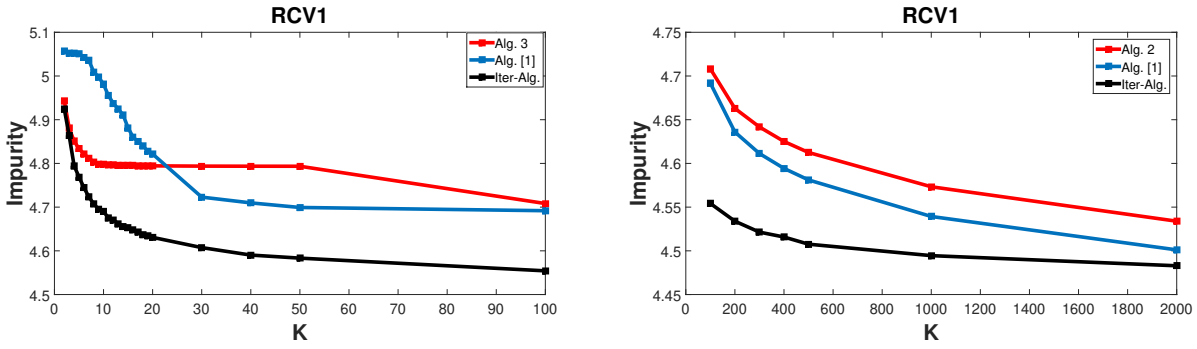


(a) (b)

Figure 6.5: Simulation results using 20NEWS dataset: (a) Algorithm 3 vs. the proposed algorithm in [1] when $K < N$; (b) Algorithm 2 vs. the proposed algorithm in [1] when $K \geq N$.

and $K \geq N$, respectively, with the impurity produced by running the iterative algorithms 100 times from 100 randomly starting points. The details of these iterative algorithms can be viewed in the Appendix 6.8.10 (Algorithm 4 for finding the optimal entropy impurity and Algorithm 5 for finding the optimal Gini index impurity, respectively). It is worth noting that Algorithm 4 is identical to the Divisible Clustering algorithm of Dhillon et al. [75] for finding the optimal entropy impurity. Although these iterative algorithms do not guarantee to find a globally optimal solution, their performances were shown in [75] to outperform the famous Agglomerative Clustering method in [95], [96].

20NEWS and RCV1 datasets are available online in `https://scikit-learn.org/stable/datasets/index.html#newsgroups-dataset`, and in `http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm`, respectively. In detail, 20NEWS includes 18.846 documents evenly divided into 20 disjoint classes and RCV1 includes 804,414 documents assigned to 103 different classes. Since both our algorithms and iterative algorithms use the joint distribution dataset, one wants to normalize the raw data in 20NEWS and RCV1 to a joint distribution $p(x_i, y_j)$, for example, by counting the number of times that a word y_j appears



(a) (b)

Figure 6.6: Simulation results using RCV1 dataset: (a) Algorithm 3 vs. the proposed algorithm in [1] when $K < N$; (b) Algorithm 2 vs. the proposed algorithm in [1] when $K \geq N$.

in document x_i. For convenience, we utilize the normalized datasets in [1]. After normalized, the dataset 20NEWS contains $M = 51840$ vectors of dimension $N = 20$ while the dataset RCV1 has $M = 170946$ vectors of dimension $N = 103$. The code as well as the datasets for testing are available in `https://github.com/hoangle96/linear_clustering`.

Next, we run the proposed algorithms (Algorithm 2 and 3 corresponding to the case of $K \geq N$ and $K < N$, respectively), the iterative algorithm, and the ratio-greedy in [1] for $K = 2, 3, 4, 5, \ldots, 2000$ using both 20NEWS dataset and RCV1 dataset. The entropy impurity of these algorithms are provided in Table 6.1 and Table 6.2 for 20NEWS and RCV1 datasets, respectively. Fig. 6.5 and Fig. 6.6 illustrate the impurity provided by our proposed Algorithm 2 and 3, the iterative algorithms, and the algorithm in [1] for 20NEWS and RCV1 datasets. As seen, the impurities provided by our proposed algorithms are very close to the impurity obtained from the iterative algorithms (see *Alg. 2,3/Iter-Alg.* columns in Table 6.1 and Table 6.2 and assuming that the iterative algorithm obtains a globally optimal solution). Particularly, the impurities provided by our proposed algorithms are at most 1.0181 time larger than the impurity provided by running the iterative algorithms 100 times using 20NEWS dataset and at most 1.0459

time larger for RCV1 dataset. In addition, the impurities provided by our algorithms (the red curves) are comparable to the impurities obtained by the proposed algorithm in [1] (the blue curves) as illustrated in Fig. 6.5 and Fig. 6.6. Particularly, Fig. 6.5 and Fig. 6.6 point out that our proposed algorithms outperforming the proposed algorithm in [1] if the number of partitions K is small, however, when K is large (for example $K \geq 19$ for 20NEWS and $K \geq 24$ for RCV1), the algorithm in [1] provides lower impurities. Finally, it is worth noting that our Algorithm 2 and 3 are linear in M while the proposed algorithm in [1] has a polynomial time complexity.

6.7 Conclusion

In this chapter, we proposed a guaranteed bounded linear time algorithm for minimizing a wide class of impurity function including entropy and Gini index. In some cases, we showed that the proposed algorithm is better than the state-of-art algorithms in both terms of computational complexity and the quality of partitioned outputs. Our upper bound and lower bound generalize two well-known results in information theory and signal processing, specially the Fano's inequality and the Boyd-Chiang upper bound of channel capacity. Both the theoretical and numerical results are provided to illustrate the advantages of our algorithm.

20NEWS dataset						
K	$e^{\max}$	$R(e^{\max})$	Alg. 2,3	Alg. in [1]	Iter-Alg.	Alg. 2,3/Iter-Alg.
2	0.0815	1.1915	4.2382	4.2934	4.1630	1.0181
3	0.1002	1.2933	4.2006	4.2818	4.1607	1.0096
4	0.1179	1.3844	4.1595	4.2741	4.1127	1.0114
5	0.1313	1.4512	4.1214	4.2576	4.1020	1.0047
6	0.1425	1.5062	4.0936	4.2447	4.0845	1.0022
7	0.1558	1.5698	4.0787	4.2337	4.0379	1.0101
8	0.1659	1.6174	4.0618	4.2196	4.0198	1.0105
9	0.1760	1.6642	4.0474	4.2006	4.0074	1.0100
10	0.1856	1.7087	4.0339	4.1824	3.9980	1.0090
11	0.1950	1.7517	4.0253	4.0941	3.9792	1.0116
12	0.2031	1.7885	4.0128	4.0822	3.9755	1.0094
13	0.2108	1.8234	4.0030	4.0643	3.9654	1.0095
14	0.2175	1.8535	3.9900	4.0386	3.9590	1.0078
15	0.2239	1.8824	3.9826	4.0047	3.9490	1.0085
16	0.2294	1.9070	3.9783	3.9845	3.9459	1.0082
17	0.2338	1.9264	3.9747	3.9635	3.9253	1.0126
18	0.2377	1.9437	3.9708	3.9585	3.9211	1.0127
19	0.2408	1.9578	3.9673	3.9658	3.9093	1.0148
20	0.2420	1.9630	3.9658	3.9658	3.9043	1.0158
30	0.2420	1.9630	3.9077	3.9081	3.8709	1.0095
40	0.2420	1.9630	3.8891	3.8798	3.8526	1.0095
50	0.2420	1.9630	3.8774	3.8726	3.8404	1.0096
100	0.2420	1.9630	3.8584	3.8443	3.8076	1.0134
200	0.2420	1.9630	3.8369	3.8281	3.7852	1.0137
300	0.2420	1.9630	3.8320	3.8172	3.7730	1.0156
400	0.2420	1.9630	3.8271	3.8088	3.7656	1.0163
500	0.2420	1.9630	3.8219	3.8035	3.7592	1.0167
1000	0.2420	1.9630	3.8090	3.7829	3.7427	1.0177
2000	0.2420	1.9630	3.7873	3.7581	3.7264	1.0163

Table 6.1: Simulation results using 20NEWS dataset for $K = 2, 3, 4, 5, \ldots, 2000$.

RCV1 dataset						
K	$e^{\max}$	$R(e^{\max})$	Alg. 2,3	Alg. in [1]	Iter-Alg.	Alg. 2,3/Iter-Alg.
2	0.1882	2.5376	4.9438	5.0571	4.9249	1.0038
3	0.2096	2.6681	4.8806	5.0525	4.8637	1.0035
4	0.2135	2.6915	4.8510	5.0520	4.7950	1.0117
5	0.2149	2.6999	4.8337	5.0512	4.7689	1.0136
6	0.2153	2.7024	4.8216	5.0429	4.7448	1.0162
7	0.2160	2.7068	4.8121	5.0367	4.7237	1.0187
8	0.2163	2.7081	4.8034	5.0089	4.7072	1.0204
9	0.2170	2.7124	4.7985	4.9976	4.6951	1.0220
10	0.2171	2.7131	4.7980	4.9810	4.6897	1.0231
11	0.2173	2.7142	4.7969	4.9551	4.6756	1.0260
12	0.2174	2.7148	4.7965	4.9373	4.6710	1.0269
13	0.2174	2.7152	4.7961	4.9245	4.6619	1.0288
14	0.2176	2.7161	4.7957	4.9107	4.6551	1.0302
15	0.2176	2.7164	4.7953	4.8811	4.6526	1.0307
16	0.2178	2.7173	4.7951	4.8607	4.6483	1.0316
17	0.2179	2.7178	4.7949	4.8497	4.6432	1.0327
18	0.2179	2.7180	4.7948	4.8403	4.6364	1.0342
19	0.2180	2.7183	4.7946	4.8277	4.6344	1.0346
20	0.2180	2.7187	4.7944	4.8215	4.6308	1.0353
30	0.2184	2.7207	4.7938	4.7229	4.6074	1.0405
40	0.2184	2.7212	4.7936	4.7097	4.5903	1.0443
50	0.2185	2.7214	4.7935	4.6992	4.5833	1.0459
100	0.2185	2.7215	4.7082	4.6918	4.5542	1.0338
200	0.2185	2.7215	4.6629	4.6359	4.5339	1.0285
300	0.2185	2.7215	4.6417	4.6113	4.5214	1.0266
400	0.2185	2.7215	4.6251	4.5943	4.5158	1.0242
500	0.2185	2.7215	4.6125	4.5811	4.5076	1.0233
1,000	0.2185	2.7215	4.5733	4.5393	4.4945	1.0175
2,000	0.2185	2.7215	4.5339	4.5011	4.4829	1.0114

Table 6.2: Simulation results using RCV1 dataset for $K = 2, 3, 4, 5, \ldots, 2000$.

6.8 Appendix

6.8.1 Improvement of Algorithm in [1]

In Section V [1], Cicalese et al. proposed an algorithm that provably achieves near-optimal partition for entropy impurity. This algorithm has two steps: (1) performing a projection to transfer the multidimensional data back to a 2-dimensional data, and (2) using dynamic programming to find the optimal partition in 2-dimensional data based on the idea in [44].

Cicalese et al. proved that the running time of the algorithm in [1] is polynomial, however, no precise complexity is constructed. Since the running time of projection the original data to a 2-dimension data is NM and the running time of finding the optimal partition in 2-dimensional space using the method in [44] is $O(M^3)$, the time complexity of the algorithm in [1] should be at least $O(NM + M^3)$.

Based on the well-known SMAWK algorithm [87], we show that the running time of the algorithm in [1] can be further reduced from $O(M^3)$ to $O(M \log M)$. Indeed, the SMAWK algorithm can be applied to reduce the running time of algorithm in [44] to $O(KM)$ if the binary data is ordered (see [47] and [50] for detail). However, to order a data of size M, the fastest sorting technique requires the running time of $O(M \log M)$. Thus, the problem in [44] can be solved in $O(M \log M)$ that reduces the polynomial time complexity in step (2) of algorithm in [1] to $O(M \log M)$. The total time complexity of the proposed algorithm in [1], therefore, is $O(NM + M^3)$.

6.8.2 Jensen's Inequality

Jensen inequality states that for a random variable T, then $\mathbb{E}[f(T)] \geq f(\mathbb{E}[T])$ if $f(x)$ is convex, and $\mathbb{E}[f(T)] \leq f(\mathbb{E}[T])$ if $f(x)$ is concave.

Now, let $T \in \{t_1, t_2, \ldots, t_K\}$ be a random variable with the uniform distribution $\frac{1}{K}$. If $f(x)$ is concave, using Jensen's inequality:

$$f(\frac{\sum_{i=1}^{K} t_i}{K}) \geq \sum_{i=1}^{K} \frac{1}{K} f(t_i),$$

which is equivalent to

$$K f(\frac{\sum_{i=1}^{K} t_i}{K}) \geq \sum_{i=1}^{K} f(t_i). \tag{6.45}$$

Thus, (6.13) is a direct result of (6.45) using $t_i = p(x_i|z_j)$ and $K = N - 1$.

6.8.3 Fano's Inequality

If the impurity function is entropy i.e., $f(x) = -x \log x$, then

$$I_Q = \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_k) \Big(- p(x_i|z_k) \log(p(x_i|z_k)) \Big) = \sum_{k=1}^{K} p(z_k) H(X|z_k) = H(X|Z).$$

By plugging $f(x) = -x \log x$ into (6.16)

$$\begin{aligned} H(X|Z) &\leq -e_Q \log e_Q - (N-1) \frac{1-e_Q}{N-1} \log \frac{1-e_Q}{N-1} \\ &= -[e_Q \log e_Q + (1-e_Q) \log(1-e_Q)] + (1-e_Q) \log(N-1) \qquad (6.46) \\ &= H(1-e_Q) + (1-e_Q) \log(N-1), \qquad (6.47) \end{aligned}$$

with (6.46) is due to a bit of algebra, (6.47) is due to the binary entropy function is symmetric, i.e., $H(e_Q) = H(1-e_Q) = -[e_Q \log e_Q + (1-e_Q) \log(1-e_Q)]$.

Let us now consider X and Z as two random variables that represent the input and the output symbols of a communication channel. Errors might occur during transmissions. Suppose

that the receiver estimates the transmitted symbol based on the received z_k as x_{k^*} where $k^* = \arg\max_i p(x_i|z_k)$ (maximum likelihood decoding). Thus, the error probability of this decoding scheme is $P_e = 1 - \sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k) = 1 - e_Q$. Then,

$$\begin{aligned} H(X|Z) &\leq H(1-e_Q) + (1-e_Q)\log(N-1) \\ &= H(P_e) + P_e \log(|X|-1), \end{aligned}$$

which is identical to the well-known Fano's inequality [3].

6.8.4 Boyd-Chiang Upper Bound of Channel Capacity

The mutual information $I(X;Z)$ between channel input and channel output is defined by $I(X;Z) = H(X) - H(X|Z)$. However, from Theorem 6.4, if the impurity function is entropy i.e., $f(x) = -x\log(x)$ and $l(x) = -\log(x)$ then $H(X|Z) = I_Q \geq l(e_Q) = -\log(e_Q)$. Now, by using the uniform input distribution,

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &\leq H(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}) - l(\sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k)) && (6.48) \\ &= \log N - l(\sum_{k=1}^{K} p(x_{k^*})p(z_k|x_{k^*})) && (6.49) \\ &= \log N - l(\frac{1}{N}\sum_{k=1}^{K} p(z_k|x_{k^*})) && (6.50) \\ &= \log N + \log(\frac{\sum_{k=1}^{K} p(z_k|x_{k^*})}{N}) && (6.51) \\ &= \log(\sum_{k=1}^{K} p(z_k|x_{k^*})), && (6.52) \end{aligned}$$

with (6.48) is due to $H(X|Z) \geq l(e_Q)$ and $e_Q = \sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k)$, (6.49) is due to Bayes's theorem, (6.50) is due to the input distribution is uniform, (6.51) is due to $l(x) = -\log(x)$, (6.52) is due to a bit of algebra.

Let us now consider $X =$ and Z as two random variables that represent the input and the output symbols of a communication channel. Due to the errors during transmissions, a channel matrix A whose entry $A_{ij} = p(z_j|x_i)$ denotes the probability of the transmitter transmitted symbol x_i but the receiver decoded to symbol z_j. Now, since the input distribution is uniform, from $p(z_k)p(x_{k^*}|z_k) = p(x_{k^*})p(z_k|x_{k^*})$, then $p(z_k|x_{k^*})$ is the largest entry in k^{*th} column of channel matrix. Thus, the upper bound of channel capacity is $\log(\sum_{k=1}^{K} p(z_k|x_{k^*}))$ that is identical to the bound constructed by Boyd and Chiang [13].

6.8.5 Proof of Theorem 6.2

Proof. We show that $u(e_Q) = f(e_Q) + (N-1)f(\frac{1-e_Q}{N-1})$ is a non-increasing function.

$$\begin{aligned} u'(e_Q) &= f'(e_Q) - (N-1)\frac{1}{N-1}f'(\frac{1-e_Q}{N-1}) && (6.53) \\ &= f'(e_Q) - f'(\frac{1-e_Q}{N-1}). && (6.54) \end{aligned}$$

Since $k^* = \arg\max_i p(x_i|z_k)$ and $\sum_{i=1}^{N} p(x_i|z_k) = 1$, $p(x_{k^*}|z_k) \geq \frac{1}{N}$. Thus,

$$e_Q = \sum_{k=1}^{K} p(z_k)p(x_{k^*}|z_k) \geq \sum_{k=1}^{K} p(z_k)\frac{1}{N} = \frac{1}{N}.$$

Therefore, $1 - e_Q \leq \frac{N-1}{N}$. Thus,

$$e_Q \geq \frac{1}{N} \geq \frac{1-e_Q}{N-1}. \tag{6.55}$$

Now since $f'(e_Q)$ is a non-increasing function due to $f(e_Q)$ is concave. Therefore,

$$u'(e_Q) = f'(e_Q) - f'(\frac{1-e_Q}{N-1}) \leq 0. \tag{6.56}$$

Or, $u(e_Q)$ is a non-increasing function.

Finally, it is possible to verify that if $e_Q = \frac{1}{N}$ or $e_Q = 1$, then the upper bound is tight i.e., $u(e_Q) = I_Q$ for both the entropy impurity and the Gini index impurity. Indeed, if $e_Q = \frac{1}{N}$, $p(x_i|z_k) = \frac{1}{N}$ $\forall i, k$ and then $u(e_Q) = I_Q = Nf(\frac{1}{N})$. If $e_Q = 1$, $p(x_{k^*}|z_k) = 1$ and $p(x_i|z_k) = 0$ $\forall$ $i \neq k^*$ and then $u(e_Q) = I_Q = 0$. □

6.8.6 Proof of Theorem 6.3

6.8.6.1 Proof of Theorem 6.3-(a)

Proof. We first consider the case when $K = N$, we show that $e_{Q_{e^{\max}}} \geq e_Q, \forall Q$. We have:

$$\begin{aligned}
e_{Q_{e^{\max}}} &= \sum_{j^*=1}^{K} p(z_{j^*}) \max_i p(x_i|z_{j^*}) \\
&= \sum_{j^*=1}^{K} \max_i p(x_i, z_{j^*}) \\
&= \sum_{j^*=1}^{K} \sum_{j:Q(y_j)=z_{j^*}} \max_i p(x_i, y_j) \\
&\geq \sum_{j=1}^{K} \sum_{j:Q(y_j)=z_j} p(x_i, y_j) \\
&= e_Q.
\end{aligned}$$

Note that $Q(y_j) = z_j$ in the index of the sum in the last equation represents any arbitrary partition scheme.

We now show that if a quantizer Q produces $e_{\max} = \max_Q e_Q$ then it must has the structure of $Q_{e^{\max}}$. We will prove this by contradiction. Suppose that a quantizer Q produces the partitions $z_1, z_2, \ldots, z_K$ that has $e^{\max}$, but there exists a y_n that is partitioned to z_l, such that $l \neq \arg\max_{1 \leq i \leq N} p(x_i, y_n)$. Let $m = \arg\max_{1 \leq i \leq N} p(x_i, y_n)$. Now, let consider a quantizer Q' which is constructed from quantizer Q by moving y_n from z_l to z_m. This new quantizer Q' produces a new partition $\{z'_1, \ldots, z'_l, \ldots, z'_m, \ldots, z'_K\}$ with $z'_k = z_k$, $\forall\ k \neq l, m$, with corresponding $p'(x_i, z_k)$ and $p'(x_i|z_k)$.

From the definition of e_Q, we have:

$$\begin{aligned}
e_Q - e'_Q &= p(z_l)p(x_{l*}|z_l) + p(z_m)p(x_{m^*}|z_m) - p'(z_l)p'(x_{l*}|z_l) - p'(z_m)p'(x_{m^*}|z_m) \\
&= p(x_{l*}, z_l) + p(x_{m^*}, z_m) - p'(x_{l*}, z_l) - p'(x_{m^*}, z_m) \\
&= \sum_{j:Q(y_j)=z_l} p(x_{l*}, y_j) + \sum_{j:Q(y_j)=z_m} p(x_{m*}, y_j) \\
&- \sum_{j:Q(y_j)=z_l} p'(x_{l*}, y_j) - \sum_{j:Q(y_j)=z_m} p'(x_{m*}, y_j) \\
&= p(x_{l*}, y_n) - p(x_{m*}, y_n).
\end{aligned}$$

Since by assumption that $p(x_m, y_n) > p(x_l, y_n)$, we have $e_{Q'} < e_Q$ which is a contradiction. Thus, any partition scheme that achieves $e^{\max}$ must have the structure of maximum likelihood of $Q_{e^{\max}}$. □

6.8.6.2 Proof of Theorem 6.3-(b)

Proof. Theorem 6.3-(a) handled the case when $K = N$ and showed that any partition scheme that achieves $e^{\max}$ must have the structure of maximum likelihood of $Q_{e^{\max}}$. On the other hand, Theorem 6.3-(b) finds the partition that achieves $e^{\max}$ when $K > N$. Interestingly, we show that the mapping of $Q_{e^{\max}}$ in Theorem 6.3-(a) that partitions the data to N-nonempty partitions and $K - N$ empty partitions still produces $e^{\max}$. For detail. let $j^* = \arg\max_i p(x_i, y_j)$ and define quantizer $Q_{e^{\max}}$ with the following structure:

$$Q_{e^{\max}}(y_j) = z_{j^*}, \tag{6.57}$$

then $Q_{e^{\max}}$ produces $e^{\max} = \max_Q e_Q$ even if $K > N$. Moreover, due to the mapping in (6.57), $Q_{e^{\max}}$ produces N nonempty partitions and $K - N$ empty partitions.

Indeed, suppose the quantizer $Q_{e^{\max}}$ produces K' nonempty partitions and $K \geq K' > N$. We show that existing another quantizer Q having exactly N nonempty partitions which still can produces the same $e^{\max}$ as $Q_{e^{\max}}$. Now, since $K' > N$, existing two partition z_i, z_j such that:

$$i^* = j^* = \arg\max_{1 \leq t \leq N} p(x_t|z_i) = \arg\max_{1 \leq t \leq N} p(x_t|z_j).$$

Next, consider a new quantizer Q that maps two partitions z_i and z_j into a single partition z_k, we show that Q still provide the same $e^{\max}$ as $Q_{e^{\max}}$. Indeed, since $i^* = j^*$ and $z_i \cup z_j = z_k$, $z_i \cap z_j = \emptyset$, we have:

$$i^* = j^* = k^* = \arg\max_{1 \leq t \leq N} p(x_t|z_k),$$

and

$$p(x_{k^*}, z_k) = p(x_{i^*}, z_i) + p(x_{j^*}, z_j).$$

Thus,

$$\begin{aligned} p(z_k)p(x_{k^*}|z_k) &= p(x_{k^*}, z_k) = p(x_{i^*}, z_i) + p(x_{j^*}, z_j) && (6.58) \\ &= p(z_i)p(x_{i^*}|z_i) + p(z_j)p(x_{j^*}|z_j). && (6.59) \end{aligned}$$

By definition of e_Q in (6.6) and noting that Q is identical to $Q_{e^{\max}}$ except that two partitions z_i and z_j are grouped into a single partition z_k, $e_Q = e^{\max}$. By induction method, after at most $K' - N$ times grouping, existing a quantizer Q having exactly N nonempty partitions which still can produce $e^{\max}$. Moreover, this quantizer satisfies the mapping in (6.57). The proof is complete. □

6.8.7 Proof of Theorem 6.7

Proof. For the Gini index impurity function, $f(x) = x(1-x)$ and $l(x) = 1-x$. Thus,

$$\begin{aligned} R(e^{\max}) &= \frac{f(e^{\max}) + (N-1)f(\frac{1-e^{\max}}{N-1})}{l(e^{\max})} \\ &= \frac{e^{\max}(1-e^{\max}) + (N-1)\frac{1-e^{\max}}{N-1}(1-\frac{1-e^{\max}}{N-1})}{1-e^{\max}} \quad (6.60) \\ &= e^{\max} + 1 - \frac{1-e^{\max}}{N-1} \quad (6.61) \\ &\leq e^{\max} + 1 \quad (6.62) \\ &\leq 2, \quad (6.63) \end{aligned}$$

with (6.60) due to $f(x) = x(1-x)$ and $l(x) = 1-x$, (6.61) and (6.62) due to a bit of algebra, (6.63) due to $e^{\max} \leq 1$. Noting that one can use $e^{\max}+1$ as another approximation for Gini index impurity. □

6.8.8 Proof of Theorem 6.8

Proof. For entropy impurity, $f(x) = -x\log(x)$ and $l(x) = -\log(x)$, plugin the upper bound and the lower bound in Theorem 6.1 and Theorem 6.4, we have

$$R(e^{\max}) = \frac{H(e^{\max}) + (1-e^{\max})\log(N-1)}{-\log(e^{\max})}. \quad (6.64)$$

Since $\log(N-1) < \log N$, we have

$$R(e^{\max}) = \frac{H(e^{\max}) + (1-e^{\max})\log(N-1)}{-\log(e^{\max})} < \frac{H(e^{\max}) + (1-e^{\max})\log N}{-\log(e^{\max})}. \quad (6.65)$$

To prove Theorem 6.8, we want to show that the inequality below holds

$$\frac{H(e^{\max}) + (1 - e^{\max}) \log N}{-\log(e^{\max})} \leq \log^2 N, \forall N \geq N^{\min}. \tag{6.66}$$

This is equivalent to show that

$$\log^2 N(-\log(e^{\max})) - (1 - e^{\max}) \log N - H(e^{\max}) \geq 0, \forall N \geq N^{\min}.$$

Indeed, using a bit of algebra,

$$\begin{aligned}
& \log^2 N(-\log(e^{\max})) - (1 - e^{\max}) \log N - H(e^{\max}) \\
= \; & -\log(e^{\max})[\log^2 N - 2 \log N \frac{1 - e^{\max}}{2(-\log(e^{\max}))} + (\frac{1 - e^{\max}}{2(-\log(e^{\max}))})^2 - \frac{H(e^{\max})}{-\log(e^{\max})} - (\frac{1 - e^{\max}}{2(-\log(e^{\max}))})^2] \\
= \; & -\log(e^{\max})[(\log N - \frac{1 - e^{\max}}{2(-\log(e^{\max}))})^2 - \frac{4H(e^{\max})(-\log(e^{\max})) + (1 - e^{\max})^2}{(-2\log(e^{\max}))^2}] \\
= \; & -\log(e^{\max})[(\log N - \frac{1 - e^{\max}}{2(-\log(e^{\max}))})^2 - (\frac{\sqrt{4H(e^{\max})(-\log(e^{\max})) + (1 - e^{\max})^2}}{-2\log(e^{\max})})^2].
\end{aligned}$$

Now, if

$$\log N \geq \frac{1 - e^{\max}}{-2\log(e^{\max})} + \frac{\sqrt{4H(e^{\max})(-\log(e^{\max})) + (1 - e^{\max})^2}}{-2\log(e^{\max})} = S(e^{\max}), \tag{6.67}$$

then (6.66) holds. Thus, $R(e^{\max}) < \log^2 N$ holds if $N \geq 2^{S(e^{\max})} = N^{\min}$. □

6.8.9 Well-known results on minimizing impurity partitions

This section summarizes two important results that were stated in [53], [51]. To summarize these well-known results, we first need to rewrite the original impurity function which is defined in (6.1).

Reformulation of the impurity function:

For ease of analysis, in this chapter, we rewrite the impurity function I_Q in term of the joint pmf $\mathbf{p}_{\mathbf{x},z_k} = [p(x_1, z_k), p(x_2, z_k), \ldots, p(x_N, z_k)]$ of X and $Z = z_k$, for $k = 1, 2, \ldots, K$. Specifically,

$$\begin{aligned} I_Q &= \sum_{k=1}^{K}\sum_{i=1}^{N} p(z_k) f(p(x_i|z_k)) = \sum_{k=1}^{K} p(z_k) \sum_{i=1}^{N} f(p(x_i|z_k)) && (6.68)\\ &= \sum_{k=1}^{K}\sum_{i=1}^{N} p(x_i, z_k) \sum_{i=1}^{N} f\Big(\frac{p(x_i, z_k)}{\sum_{i=1}^{N} p(x_i, z_k)}\Big) && (6.69)\\ &= \sum_{i=1}^{K} F(\mathbf{p}_{\mathbf{x},z_k}), && (6.70) \end{aligned}$$

where

$$F(\mathbf{p}_{\mathbf{x},z_k}) = \sum_{i=1}^{N} p(x_i, z_k) \sum_{i=1}^{N} f\Big(\frac{p(x_i, z_k)}{\sum_{i=1}^{N} p(x_i, z_k)}\Big) \tag{6.71}$$

is a function of the joint distribution vector $\mathbf{p}_{\mathbf{x},z_k} = [p(x_1, z_k), p(x_2, z_k), \ldots, p(x_N, z_k)]$ which specifies the impurity measured in the cluster that represents $Z = z_k$. Clearly that the total impurity I_Q produced by quantizer Q, is the summation overall the impurity in each cluster. Thus, I_Q also can be viewed as a function of the joint distribution vector $\mathbf{p}_{\mathbf{x},z_k}$ for $k = 1, 2, \ldots, K$.

Theorem 6.9. ***(Impurity gain after partition is always non-negative)***

If $\mathbf{p}_{\mathbf{x},z_i} = \mathbf{p}_{\mathbf{x},z_j} + \mathbf{p}_{\mathbf{x},z_k}$, *then*

$$F(\mathbf{p}_{\mathbf{x},z_i}) \geq F(\mathbf{p}_{\mathbf{x},z_j}) + F(\mathbf{p}_{\mathbf{x},z_k}). \tag{6.72}$$

In other words, for an arbitrary set A, if $A = B \cup C$ and $B \cap C = \emptyset$, then the impurity in A is larger or at least equal the total impurity in B and C.

The proof can be viewed in [53], Proposition 1. Since the impurity gain after partitioning is always non-negative, the optimal impurity in K'-partition is always less than or at least equal the

optimal impurity in K-partition if $K' \geq K$. We use this property to prove that the theoretical bound for K-partition is always better or at least equal the theoretical bound for N-partition if $K > N$.

Theorem 6.10. ***(Necessary optimality condition)***

Let Q be a quantizer with an induced K-partition corresponding to K joint pmf vectors $\mathbf{p}_{\mathbf{x},z_i}$, $i = 1, 2, \ldots, K$. For each partition z_k, $k = 1, 2, \ldots, K$, define:

$$\mathbf{c}_k = \frac{\mathrm{d}F(\mathbf{p}_{\mathbf{x},z_k})}{\mathrm{d}\mathbf{p}_{\mathbf{x},z_k}}, \tag{6.73}$$

Define the "distance" from a data point y_j to the cluster z_k as:

$$d(y_j, z_k) = \mathbf{c}_k^T \mathbf{p}_{\mathbf{x},y_j}, \tag{6.74}$$

then an optimal quantizer Q^* that quantizes y_j to z_k must have $d(y_j, z_k) \leq d(y_j, z_l)$, $l \neq k$.

The proof can be viewed in [53], Theorem 1. Based on the necessary optimality condition, a locally optimal solution can be found using an iterative algorithm that begins from a randomly assigned partition and alternatively updates the cluster members based on their distances. This algorithm is very similar to a k-means algorithm using a "distance" $d(y_j, z_k)$ which is defined in (6.74). The running time of this iterative algorithm, therefore, is $O(TKMN)$ where T is the number of iterations. The detail of the iterative algorithms for entropy impurity and Gini index impurity can be found in Appendix 6.8.10. Since the condition in (6.74) is necessary but not sufficient, the iterative algorithm ensures a locally optimal solution rather than a globally optimal solution. On the other hand, the proposed algorithm in this chapter guarantees a near-global optimal solution while it does not guarantee a locally optimal solution. To improve the splitting quality a bit, the iterative algorithm can be applied over the partitions produced by

our approximation algorithms to achieve the local optimal solution at the cost of an additional time complexity $O(TKNM)$. Similar to the approach in [1], to evaluate the performance of the proposed algorithms, one can approximate the global optimal partition by running the iterative algorithms from many randomly starting points and select the best solution. The detail of this approach can be viewed in Section 6.6.

6.8.10 Finding the optimal partition via iterative algorithms

It is well-known that the problem of finding the optimal impurity partition can be solved using iterative algorithms [53], [68]. For example, finding an optimal quantizer that minimizes the entropy impurity is equivalent to finding the optimal partition that minimizes the KL-divergence distance [1], [54]. Thus, existing iterative algorithms [54], [75] can be applied to find a locally optimal partition that minimizes entropy impurity. These algorithms are based on the famous k-means algorithms which use the KL-divergence as the distance metric. For convenience, we refer the reader to [54], [75] for more details of the iterative algorithms that find the optimal entropy impurity partitions. The pseudo code is shown in Algorithm 4. Similarly, a locally optimal quantizer that minimizes the Gini index impurity can be determined by an iterative algorithm using a suitable distance metric. Based on the general iterative algorithms in [53], [68], Algorithm 5 is constructed to find the locally optimal solution for Gini index impurity.

We recall that

$$\mathbf{p}_{\mathbf{x},z_k} = (p(x_1, z_k), p(x_2, z_k), \ldots, p(x_N, z_k)),$$

$$\mathbf{p}_{\mathbf{x}|z_k} = (p(x_1|z_k), p(x_2|z_k), \ldots, p(x_N|z_k)),$$

$$\mathbf{p}_{\mathbf{x},y_j} = (p(x_1, y_j), p(x_2, y_j), \ldots, p(x_N, y_j)),$$

and

$$\mathbf{p}_{\mathbf{x}|y_j} = (p(x_1|y_j), p(x_2|y_j), \ldots, p(x_N|y_j)).$$

Algorithm 4 and Algorithm 5 work as follows. In the initial step, these algorithms randomly assign y_j to z_k using a random quantizer Q. In step 1, based on the initial random clustering, the joint pmf of X and Z $\mathbf{p}_{\mathbf{x},z_k}$ and the conditional pmf of X and Z $\mathbf{p}_{\mathbf{x}|z_k}$ are computed $\forall k$. In step 2, from $\mathbf{p}_{\mathbf{x},z_k}$, $\mathbf{p}_{\mathbf{x}|z_k}$, $\mathbf{p}_{\mathbf{x},y_j}$ and $\mathbf{p}_{\mathbf{x}|y_j}$, the distance $d(y_j, z_k)$ between y_j and z_k is computed. Noting that the distance $d(y_j, z_k)$ in (6.79) and (6.83) are constructed separately from the same general form in [91]. Based on $d(y_j, z_k)$, the membership of y_j to each z_k is updated such that $Q(y_j) = z_k$ if $d(y_j, z_k)$ is the smallest over all z_k. Algorithm 4 and Algorithm 5 are similar to the famous k-means algorithm with the computational complexity of $O(TNKM)$ where T is the number of iterations.

Algorithm 4 Iterative algorithm finding optimal partitions for entropy impurity.

1: **Input**: Dataset $Y = \{y_1, \ldots, y_M\}$ and $p(x_i, y_j)$.
2: **Output**: Partition $Z = \{z_1, z_2, \ldots, z_K\}$.
3: **Initialization**: Randomly cluster y_j into K clusters $z_1, z_2, \ldots, z_K$, i.e., choose a random quantizer Q.
4: **Step 1**: Compute $\mathbf{p}_{\mathbf{x},z_k}$, $\mathbf{p}_{\mathbf{x}|z_k}$ and $\mathbf{p}_{\mathbf{x}|y_j}$.

$$\mathbf{p}_{\mathbf{x},z_k} = \sum_{j:Q(y_j)=z_k} \mathbf{p}_{\mathbf{x},y_j}. \tag{6.75}$$

$$\mathbf{p}_{\mathbf{x}|z_k} = \frac{\mathbf{p}_{\mathbf{x},z_k}}{\mathbf{p}_{\mathbf{x},z_k}^T \mathbf{1}}. \tag{6.76}$$

$$\mathbf{p}_{\mathbf{x}|y_j} = \frac{\mathbf{p}_{\mathbf{x},y_j}}{\mathbf{p}_{\mathbf{x},y_j}^T \mathbf{1}}. \tag{6.77}$$

5: **Step 2**: For each y_j, compare the distance $d(y_j, z_k)$ from y_j to each partition z_k. y_j belongs to the partition with the smallest distance.

$$Q(y_j) = \arg\min_{z_k} d(y_j, z_k), \tag{6.78}$$

where

$$\begin{aligned} d(y_j, z_k) &= D_{KL}(\mathbf{p}_{\mathbf{x}|y_j} || \mathbf{p}_{\mathbf{x}|z_k}) & (6.79) \\ &= \sum_{i=1}^{N} p(x_i|y_j) \log(\frac{p(x_i|y_j)}{p(x_i|z_k)}). & (6.80) \end{aligned}$$

6: **Step 3**: Go to Step 1 until the partitions $z_1, z_2, \ldots, z_K$ stop changing (membership of all y_j does not change), or the maximum number of iterations has been reached.

Algorithm 5 Iterative algorithm finding optimal partitions for Gini index impurity.

1: **Input**: Dataset $Y = \{y_1, \ldots, y_M\}$ and $p(x_i, y_j)$.
2: **Output**: Partition $Z = \{z_1, z_2, \ldots, z_K\}$.
3: **Initialization**: Randomly cluster y_j into K clusters $z_1, z_2, \ldots, z_K$, i.e., choose a random quantizer Q.
4: **Step 1**: Compute the joint distribution $\mathbf{p}_{\mathbf{x},z_k}$ in each cluster z_k.

$$\mathbf{p}_{\mathbf{x},z_k} = \sum_{j:Q(y_j)=z_k} \mathbf{p}_{\mathbf{x},y_j}. \tag{6.81}$$

5: **Step 2**: For each y_j, compare the distance $d(y_j, z_k)$ from y_j to each partition z_k. y_j belongs to the partition with the smallest distance.

$$Q(y_j) = \arg\min_{z_k} d(y_j, z_k), \tag{6.82}$$

where

$$d(y_j, z_k) = \sum_{i=1}^{N} p(x_i, y_j)\Big(1 - 2\frac{p(x_i, z_k)}{\sum_{n=1}^{N} p(x_n, z_k)} + \frac{\sum_{n=1}^{N} p^2(x_n, z_k)}{(\sum_{n=1}^{N} p(x_n, z_k))^2}\Big). \tag{6.83}$$

6: **Step 3**: Go to Step 1 until the partitions $z_1, z_2, \ldots, z_K$ stop changing (membership of all y_j does not change), or the maximum number of iterations has been reached.

Chapter 7: Conclusion

The works in this dissertation aim to (1) finding the closed-form expression for a good upper bound on capacities of discrete memoryless channels together with optimality conditions for which the upper bound is precisely the channel capacity, (2) designing the optimal quantizer that maximizes mutual information between the input and the quantized-output of a communication channel, (3) finding the optimal quantizer structure maximizing mutual information under the quantized-output constraints, (4) finding the maximum value of mutual information (channel capacity) over both the input distribution and the quantization parameters, and (5) designing efficient algorithms for finding the optimal impurity partition. Particularly, the closed-form expression for capacities and upper bounds of discrete memoryless channels are investigated in Chapter 2. Chapter 3 is dedicated to finding the optimal structure of quantizers that maximizing the mutual information between the input and the quantized-output. Chapter 4 extends the results in Chapter 3 to design the optimal quantizers that maximize mutual information between the input and the quantized-output under quantized-output constraints. Chapter 5 establishes the fundamental results for finding the channel capacity over both the input distribution and the quantization parameter variables. Finally, a guaranteed approximation algorithm for minimizing a wide class of impurity function is proposed in Chapter 6. In the future, based on the results of this dissertation, I would like to investigate a few problems related to what I am pursuing. These problems will focus on both communication theory and information theory together with its applications in machine learning and signal processing, including but not limited to develop error-correcting code for 5G and 6G telecommunication networks, efficient learning algorithms over noisy channels, and low-complexity approximation algorithms for information-theoretic learning frameworks.

Bibliography

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

[2] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

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

[4] Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

[5] Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

[6] Saburo Muroga. On the capacity of a discrete channel, mathematical expression of capacity of a channel which is disturbed by noise in its every one symbol and expressible in one state diagram. *Journal of the Physical Society of Japan*, 8(4):484–494, 1953.

[7] Claude Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):8–19, 1956.

[8] B Robert. Ash. information theory, 1990.

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

[10] Keye Martin, Ira S Moskowitz, and Gerard Allwein. Algebraic information theory for binary channels. *Theoretical Computer Science*, 411(19):1918–1927, 2010.

[11] Xue-Bin Liang. An algebraic, analytic, and algorithmic investigation on the capacity and capacity-achieving input probability distributions of finite-input–finite-output discrete memoryless channels. *IEEE Transactions on Information Theory*, 54(3):1003–1023, 2008.

[12] Paul Cotae, Ira S Moskowitz, and Myong H Kang. Eigenvalue characterization of the capacity of discrete memoryless channels with invertible channel matrices. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1–6. IEEE, 2010.

[13] Mung Chiang and Stephen Boyd. Geometric programming duals of channel capacity and rate distortion. *IEEE Transactions on Information Theory*, 50(2):245–258, 2004.

[14] Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[15] Abhishek Sinha. Convex optimization methods for computing channel capacity. 2014.

[16] Frédéric Dupuis, Wei Yu, and Frans MJ Willems. Blahut-arimoto algorithms for computing channel capacity and rate-distortion with side information. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, page 179. IEEE, 2004.

[17] Gerald Matz and Pierre Duhamel. Information geometric formulation and interpretation of accelerated blahut-arimoto-type algorithms. In *Information theory workshop, 2004. IEEE*, pages 66–70. IEEE, 2004.

[18] Yaming Yu. Squeezing the arimoto–blahut algorithm for faster convergence. *IEEE Transactions on Information Theory*, 56(7):3149–3157, 2010.

[19] Bernd Meister and Werner Oettli. On the capacity of a discrete, constant channel. *Information and Control*, 11(3):341–351, 1967.

[20] Masakazu Jimbo and Kiyonori Kunisawa. An iteration method for calculating the relative capacity. *Information and Control*, 43(2):216–223, 1979.

[21] Thai Duong, Duong Nguyen-Huu, and Thinh Nguyen. Location assisted coding (lac) embracing interference in free space optical communications. In *Proceedings of the 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, pages 107–114, 2015.

[22] Claude E Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois press, 1998.

[23] T Cover. An achievable rate region for the broadcast channel. *IEEE Transactions on Information Theory*, 21(4):399–404, 1975.

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

[25] Eric W Weisstein. Gershgorin circle theorem. 2003.

[26] Miroslav Fiedler and Vlastimil Pták. Diagonally dominant matrices. *Czechoslovak Mathematical Journal*, 17(3):420–433, 1967.

[27] Thuan Nguyen and Thinh Nguyen. Relay-miso channel. Available at `http://ir.library.oregonstate.edu/concern/articles/tb09jb69h`, 2018.

[28] Thomas Cover and A EL Gamal. Capacity theorems for the relay channel. *IEEE Transactions on information theory*, 25(5):572–584, 1979.

[29] Boris Rankov and Armin Wittneben. Achievable rate regions for the two-way relay channel. In *Information theory, 2006 IEEE international symposium on*, pages 1668–1672. IEEE, 2006.

[30] Stephen Boyd, Seung-Jean Kim, Lieven Vandenberghe, and Arash Hassibi. A tutorial on geometric programming. *Optimization and engineering*, 8(1):67, 2007.

[31] Fady Alajaji and Po-Ning Chen. *An Introduction to Single-User Information Theory*. Springer, 2018.

[32] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[33] Rayleigh quotient and the min-max theorem. Available at `http://www.math.toronto.edu/mnica/hermitian2014.pdf`, 2014.

[34] Charles R Johnson. A gersgorin-type lower bound for the smallest singular value. *Linear Algebra and its Applications*, 112:1–7, 1989.

[35] YP Hong and C-T Pan. A lower bound for the smallest singular value. *Linear Algebra and its Applications*, 172:27–32, 1992.

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

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

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

[39] 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.

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

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

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

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

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] X. He, K. Cai, W. Song, and Z. Mei. Dynamic programming for quantization of q-ary input discrete memoryless channels. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 450–454, 2019.

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

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

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

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

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

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

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

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

[64] 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.

[65] 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.

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

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

[68] 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.

[69] 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.

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

[71] 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.

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

[73] 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.

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

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

[76] R Silverman. On binary channels and their cascades. *IRE Transactions on Information Theory*, 1(3):19–27, 1955.

[77] T. Nguyen and T. Nguyen. On bounds and closed-form expressions for capacities of discrete memoryless channels with invertible positive matrices. *IEEE Transactions on Vehicular Technology*, 69(9):9910–9920, 2020.

[78] Thuan Nguyen and Thinh Nguyen. Thresholding quantizer design for mutual information maximization under output constraint. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5. IEEE, 2020.

[79] 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.

[80] Thuan Nguyen and Thinh Nguyen. On binary quantizer for maximizing mutual information. *IEEE Transactions on Communications*, pages 1–1, 2020.

[81] Brian M Kurkoski and Hideki Yagi. Finding the capacity of a quantized binary-input dmc. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 686–690. IEEE, 2012.

[82] Minh N Vu, Nghi H Tran, Dissanayakage G Wijeratne, Khanh Pham, Kye-Shin Lee, and Duy HN Nguyen. Optimal signaling schemes and capacity of non-coherent rician fading channels with low-resolution output quantization. *IEEE Transactions on Wireless Communications*, 18(6):2989–3004, 2019.

[83] M. Ranjbar, N. H. Tran, M. N. Vu, T. V. Nguyen, and M. Cenk Gursoy. Capacity region and capacity-achieving signaling schemes for 1-bit adc multiple access channels in rayleigh fading. *IEEE Transactions on Wireless Communications*, 19(9):6162–6178, 2020.

[84] Eric E Majani and H Rumsey. Two results on binary-input discrete memoryless channels. In *Information Theory, 1991 (papers in summary form only received), Proceedings. 1991 IEEE International Symposium on (Cat. No. 91CH3003-1)*, pages 104–104. IEEE, 1991.

[85] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[86] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[87] 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.

[88] Suvrit Sra, Stefanie Jegelka, and Arindam Banerjee. Approximation algorithms for bregman clustering co-clustering and tensor clustering. 2008.

[89] Kamalika Chaudhuri and Andrew McGregor. Finding metric structure in information theoretic clustering. In *COLT*, volume 8, page 10. Citeseer, 2008.

[90] Andrea Vattani. K-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, 45(4):596–616, 2011.

[91] Thuan Nguyen and Thinh Nguyen. Minimizing weighted concave impurity partition under constraints. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, submitted.

[92] Thuan Nguyen and Thinh Nguyen. Communication-channel optimized partition. *In Global Communications Conference (GLOBECOM), IEEE*, 2020.

[93] 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.

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

[95] Noam Slonim and Naftali Tishby. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research*, volume 1, page 200, 2001.

[96] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *Advances in neural information processing systems*, pages 617–623, 2000.