Divergence Covering

by

Jennifer Tang

Master of Science, Massachusetts Institute of Technology (2015)
Bachelor of Science, Princeton University (2013)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author .....
Department of Electrical Engineering and Computer Science
January 26, 2022

Certified by .....
Yury Polyanskiy
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Abstract

A longstanding problem of interest is that of finding covering numbers. A very important measure between probability distributions is Kullback-Leibler (KL) divergence. Both topics have been massively studied in various contexts, and in this thesis we focus on studying the problem when the two concepts are combined. This combination yields interesting techniques for providing useful bounds on a number of important problems related to information theory. Our goal is to explore covering the probability simplex in terms of KL divergence. Various properties of KL divergence (e.g. it is not a metric, not symmetric, and can easily blow up to infinity) make it unintuitive and difficult to analyze using traditional methods. We look at covering discrete large-alphabet probabilities both with worst-case divergence distance and average-case divergence distance and examine the implications of these divergence covering numbers. One implication of worst-case divergence covering is finding how to communicate probability distributions with limited communication bandwidth. Another implication is in universal compression and universal prediction, where the divergence covering number provides upper bounds on minimax risk. A third application is computing capacity of the noisy permutation channel. We then use average-case divergence covering to study efficient algorithms for quantizing large-alphabet distributions in order to save storage space.

Dedication

To mom, dad and Alan.

Acknowledgements

First, thank you to my advisor Prof. Yury Polyanskiy for his guidance, on both technical and non-technical subjects, as well as for his brilliant energy which was the inspirational force that led me to projects in this work. Next, I want to thank Prof. Greg Wornell for both for his wisdom on research and for all that I learned from him about teaching classes. Thanks to Prof. Meir Feder and Prof. Sasha Rakhlin for the many discussions that influenced the direction of my research. (Additional thanks to Prof. Feder for working with me at Tel Aviv University.) Thanks to all the professors, staff, and admins in LIDS for their positive interactions, among these are Prof. Guy Bresler, Dr. Mardavij Roozbehani, Prof. Caroline Uhler, Prof. Stefanie Jegelka, Prof. Asu Ozdaglar and Prof. John Tsitsiklis. Special thanks to Prof. Devavrat Shah for getting me involved with data science. Thank you to the Department of Electrical Engineering and Computer Science, and particularly Prof. Polina Golland, Prof. Lizhong Zheng and Prof. Al Oppenheim. Special thanks to Prof. Anthony Philippakis for his expertise on working with biological data.

I would like to thank many of my wonderful colleagues and friends for being part of my journey throughout the years. To those in LIDS: Omer Tanovic, Chenyang Yuan, Suhas Kowshik, Austin Collins, Hajir Roozbehani, Or Ordentlich, Ziv Goldfeld, Shreya Saxena, Qingqing Huang, Christina Lee, Quan Li, Hamza Fawzi, Igor Spasojevic, Jason Altschuler, and the excellent Flora Meng. A special thanks to Anuran Makur for inspiring new projects and Qian Yu for his proof ideas. To my most recent collaborators: Gary Lee, Amir Weiss, Yuheng Bu, and Alejandro Lancho. To those in the Theory of Computing group who adopted me into their circle: Daniel Grier, Luke Schaeffer, Siddhartha Jayanti, Dylan McKay, Gautam Kamath, Govind Ramnarayan and many many others. And to these amazing people: Amy Ousterhout, Danielle Pace, Mandy Korpusik, Mengfei Wu, Ramya Ramakrishnan, and Guha Balakrishnan.

Finally, I would like to thank my family. Thanks to my mother and father for supporting me in pursuing my goals and always being there when I need them. An additional thanks to my mother, father, and my uncle for housing and feeding me through the pandemic as I worked. Thanks to my brother for always being available to discuss any topic with me. Most of all, I would like to thank Aviv Adler for not only being a collaborator on some of the work in this thesis but also for being my greatest companion through all the ups and downs.

Contents

Chapter 1

Introduction

Consider the following questions:

1. You are gambling with the devil. The devil picks a secret probability distribution on K letters. He reveals a letter selected randomly and independently from his distribution one at a time. You must, before each letter is revealed, predict the likelihood of what that letter will be. You will be penalized with log-loss. The devil's secret probability does not change, so you can learn from the past letters to predict the current one. What is your strategy for prediction when the number of letters K is as large as the sequence you are trying to predict?

2. You send data over the network but, due to a quirk, your data symbols arrive in a random order. Usually, when you encode your message, your data symbols appear in a string and the order of the symbols matters for decoding. Now the receiver has an unordered set of symbols. How much information are you able to communicate with this scheme?

3. You have determined the frequency of words in various books by different authors. This way, the next time an author writes a book, you are ready to use your Huffman encoder or arithmetic encoder to compress it. However, storing all the different frequencies for each author takes too much space on your limited hard drive. You would like to compress these frequencies to a small number of bits, while still keeping them useful for compression tasks. What is the best way to compress these frequencies?

While these are three seemingly different questions about different topics, it turns out that they are united by a common theme: they are all applications of divergence covering. Specifically, we mean a Kullback-Leibler (or KL) divergence. The goal of this thesis is to explore divergence covering and its consequences, which include finding a solution to each of the three questions.

We begin with a gentle introduction first discussing the idea of covering and then divergence. A *covering number* for a space A, distance metric $D(x, y)$, and distance ε is defined as the size of the smallest set of points $\mathcal{C} \subset A$ so that all points in A are within a distance of ε from a point in $\mathcal{C}$. Defining the covering number as $M(A, D, \varepsilon)$ formally, we have

$$M(A, D, \varepsilon) = \inf\{|\mathcal{C}| : \min_{y \in \mathcal{C}} D(x, y) \leq \varepsilon, \forall x \in A\} \tag{1.1}$$

Typically, covering numbers are studied when D is a norm, and most commonly the L_2 or Euclidean norm is considered. For the Euclidean norm, while there are many nuances to the problem of finding the covering number that still require more research, a first order approximation can usually be given by using a volume argument: take the volume of the space of A and divide that by the volume of a Euclidean ball with radius ε. Special adjustments might have to be made to ensure the method gets proper upper and lower bounds on the covering number, but in general this volume argument works when D is a metric on a finite-dimensional space.

But what if the notion of distance we are interested in is not a norm or even a metric. This is what occurs when the distance used for the covering number problem is the KL divergence. As its name suggests,

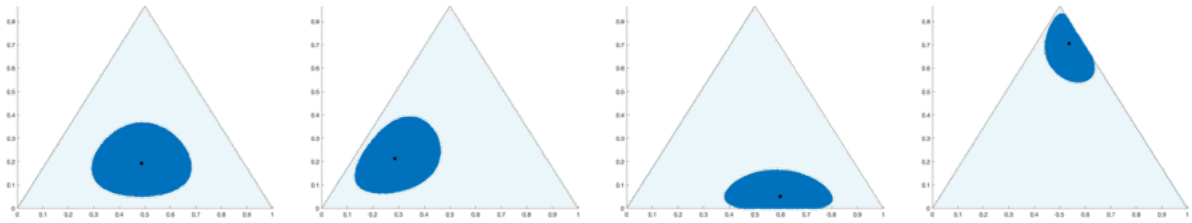


Figure 1.1: Example of divergence balls centered at different points in $\triangle_2$ with the same radius.

KL divergence is a divergence and not a distance metric in the technical sense, though colloquially, we say it measures the "distance" between two probability distributions. Studying this as the distance for covering implies that the space A we are considering must be the space of probabilities. We will be focusing on the case when A is the $(K-1)$-dimensional simplex representing the space of discrete probabilities on K outcomes.

1.0.1 KL Divergence Basics

For probability distributions P and Q over a discrete alphabet $\mathcal{A}$, the *KL divergence* is defined as

$$D(P\|Q) = \sum_{a\in\mathcal{A}} P(a)\log\frac{P(a)}{Q(a)} \tag{1.2}$$

Unlike a metric, the KL divergence is not symmetric and does not obey the triangle inequality. We thus need to specify that we want probability distribution P, the first argument in (1.2), to be covered, and that Q, the second argument in (1.2) is doing the covering. We call the set of Q used to cover the space the set of *centers*.

Like a metric, $D(P\|Q) \geq 0$ with equality if and only if $P = Q$. We can define a *divergence ball* of radius ε around Q to be all P in the simplex so that $D(P\|Q) \leq \varepsilon$. Unlike Euclidean balls, KL divergence balls cover a different volume of points depending on where Q is. For an illustration of this, see Figure 1.1. The ball around Q covers more volume when Q is closer to the center of simplex. In fact, if there is an a such that $Q(a) = 0$, then the ball of radius ε around Q can only contain P where $P(a) = 0$. This unevenness in ball volume makes using the volume argument (discussed above) to compute covering numbers not straightforward (we will be using this argument in places, but with appropriate considerations that need to be discussed). In addition to covering numbers, we are interested in examining where to best place these KL divergence balls. Figure 1.2 gives a picture of randomly selected centers and which points are closest to which center.

1.0.2 Definitions

We will use the notation $\triangle_{K-1}$ to mean a $(K-1)$-dimensional simplex (and thus the alphabet size here is K).

Definition 1 (Simplex)**.**

$$\triangle_{K-1} = \left\{\boldsymbol{x} \in \mathbb{R}^K, x_i \geq 0 : \sum_{i=1}^{K} x_i = 1\right\} \tag{1.3}$$

One of our main focuses is to explore how many covering centers are needed to cover the probability simplex. There are two notions we can define of what it means to be covered, *worst-case* and *average-case*. Worst-case covering means that the distance of any $P \in \triangle_{K-1}$ to a covering center is smaller than some value ε. This covering number for worst-case divergence covering is given by the following:

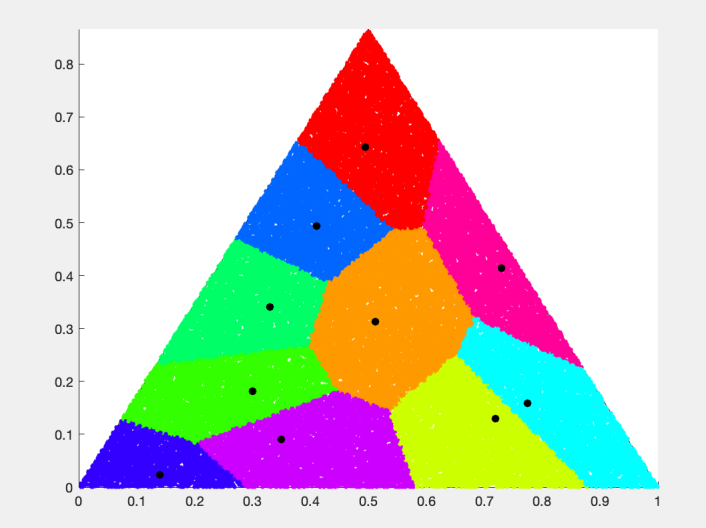


Figure 1.2: Illustration of a 2-dimensional simplex and which points on the simplex are closest to which centers in KL divergence. Notice that the regions have very different boundaries than what would have occurred with Euclidean distance (but the boundaries are still straight lines).

Definition 2 (Worst-Case Divergence Covering of Simplex)**.** *Let K be the alphabet size and $\varepsilon > 0$.*

$$M(K,\varepsilon) = \inf\{m : \exists\{Q_{(1)},...,Q_{(m)}\}\ s.t \max_{P\in\triangle_{K-1}} \min_{Q_{(i)}} D(P||Q_{(i)}) \leq \varepsilon\} \tag{1.4}$$

We will call ε the *radius*. The set of centers is denoted as $\mathcal{Q} = \{Q_{(1)},...,Q_{(m)}\}$. We will also refer to $M(K,\varepsilon)$ as the *covering number*.

For average-case divergence covering, we instead consider how many covering centers are needed so that the expected KL divergence between points P and one of the centers is less than ε, under some prior W over the probability simplex.

Definition 3 (Average-Case)**.** *Given a distribution W over the simplex $\triangle_{K-1}$,*

$$M^*(W,\varepsilon) = \inf\{|\mathcal{Q}| : \mathbb{E}_{\pi\sim W}[\min_{Q\in\mathcal{Q}} D(\pi\|Q)] \leq \varepsilon\}\,. \tag{1.5}$$

The number of centers needed in the average-case is a lower bound for worst-case divergence covering. We will use the term *divergence covering* to mean worst-case divergence covering and only add "worst-case" to clarify the difference from average-case.

1.1 Worst-Case KL Divergence Covering

1.1.1 Divergence Covering Results

We show in this work that the divergence covering number can be bounded by

$$\left(c\frac{1}{\varepsilon}\right)^{\frac{K-1}{2}} \leq M(K,\varepsilon) \leq \left(C\frac{\log K}{\varepsilon}\right)^{\frac{K-1}{2}} \tag{1.6}$$

for positive constants c and C. This is our most general bound which applies for any $\varepsilon < \log K$ (note that all points in $\triangle_{K-1}$ can be covered with one point using radius $\log K$). The exponent of $(K-1)/2$ implies that covering the simplex with divergence balls of radius ε is similar to covering a space with Euclidean balls of radius $\sqrt{\varepsilon}$.

Our bound in (1.6) is not best bound for all possible ε and K. For a region of small ε, we can remove the $\log K$ term (though a term with ε shows up). For very large ε, say example when $\varepsilon = \frac{1}{2}\log K$, we can find *subexponential* bounds where the exponent of $1/\varepsilon$ in the divergence covering number is less than linear in K. Our divergence covering bounds will be explored in Chapter 2 and the subexponential ones will be explored in Chapter 3.

In the next sections, we introduce some of the applications of the divergence covering number. We will start by discussing our initial motivation (the Cloud Communication Problem) and then move to other applications, including the three problems mentioned earlier.

1.1.2 Application to Cloud Communication Problem

Suppose that a company has a centralized computing and data storage component, for example a cloud, and also has many subsidiaries around the world which have limited computing and storage abilities.

Overall, the task of the centralized cloud and all the subsidiaries is to compress their users' information, so that this information can be stored or transmitted efficiently. To use a compression algorithm, such as Huffman or arithmetic coding, each subsidiary would need to know the probability distribution of the symbols the users are using.

The centralized cloud has access to all the data and can therefore determine the statistics of the users' data. It can, on any given day, model the current trends of the data and determine the best probability distribution to use to compress the users' data. Each subsidiary on the other hand has no way of determining which probability distribution to use. They rely on communication from the centralized cloud for this information. But to save on cost, we want to limit the communication bandwidth between the centralized cloud and all the subsidiaries. The task is then to communicate a probability distribution to the subsidiaries using a limited number of bits.

Limiting the number of bits has a trade-off with how precisely a probability distribution can be described. If we limit the number of bits to b, then we can communicate at most 2^b different probability distributions. Let $\mathcal{Q}$ be this set of 2^b probabilities chosen as candidates to describe. Then if the centralized cloud discovers that P is the distribution that best describes the data to lower expected compression length, we know from [1] that the best choice of $Q^* \in \mathcal{Q}$ to send to the subsidiaries is

$$Q^* = \arg\min_{Q\in\mathcal{Q}} D(P\|Q)\,. \tag{1.7}$$

Ideally $\mathcal{Q}$ is chosen so that each P is not far from some $Q \in \mathcal{Q}$ in KL divergence distance. An equivalent way of looking at the problem is to fix the maximum distance $D(P\|Q)$ to be less than ε, and to determine how many bits our communication system has to use. The number of bits is given by $\log_2|\mathcal{Q}|$, and $|\mathcal{Q}|$ is the covering number for KL divergence.

Using our results we can show the following as an example: Suppose that the alphabet size is $K = 30,000$, which is a typical vocabulary sized used for modeling English words, and we want $D(P\|Q)$ to be less than 1 bit (so $\varepsilon = 1$). Using (1.6), get that

$$\log_2 M(30000, 1) \approx 173000 \tag{1.8}$$

then at most $173,000$ bits of communication are needed from the centralized cloud to the subsidiary. This means we can send Q, the approximation to P, using less than $173000/30000 \approx 5.7$ bits for each word in the vocabulary. Increasing or decreasing the bound on $D(P\|Q)$ leads to different tradeoffs in numbers of bits, characterized by our bounds.

1.1.3 Application to Universal Compression and Universal Prediction

An important question which has been studied in depth is that of finding minimax redundancy and minimax regret for iid discrete alphabets sources. We will also look into this problem by showing how divergence covering can be used to find upper bounds on both quantities. The goal is to develop the framework which can be used to obtain non-asymptotic results for any alphabet size and sequence length. The upper bounds we compute from our results are tight enough to get all the first order terms of known regret bounds.

In the same chapter, using similar techniques, we show how divergence covering can be used for finding regret for the class of patterns. Our resulting regret is (for length n sequences)

$$cn^{1/3}\log^{4/3}(n) \tag{1.9}$$

improving the exponent on the logarithmic factor compared to known results [2].

1.1.4 Application to Permutation Channels

Another application of our divergence covering results is that we can use it to analyze the noisy permutation channel. The noisy permutation channel is used to represent communication channels where symbols might not arrive in the same order they were sent (as in our second question). It can be be used to model applications like multipath communication and biological storage systems.

The structure of the noisy permutation channel can be represented by the Markov chain

$$X^n \to Z^n \to Y^n \tag{1.10}$$

where each X^n, Y^n, Z^n is a length n sequence. The transformation between X^n to Z^n is a discrete memoryless channel given by the transition probabilities $P_{Z|X}$. The sequence Y^n is a uniformly random permutation of Z^n. The goal is to determine the capacity of the noisy permutation channel for different transition probabilities $P_{Z|X}$.

While it might not seem obvious at first, it turns out that KL divergence is exactly the correct distance to study the noisy permutation channel with. Covering with KL divergence allows us to get converse bounds (or upper bounds on the capacity). We will be able to resolve the conjecture in [3] and show that as a function of the transition probabilities $P_{Z|X}$, the capacity of the permutation channel is

$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{\mathsf{rank}(P_{Z|X}) - 1}{2}\,. \tag{1.11}$$

where $\mathsf{rank}(\cdot)$ gives the rank of the matrix. We dedicate Chapter 5 to exploring the noisy permutation channel in depth.

1.2 Average-Case Divergence Covering

Worst-case divergence covering is the correct criteria to use when it is necessary to make sure every possible probability distribution P in $\triangle_{K-1}$ is close to a center. However, when working with a collection of P drawn randomly, it is sometimes sufficient that P is just on average close to one of the centers. Certain distributions P might be very rare and it may not be necessary to dedicate so many centers to covering rare distributions.

Our average-case divergence covering problem can be restated as a quantization problem where the loss function is the KL divergence. The goal is to find centers, or reconstruction points, so that the expected loss is less than ε. Thus, an application of average-case divergence covering is our third question about lossy compression of probability distributions.

We will focus on solutions where the covering (or quantization) centers can be determined by looking at each coordinate of the probability distribution separately. This assumption greatly simplifies the quantization steps and makes the solution practical for implementation.

1.2.1 Rate Distortion Single-Letterization

Since our average-case divergence covering problem is also a quantization problem, it is natural to look at the problem through the lens of rate distortion (which we can do because we are looking at each coordinate separately). As we mentioned above, for average-case divergence covering we need a prior W over the probability simplex. We will be particularly interested in the case when the prior W is a symmetric Dirichlet distribution. This class of priors covers the uniform prior and the Jeffreys' prior (which is known to be asymptotically worst-case for certain problems). We develop our rate distortion problem for average-case

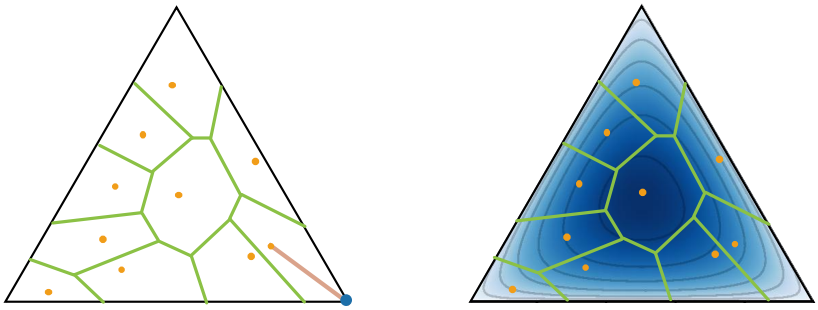


Figure 1.3: Illustration of the concept of worst-case covering (left) and average-case covering (right). For a set of centers, the radius in worst-case covering is determined by the largest KL divergence distance from any point to its nearest center (we illustrate this distance with a red line). In contrast, for average-case covering, the radius is determined by the average KL distance from a point to its nearest center. This average is computed with respect to prior W (the prior is illustrated by the shade of blue).

divergence covering in Chapter 6 where we determine upper that if our prior W is a symmetric Dirichlet distribution with parameter α, then approximately

$$\log M^*(W, \varepsilon) \approx \frac{K}{2} \log \frac{c(\alpha)}{\varepsilon} . \tag{1.12}$$

Our results show that the number of quantization points needed *per symbol* does not increase with K. An implementable method based on our proof for the case when W is the uniform prior (this is the same as a symmetric Dirichlet with parameter $\alpha = 1$) is called the EDI (exponential density intervals) method. It's performance is illustrated in Figure 1.4. We show that if we fix the number of quantization points per symbol, then the KL divergence loss of the EDI method is nearly constant with K, agreeing with (1.12).

1.2.2 Efficient Quantization via Companding

While showing some of our theoretical results for average-case divergence covering with Dirichlet priors, we discovered that perhaps very simple coordinate-based quantization schemes exist for our average KL divergence loss. We describe these schemes as *companders*, based on companding, a type of transform coding used in signal processing. We discuss what our companding scheme is in Chapter 7.

In our study of companders for average-case KL divergence, we discovered that there exists a minimax compander which is the best compander to use against the worst-case prior W chosen on the probability simplex. Our analysis gives asymptotic guarantees on the performance of the minimax compander and other companders. Impressively, our experimental results with our companders give performances very close to what is expected theoretically, despite the fact that the parameters are far from the asymptotic regime.

Typically, if each value in a probability distribution needs to be stored only using 8-bits per value, the most direct solution is to quantize the value using uniform (equally-spaced) bins. Our experiments on both randomly generated data and real data show that this solution (which we call truncation) achieves a KL divergence on the order of 10^{-1}. However, using our minimax companders, an operation with an insignificant amount of overhead, we can reduce the KL divergence between the true distribution and stored distribution to 10^{-4}, a massive reduction in loss. These results are in Figure 1.4. These experiments show that indeed well-defined companders are excellent choices for quantization with respect to average KL divergence. They give a practical and efficient solution to our third problem.

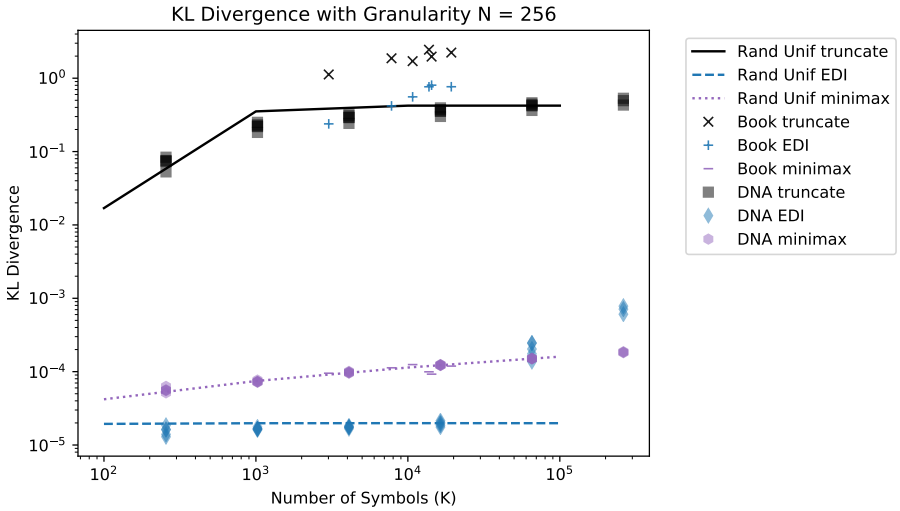


Figure 1.4: This plot shows different methods used to quantize probability distributions with respect to KL divergence. The number of quantization levels (the granularity) is fixed to $N = 256$ per symbol in the alphabet size of K. This plot shows two things. First, the EDI method with uniformly generated probabilities is constant even as K grows. Second, our method, the minimax compander, is able to get a few orders of magnitude improvement compared to truncation (the default method). Though it is not as good as the EDI on randomly drawn probabilities, the minimax compander is better on real data compared to the EDI.

1.3 Other Results: Divergence Covering of Subsets of the Simplex

In addition to finding divergence covering of the whole simplex (as in Definition 2), we also explore divergence coverings of certain subsets of the simplex. The formal definition for this is as one would expect:

Definition 4. *For a subset* $\mathcal{B} \subset \Delta_{K-1}$

$$M(K, \varepsilon, \mathcal{B}) = \inf\{m : \exists\{Q_{(1)}, ..., Q_{(m)}\} \ s.t \ \max_{P \in \mathcal{B}} \min_{Q_{(i)}} D(P||Q_{(i)}) \leq \varepsilon\}. \tag{1.13}$$

Some of the key results we derive in this work relies on covering a subset of the simplex. One occurs in the problem of finding regret on the class of patterns. Here, the subset $\mathcal{B}$ is the set of probabilities in $\triangle_{K-1}$ which are in sorted order. This means that if probability P is such that symbol $i \in [K]$ occurs with probability p_i, then if $i < j$, we have $p_i \geq p_j$. For this subset (with another condition that non-zero p_i must not be too small) we can show (ignoring logarithmic factors) that $\log M(K, n^{-2/3}, \mathcal{B}) \leq cn^{1/3}$. Finding the covering number is essential to determining Equation (1.9).

Another instance subset covering appears is in the application to the noisy permutation channel where we need to cover a lower dimensional linear subspace of the simplex. In such a case, if the linear subspace has dimension ℓ, then the divergence covering of the subspace, $M(K, \varepsilon, \mathcal{B})$ can be bounded by $\binom{K}{\ell} M(\ell, \varepsilon)$. We discuss this in Chapter 5 while discussing the noisy permutation channel.

Chapter 2

Divergence Covering

2.1 Introduction

In this chapter, our primary task is to determine upper and lower bounds on the (worst-case) *divergence covering number* $M(K, \varepsilon)$ defined in Definition 2. We will primarily focus on the case when the covering radius ε is small, meaning (roughly) constant in size even as the dimension of the simplex we are covering grows. (Chapter 3 will focus on the case when ε is large.)

A secondary goal is to examine how the divergence covering centers $\mathcal{Q}$ should be placed throughout the simplex in order to minimize the number of centers used. For this purpose, we include some analysis which give suboptimal bounds for $M(K, \varepsilon)$. The first of these is an analysis of when the divergence covering centers are on a uniform grid. This placement of centers is likely the default, and therefore it is informative to understand its performance. The second suboptimal choice of centers we give is one similar to placing centers on a grid, but the grid is unevenly spaced throughout the simplex.

The choice of centers which gives the upper bound that is closest to the known lower bound is one where the centers are chosen randomly. We also show two different techniques for choosing these random centers. All of these are summarized below in Section 2.1, after a discussion of the applications and previous works.

Application To reiterate from the introduction, we have two main applications of divergence covering numbers. The first is the cloud communication problem: to find the trade-off between the excess coding length for communications between subsidiaries and bits needed for communication from the cloud. The second important application of divergence covering numbers (the upper bounds in particular) is that they can be used to find upper bounds on redundancy and regret for a class of iid distributions. This is discussed in Chapter 4.

Our results on redundancy and regret also tie in with the fact that worst-case divergence covering can be used as a 2-part universal code. For any length n sequence which needs to be compressed, we can determine the empirical distribution P of the sequence. For each P, we can determine the center $Q \in \mathcal{Q}$ so that $D(P\|Q) \leq \varepsilon$. The first part of the code tells us which Q to use. The second part encodes the sequence using Q.

The expected coding length (or rather the idealized code length which ignores constants occurring in practical codes) for this scheme is given by

$$\mathbb{E}[L] = \log M(K, \varepsilon) + n \min_{Q \in \mathcal{Q}} D(P\|Q) + nH(P) \tag{2.1}$$

This 2-part code is an achievable code for universal compression. It is similar to 2-part codes used in the minimum description length (MDL) literature.

Previous Works If we take the logarithm of the definition for covering number in (1.1) of Chapter 1, we get the Kolmogorov metric entropy. Kolmogorov introduced the notion of ε-entropy for function classes which are discussed in [4] and referenced by [5]. Let T be a domain and let class $\mathcal{F}$ be a set of functions $f(t)$

with $t \in T$, where $\mathcal{F}$ is compact for norm $D(f_1, f_2)$. Then an ε-net is a system $\mathcal{N}(\varepsilon)$ such that

$$\sup_{f \in \mathcal{F}} \min_{f' \in \mathcal{N}(\varepsilon)} D(f, f') \leq \varepsilon \,. \tag{2.2}$$

If $M(\mathcal{F}, D, \varepsilon)$ is the minimal size of the ε-net, then the Kolmogorov ε-entropy is

$$H_\epsilon(\mathcal{F}, D) = \log_2 M(\mathcal{F}, D, \varepsilon) \,. \tag{2.3}$$

Even though KL divergence is not a norm or metric, authors will still refer to logarithm of the covering number under KL divergence as a metric entropy. Yang and Barron[1] [6], use metric entropies to determine the minimax rates of convergence for estimating parameters of densities. While they discuss implications of finding the metric entropies and give examples for certain cases, they do not determine what the metric entropy actually is for the case of discrete probability distributions. Our divergence covering numbers will determine this metric entropy. We will be using a key technique of Yang-Barron throughout this work.

Other than the line of work from [6], we did not find works which discuss covering with KL divergence. However, there are some works which consider very similar problems.

In [7], the authors examined the problem of vector quantization for universal coding. They analyze a two-stage universal noiseless code and determine an iterative (Lloyd-like) algorithm to construct the codebook for this code. If only the first stage of their noiseless coding is used, then their construction becomes a universal lossy code. Their work applies to loss functions in general, and KL divergence is considered as a special case. The quantization points which result from the first stage code using KL divergence is analogous to our centers for covering (though this connection is more relevant to average-case divergence covering). They applied their algorithm to find quantization points for the probability simplex. However, critically, their algorithm and results do not apply when the boundary points of the simplex are included; their results only hold for some interior of the simplex. This scenario is unsatisfying for the purposes of finding centers for a divergence covering for high-dimensional spaces, where the boundary will take up a larger proportion of the space. Also, for the applications, probability distributions where a symbol occurs with probability zero are very likely. Neglecting them drastically limits the utility of the divergence covering.

The problem of ignoring the boundary of the simplex similarly comes up in [8], where the author considers an algorithm for quantizing probabilities which primarily uses lattices. Different distortion measures for quantizing are considered including KL divergence. (The author considers both average-case and worst-case quantization.) For quantizing in terms of KL, the author's main technique is to use chi-squared divergence to upper bound KL divergence. This however, still excludes the boundary points, since the equation for computing the chi-squared divergence explodes to infinity for points on the boundary.

The boundary is precisely where the difficulty lies for finding a divergence covering. Points which are close in Euclidean distance near the boundary can be very far, even infinite, in terms of KL distance. A key contribution of this work is find methods which can take care of these effects at the boundary.

Summary of Main Results First, we give the suboptimal results for when the centers are on a uniform grid. The centers points here are evenly spaced throughout the simplex. The bounds for the divergence covering number achievable with these centers is

$$\left(c\frac{1}{\varepsilon}\right)^{K-1} \leq M_{\text{evenly spaced}}(K, \varepsilon) \leq \left(C\frac{K}{\varepsilon}\right)^{K-1} \tag{2.4}$$

where[2] $0 \leq \varepsilon \leq \log K$. This bound shows that if we want the radius of our covering balls to be ε, using centers that are evenly spaced will require on the order of $(1/\varepsilon)^{K-1}$ total number of centers. We have mentioned already that this placement of centers is suboptimal. One reason this is apparent is that points near the center are very close in KL distance to a center, whereas points near the boundary are very far from a center point. An equivalent way of thinking about this is that points near the center of simplex are covered by multiple divergence balls. We can remove many of the centers near the center and still maintain the same radius ε.

[1] We will be using a key technique of Yang-Barron throughout this work.

[2] Note that the radius ε can at most be $\log K$. This is the radius achievable with one center in the exact middle of the simplex.

Next, we give a lower bound for the divergence covering number (for any placement of centers). For any $0 \leq \varepsilon \leq \log K$,

$$M(K, \varepsilon) \geq \left(\frac{1}{8\varepsilon}\right)^{(K-1)/2}. \tag{2.5}$$

When we compare this lower bound to (2.4), we see there is a very large gap in the exponent for the covering number. Centers on a uniform grid give an exponent of $K-1$ which is twice as large. It turns out that the smaller $(K-1)/2$ exponent is the correct one.

Our first result which gets the correct exponent is one which uses a grid-like structure. It is not the tightest bound, but it gives an easy and explicit placement of centers. The basic insight in this bound is that more centers need on the edges of the simplex. We call this the explicit bound: For $0 < \varepsilon < 1$,

$$M(K, \varepsilon) \leq c^{K-1} \left(\frac{K-1}{\varepsilon}\right)^{\frac{K-1}{2}} \tag{2.6}$$

for some constant c. While the exponent in this bound is tight, the extra factor of $k-1$ which is raised to the exponent form a gap from the lower bound (2.5).

To tighten this gap, we have our next upper bound on the divergence covering number. This is one of our tightest upper bounds. We call the technique we use for computing the bound the "using-Hellinger" technique, since the key is to use Hellinger divergence as an upper bound on KL divergence (we discuss divergences in general in Section 2.2.1). The technique shows that some set of centers must exist, but does not explicitly define where the centers are. The using-Hellinger upper bound is: For $\varepsilon \leq \log K$,

$$M(K, \varepsilon) \leq K \left(C \frac{\log K}{\varepsilon}\right)^{\frac{K-1}{2}} \tag{2.7}$$

for some constant C. Compared to (2.6), the using-Hellinger upper bound reduces the $K-1$ raised to the exponent to a factor of $\log K$ raised to the exponent.

KL divergence can also be bounded above using chi-squared divergence. Our last upper bound using chi-squared balls to upper bound the divergence covering number. Like the using-Hellinger upper bound, this upper bound only gives existence of points. The goal of this bound is to remove $\log K$ raised to the exponent factor entirely. This upper bound states: For $\varepsilon \leq \frac{1}{4(K+1)^2}$,

$$M(K, \varepsilon) \leq c_0 K^{3/2} \left(\frac{c_1}{\varepsilon}\right)^{\frac{K-1}{2}} \log \frac{1}{\varepsilon} \tag{2.8}$$

for constants c_0 and c_1. While this bounds removes the $\log K$ raised to the exponent, it is not shown to work for ε too large. The $\log(1/\varepsilon)$ also makes the bound worse than (2.7) when ε is too small. However, we believe that the using chi-squared bound is the "correct" approach to use to find the divergence covering number. The method for producing the centers is possibly tight, but we think it is our analysis that has limitations.

Here we give an overview of all the divergence covering number results. Other bounds, which are the subexponential and polynomial bounds, are the topics of the next chapter and are not presented in these tables.

Divergence Covering Lower Bounds	
Restrictions	Lower Bound
Evenly Spaced	$(\frac{c}{\varepsilon})^{K-1}$
None	$(\frac{c}{\varepsilon})^{\frac{K-1}{2}}$

Figure 2.1: Table comparing lower bounds for divergence covering for alphabet size K. Value of constant c will vary for each bound.

Divergence Covering Upper Bounds		
Technique	Restrictions	Upper Bound
Explicit	Evenly spaced	$(C\frac{K}{\varepsilon})^{K-1}$
	$0 < \varepsilon < 1$	$(C\frac{K}{\varepsilon})^{\frac{K-1}{2}}$
Using-Hellinger	None	$(C\frac{\log K}{\varepsilon})^{\frac{K-1}{2}}$
Using-chi-squared	$0 < \varepsilon < \frac{1}{4(K-1)^2}$	$(\log\frac{1}{\varepsilon})(\frac{C}{\varepsilon})^{\frac{K-1}{2}}$

Figure 2.2: Table comparing upper bounds for divergence covering for alphabet size K. Value of constant C will vary for each bound.

Chapter Organization First we start with some preliminaries in Section 2.2 which introduces some definitions and theorems we will need for our proofs. Our first result on evenly spaced centers is given in Section 2.3. We then discuss the lower bound on the number of centers in Section 2.4. This is followed by our upper bounds. The explicit bound is given in Section 2.5. The using-Hellinger upper bound is presented in Section 2.6. And lastly, the using-chi-squared upper bound is discussed in Section 2.7.

Notational Notes For probability distributions, in general, we can use capital letters like P or Q for either discrete or continuous distributions. If P refers to a discrete distribution on one variable, then we can write the probability of x under P as $P(x)$.

When we are working exclusively with discrete distributions in $\triangle_{K-1}$ for some K (probability simplex with alphabet size K), it is more convenient at times to use lower case letters such as p or q. Unless otherwise specified, we assume the symbols (or labels) when the alphabet is of size K is given by $1, 2, \ldots, K$. To simplify notation for this, for any integer n, we let $[n] = \{1, 2, \ldots, n\}$.

For each $i \in [K]$, we let p_i be the probability discrete distribution p has on symbol i. Thus, we can write for $p \in \triangle_{K-1}$,

$$p = (p_1, p_2, \ldots, p_K)\,. \tag{2.9}$$

When working with vectors (not necessarily probability vectors) we also bold the letters to emphasize that quantity is a vector. The ith entry of vector $\boldsymbol{a}$ is written as a_i, so similarly to above,

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_K)\,. \tag{2.10}$$

We may also bold probability vectors, but will not if the notation is clear. For continuous distributions, we also typically use lower case letters like p or f to denote the probability density function (pdf), but it should be clear from context that p is a pdf and not a discrete probability.

2.2 Set of Tools

This sections includes some background which will apply to some of the ideas in this chapter (and possibly to other chapters).

2.2.1 KL Divergence and Other Divergences

KL divergences are a type of f-divergences. Other f-divergences are Hellinger and χ^2 divergence, both of which will come up in our upper bounds in this chapter. We give some exposition on f-divergences here, starting with a definition.

Definition 5 (f-Divergence)**.** *For a convex function $f : \mathbb{R} \to \mathbb{R}$ with $f(1) = 0$, the f-divergence is*

$$D_f(P||Q) = \mathbb{E}_{X\sim Q}\left[f\left(\frac{P(X)}{Q(X)}\right)\right] \tag{2.11}$$

for distributions P, Q.

Next, we give some examples and their formulas on discrete probabilities.

Definition 6 (Various f divergences)**.** *Some important examples:*

- *KL Divergence:*

$$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{2.12}$$

- *χ^2 Divergence:*

$$D\chi^2(p||q) = \sum_i \frac{(p_i - q_i)^2}{q_i} \tag{2.13}$$

- *Total Variation (TV)*

$$TV(p, q) = \frac{1}{2}\sum_i |p_i - q_i| \tag{2.14}$$

For probabilities on the set A,

$$TV(P, Q) = \sup_{B\subset A} |P(B) - Q(B)| \tag{2.15}$$

- *Squared Hellinger*

$$D_{H^2}(p, q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2 \tag{2.16}$$

The most important inequality we will use for finding upper bounds on divergence covering the inequality which upper bounds KL divergence with χ^2 divergence. The following is from [9].

Fact 1. *For discrete probabilities p and q, we have*

$$D(p||q) \leq \sum_i \frac{(p_i - q_i)^2}{q_i} = D\chi^2(p||q) \tag{2.17}$$

Proof. We will use that $\log x \leq x - 1$.

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{2.18}$$

$$\leq \sum_i p_i \left(\frac{p_i}{q_i} - 1 \right) \tag{2.19}$$

$$= \sum_i \frac{p_i^2}{q_i} - p_i \tag{2.20}$$

$$= \sum_i \frac{p_i^2}{q_i} - p_i - p_i + q_i \tag{2.21}$$

$$= \sum_i \frac{p_i^2}{q_i} - \frac{2p_i q_i}{P_2(a)} + \frac{q_i^2}{q_i} \tag{2.22}$$

$$= \sum_i \frac{(p_i - q_i)^2}{q_i} \tag{2.23}$$

□

There is unfortunately no lower bound on KL divergence using χ^2 divergence. However, locally, any f-divergence will look like the χ^2 divergence.

Fact 2. *Locally, f-divergences look like χ^2-divergences,*

$$D_f(\epsilon P + (1-\epsilon)Q||Q) = \epsilon^2 f''(1) D_{\chi^2}(P||Q) + o(\epsilon^2) \tag{2.24}$$

Proof. We can use from Taylor expansions that

$$f(t+1) = f(1) + f'(1)t + \frac{1}{2}f''(1)t^2 + o(t^2)\,. \tag{2.25}$$

For divergences, $f(1) = 0$.

$$D_f(\epsilon P + (1-\epsilon)Q||Q) = \mathbb{E}_Q \left[f\left(\frac{\epsilon P(X) + (1-\epsilon)Q(X)}{Q(X)} \right) \right] \tag{2.26}$$

$$= \mathbb{E}_Q \left[f\left(\frac{\epsilon(P(X) - Q(X))}{Q(X)} + 1 \right) \right] \tag{2.27}$$

$$= \mathbb{E}_Q \left[f'(1) \frac{\epsilon(P(X) - Q(X))}{Q(X)} + \frac{1}{2} f''(1) \left(\frac{\epsilon(P(X) - Q(X))}{Q(X)} \right)^2 + o(\epsilon^2) \right] \tag{2.28}$$

$$= \frac{1}{2}\epsilon^2 f''(1) \mathbb{E}_Q \left[\left(\frac{P(X)}{Q(X)} - 1 \right)^2 \right] + o(\epsilon^2) \tag{2.29}$$

$$= \frac{1}{2}\epsilon^2 f''(1) D_{\chi^2}(P||Q) + o(\epsilon^2) \tag{2.30}$$

□

The following bound relates KL divergence to total variation. This will be used in the proof of our lower bound for divergence covering numbers.

Fact 3 (Pinsker's Inequality)**.**

$$2TV(P,Q)^2 \leq D_{\text{KL}}(P\|Q) \tag{2.31}$$

The next bound relates Hellinger to KL divergence.

Fact 4 (Hellinger Bound)**.**

$$\sum_{i=1}^{K} p_i \log \frac{p_i}{q_i} \leq \sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2 \frac{1}{(\sqrt{e}-1)^2} \max\left\{1, \log \frac{p_i}{q_i}\right\} \tag{2.32}$$

which implies

$$D_{\mathrm{KL}}(p\|q) \leq D_{H^2}(p\|q) \frac{1}{(\sqrt{e}-1)^2} \max\left\{1, \max_i \log \frac{p_i}{q_i}\right\}. \tag{2.33}$$

Proof. Set $c = \frac{1}{(\sqrt{e}-1)^2}$. To show this, the equivalences are:

$$\sum_{i=1}^{K} p_i \log \frac{p_i}{q_i} \leq c \sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2 \max\left\{1, \log \frac{p_i}{q_i}\right\} \tag{2.34}$$

$$\sum_{i=1}^{K} p_i \log \frac{p_i}{q_i} - p_i + q_i \leq c \sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2 \max\left\{1, \log \frac{p_i}{q_i}\right\} \tag{2.35}$$

$$\sum_{i=1}^{K} \frac{p_i}{q_i} \log \frac{p_i}{q_i} - \frac{p_i}{q_i} + 1 \leq c \sum_{i=1}^{K} \left(\sqrt{\frac{p_i}{q_i}} - 1\right)^2 \max\left\{1, \log \frac{p_i}{q_i}\right\} \tag{2.36}$$

$$\sum_{i=1}^{K} r^2 \log r^2 - r^2 + 1 \leq c \sum_{i=1}^{K} (r-1)^2 \max\{1, 2\log r\} \tag{2.37}$$

Hence we can show the inequality if

$$r^2 \log r^2 - r^2 + 1 \leq c(r-1)^2 \max\{1, 2\log r\} \tag{2.38}$$

$$\frac{r^2 \log r^2 - r^2 + 1}{(r-1)^2} \leq c \max\{1, 2\log r\} \tag{2.39}$$

When $r = \sqrt{e}$, both the left hand side of (2.39) and $2\log r$ all equal the value c. The derivative of the left hand side equation is

$$\frac{2(r^2 - 1 + r\log r^2)}{(r-1)^3} \tag{2.40}$$

The derivative is always positive, so for $r \leq \sqrt{e}$, the left hand side is less than c. For values $r > \sqrt{e}$, we have

$$\frac{r^2 \log r^2 - r^2 + 1}{(r-1)^2} \leq \frac{r^2 \log r^2 - r^2 + 1}{r^2} \leq 2\log r \tag{2.41}$$

This proves the result. □

2.2.2 Basic Volumes in High Dimensional Spaces

A very useful tool for finding bounds on covering (or packing) numbers is to look at volumes of high dimensional shapes. We include calculations for volumes of balls and simplices and a way to bound covering numbers.

We will use the notation $B_f^K(r)$ to mean a ball in K-dimensional space, with norm f and radius r.

Fact 5 (Volume of K-Dimensional Balls)**.** *Let r be the radius of the Euclidean (ℓ_2) ball $B_{\ell_2}^K(r)$, then*

$$\mathrm{vol}(B_{\ell_2}^K(r)) = \frac{\pi^{\frac{K}{2}}}{\Gamma\left(\frac{K}{2}+1\right)} r^K. \tag{2.42}$$

Let r be the radius of the ℓ_1 ball $B^K_{\ell_1}(r)$, then

$$\mathrm{vol}(B^K_{\ell_1}(r)) = \frac{2^K}{K!} r^K \,. \tag{2.43}$$

.

Fact 6 (Volume of a Simplex)**.** *For a $(K-1)$-dimensional simplex with side lengths $\sqrt{2}$ (this is chosen to match the probability simplex over K symbols), the volume is*

$$\mathrm{vol}(\triangle_{K-1}) = \frac{\sqrt{K}}{(K-1)!} \,. \tag{2.44}$$

The volume under the $(K-1)$-dimensional simplex to the origin is

$$\frac{1}{K!} \tag{2.45}$$

Fact 7. *Let $M(A, f, r)$ be the number of covers needed to cover space $A \subset \mathbb{R}^d$ with norm f and radius r. Let B be the unit ball in norm f (i.e. $B = B^K_f(1)$) and $+$ be the Minkowski sum. Then,*

$$\left(\frac{1}{r}\right)^K \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \leq M(A, f, r) \leq \frac{\mathrm{vol}\left(A + \frac{r}{2}B\right)}{\mathrm{vol}\left(\frac{r}{2}B\right)} \,. \tag{2.46}$$

Note that since KL divergence is not a norm, we cannot apply Fact 7 to our problem directly. However, what we an do upper and lower bound KL divergence with other f-divergences which do behaves as norms, then apply Fact 7 to these other f-divergences.

2.2.3 Dirichlet Distribution

The Dirichlet distribution is a very common prior to use on the simplex.

Definition 7. *The Dirichlet distribution is on support $x_1, ..., x_K$ where $x_i \in (0, 1)$ and $\sum_{i=1}^K x_i = 1$. The parameters are $\alpha_1, ..., \alpha_K$ where $\alpha_i > 0$.*

The pdf of $x = (x_1, \ldots, x_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$ is given by

$$p(x_1, x_2, \ldots, x_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1} \tag{2.47}$$

Recall for positive integers n, $\Gamma(n) = (n-1)!$. Dirichlet distributions are commonly used a priors over the probability simplex. (We also use them in later chapters for examining average-case divergence covering.) The Jeffreys' prior is a particular instance of the Dirichlet distribution when all the parameters are set to $1/2$, that is when the pdf is

$$p(x_1, x_2, \ldots, x_K) = \frac{\Gamma\left(\frac{K}{2}\right)}{\pi^{\frac{K}{2}}} \frac{1}{\sqrt{x_1 x_2 \ldots x_K}} \tag{2.48}$$

The Jeffreys' prior is an important prior on the simplex. We use in this chapter and will discuss it further in Chapter 4.

2.2.4 Other Bounds

It is useful to have easy upper and lower bounds for the binomial coefficient.

$$\frac{n^k}{k^k} \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k \tag{2.49}$$

Also useful is the following:

Fact 8 (Gautschi's Inequality)**.** *For any real positive x and $s \in (0,1)$,*

$$x^{1-s} \leq \frac{\Gamma(x+1)}{\Gamma(x+s)} \leq (x+1)^{1-s} \tag{2.50}$$

2.3 Evenly Spaced Center Points

We begin our first exploration of (worst-case) divergence covering. A reasonable approach for finding covering centers is to start with what happens if the centers are placed on a uniform grid.

We require the set of centers $\mathcal{Q}$ to be on an evenly spaced grid in the simplex. An evenly spaced grid requires that each center $q \in \mathcal{Q}$ has the form

$$q = (q_1, \ldots, q_K) = \left(\frac{a_1}{\gamma}, \ldots, \frac{a_K}{\gamma}\right) \tag{2.51}$$

where each a_i is an non-negative integer and γ is an integer specifying how fine the grid is.

Define $\mathcal{Q}(\gamma, K)$ to be the set of probabilities on K symbols where the probability of any symbol i is of the form $q_i = a_i/\gamma$ where a_i is an integer so that $1 \leq a_i \leq \gamma - (K-1)$.

We now ask, for a fixed γ, what can be say about the covering radius ε?

Upper Bound We will let our centers $\mathcal{Q}$ be exactly the points in the simplex which are in $\mathcal{Q}(\gamma, K)$.

For any p in the simplex, notice we can map p to a center $q \in \mathcal{Q}$ where at for at least $K-1$ of the symbols, $|p_i - q_i| \leq i/\gamma$. For this to work, the remaining symbol must be mapped to the value which allows q to sum up to one. If we let K be this remaining symbol, this means that at most $|p_K - q_K| \leq (K-1)/\gamma$.

$$\begin{aligned} D(P||Q) &\leq \sum_{i=1}^{K} \frac{(p_i - q_i)^2}{1/\gamma} && (2.52)\\ &\leq (K-1)\frac{1/\gamma^2}{1/\gamma} + \frac{(K-1)^2/\gamma^2}{1/\gamma} && (2.53)\\ &= (K-1)\frac{1}{\gamma} + \frac{(K-1)^2}{\gamma} && (2.54)\\ &= \frac{K(K-1)}{\gamma}\,. && (2.55) \end{aligned}$$

We can choose γ so that $\frac{K(K-1)}{\gamma} \leq \varepsilon$ which means we have the constraint $\frac{K(K-1)}{\varepsilon} \leq \gamma$.

This means that the total number of centers is (using (2.49)),

$$\binom{\gamma - 1}{K-1} \leq \left(\frac{\gamma e}{K-1}\right)^{K-1} \approx \left(e\frac{K}{\varepsilon}\right)^{K-1}\,. \tag{2.56}$$

(The approximation is due to integer constraints on γ.)

Lower Bound Because we are dealing with equally spaced points, like in the upper bound, we will assume that all centers have each coordinate of the form $\frac{a}{\gamma}$ for non-negative integers a. It is possible to have a be 0, but a center with a value of 0 will have infinite divergence with any point without a 0 in that coordinate.

To find a lowerbound, it is only necessary to check the distance of some point, say $p = (\delta, \delta, ..., \delta, 1-(K-1)\delta)$ with the $q = \left(\frac{1}{\gamma}, \frac{1}{\gamma}, ..., \frac{1}{\gamma}, \frac{\gamma-(K-1)}{\gamma}\right)$. If $\delta << \frac{1}{\gamma}$, then q is the closest center point to p.

$$\lim_{\delta \to 0} D(p||q) = \lim_{\delta \to 0} (K-1)\delta \log \frac{\delta}{1/\gamma} + (1-(K-1)\delta) \log \frac{1-(K-1)\delta}{1-\frac{K-1}{\gamma}} \tag{2.57}$$

$$= -\log\left(1 - \frac{K-1}{\gamma}\right) \tag{2.58}$$

$$> \frac{K-1}{\gamma}\,. \tag{2.59}$$

Hence there is a point with non-zero coordinates arbitrarily close to having divergence of $\frac{K-1}{\gamma}$ with the closest possible center. We can set $\gamma \geq \frac{K-1}{\varepsilon}$ so that (2.59) can be less than ε. Similar to the above, the number of centers with equally spaced centers is

$$\frac{(\gamma-1)^{K-1}}{(K-1)^{K-1}} \leq \binom{\gamma-1}{K-1}. \tag{2.60}$$

Then there are at least

$$\frac{(\gamma-1)^{K-1}}{(K-1)^{K-1}} \geq \frac{\left(\frac{K-1}{2\varepsilon}\right)^{K-1}}{(K-1)^{K-1}} = \left(\frac{1}{2\varepsilon}\right)^{K-1} \tag{2.61}$$

centers.

Putting thesis upper and lower bounds together gives the following:

Proposition 1. *If we restrict the covering centers to be evenly spaced in the $K-1$ dimensional simplex, then*

$$\left(c\frac{1}{\varepsilon}\right)^{K-1} \leq M_{\textit{evenly spaced}}(K, \varepsilon) \leq \left(C\frac{K}{\varepsilon}\right)^{K-1}. \tag{2.62}$$

Here the covering number depends on the inverse of ε to the power of $K-1$. This exponent is twice that of the exponent we get in the next lower bound.

2.4 Divergence Covering Lower Bound

Here, we show our first and only lower bound.

Proposition 2. *For any $0 \leq \varepsilon \leq \log K$,*

$$M(K, \varepsilon) \geq \left(\frac{1}{8\varepsilon}\right)^{(K-1)/2}. \tag{2.63}$$

Proof. Pick any $q^{(1)}, \ldots, q^{(M)}$ which are centers for a covering in KL divergence for the $(K-1)$-dimensional simplex. For any probability $p \in \triangle_{K-1}$, Pinsker's inequality (2.31) gives

$$\min_j 2TV(p, q^{(j)})^2 \leq \min_j D(p\|q^{(j)}) \leq \varepsilon \tag{2.64}$$

$$\implies \min_j 2\frac{1}{4}\left(\sum_{i=1}^{K} |p - q^{(j)}|\right)^2 \leq \varepsilon \tag{2.65}$$

$$\implies \min_j \sum_{i=1}^{K} |p_i - q_i^{(j)}| \leq \sqrt{2\varepsilon} \tag{2.66}$$

Hence any divergence covering with radius ε should also give a ℓ_1-norm covering of the simplex with radius $\sqrt{2\varepsilon}$. Using (2.46), we can lower bound the number of centers needed in an ℓ_1-norm covering with

$$M \geq \frac{\mathrm{vol}(\triangle_{K-1})}{\mathrm{vol}(B_{\ell_1}^{K-1}(1))(\sqrt{2\varepsilon})^{K-1}} = \frac{\frac{1}{(K-1)!}}{\frac{2^{K-1}}{(K-1)!}(\sqrt{2\varepsilon})^{K-1}} = \left(\frac{1}{8\varepsilon}\right)^{\frac{K-1}{2}} . \tag{2.67}$$

To make this calculation, we projected the the simplex $\triangle_{K-1}$ in $\mathbb{R}^K$ to the space $\mathbb{R}^{K-1}$. The projected simplex is equivalent to the space connecting $\triangle_{K-2}$ to the origin. To compute the volume of this projected simplex, we used (2.45). The volume of the ℓ_1 ball in $\mathbb{R}^{K-1}$ is given by (2.43). □

The $(K-1)/2$ is indeed the correct exponent of $1/\varepsilon$ for divergence covering numbers. This is verified with the upper bound in the next section.

2.5 Divergence Covering Explicit Upper Bound

To match the lower bound, we give is our first divergence covering upper bound which achieves $(K-1)/2$. While it does not give the best bound, we include it because the proof explicitly defines the set of centers needed (instead of arguing that they must exist), which can be useful for practical implementation. It also illustrates how careful placement of centers can improve the exponent compared to that of the evenly spaced covering.

Theorem 1 (Explicit Upper Bound on Divergence Covering)**.** *For $0 < \varepsilon < 1$,*

$$M(K, \varepsilon) \leq c^{K-1} \left(\frac{K-1}{\varepsilon}\right)^{\frac{K-1}{2}} \tag{2.68}$$

for some constant c.

Remark 1. *The constant c above is shown to be less than 7.*

In order to show Theorem 1, we will start with a lemma about comparing the divergence of two points which lie on a line between a and b.

Lemma 1. *For any $a, b \in \Delta_{K-1}$ and $\lambda_1, \lambda_2 \in [0,1]$ we have that*

$$D(\lambda_1 a + (1-\lambda_1)b || \lambda_2 a + (1-\lambda_2)b) \leq \frac{(\lambda_1 - \lambda_2)^2}{\lambda_2(1-\lambda_2)} . \tag{2.69}$$

Proof. (When a scalar λ_i is used in $D(\cdot\|\cdot)$, it is notation for a Bernoulli distribution with probabilities λ_i and $1-\lambda_i$.) Fact 1 implies that

$$D(\lambda_1 || \lambda_2) \leq \frac{(\lambda_1 - \lambda_2)^2}{\lambda_2} + \frac{(\lambda_1 - \lambda_2)^2}{1-\lambda_2} = \frac{(\lambda_1 - \lambda_2)^2}{\lambda_2(1-\lambda_2)} . \tag{2.70}$$

Then using data processing inequality, we get

$$D(\lambda_1 a + (1-\lambda_1)b || \lambda_2 a + (1-\lambda_2)b) \leq D(\lambda_1 || \lambda_2) \leq \frac{(\lambda_1 - \lambda_2)^2}{\lambda_2(1-\lambda_2)} . \tag{2.71}$$

□

To define our covering centers for the simplex, we will start with a set of scalars. Let

$$\Lambda(\varepsilon) \triangleq \left\{ \varepsilon i^2 : \text{ for } i \in \mathbb{Z}_{>0}, \varepsilon i^2 < \frac{1}{2} \right\} \cup \left\{ 1 - \varepsilon i^2 : \text{ for } i \in \mathbb{Z}_{>0}, \varepsilon i^2 < \frac{1}{2} \right\} \cup \left\{ \frac{1}{2} \right\} \tag{2.72}$$

Define

$$\Lambda_2(\varepsilon) = \{(\lambda, 1-\lambda) : \lambda \in \Lambda(\varepsilon)\} \tag{2.73}$$

For each k, let $u_k \in \Delta_{k-1}$ be $u_k = (0, \ldots, 0, 1)$. For each $q \in \Delta_{k-2}$, let $\hat{q}$ be the corresponding $\hat{q} \in \Delta_{k-1}$ where $\hat{q} = (q_1, \ldots, q_{k-1}, 0)$.

For each $q \in \Delta_{k-2}$, define $q^{(\lambda)}$ such that

$$q^{(\lambda)} = \lambda u_k + (1-\lambda)\hat{q}\,. \tag{2.74}$$

For $k > 2$, recursively define

$$\Lambda_k(\varepsilon) \stackrel{\triangle}{=} \bigcup_{\lambda \in \Lambda\left(\frac{\varepsilon}{k}\right)} \left\{ q^{(\lambda)} : q \in \Lambda_{k-1}\left(\frac{k-1}{k}\varepsilon\right) \right\} \tag{2.75}$$

Lemma 2. *For any $p \in \Delta_{k-1}$,*

$$\min_{q \in \Lambda_k(\varepsilon)} D(p\|q) \leq \gamma\varepsilon \tag{2.76}$$

where γ is a constant.

Proof. We will show this by using induction. First, for any $p \in \Delta_1$, we want to show that

$$\min_{q \in \Lambda_2(\varepsilon)} D(p\|q) \leq \gamma\varepsilon\,. \tag{2.77}$$

Suppose that $p \in \Delta_1$ and $p_1 < 1/2$. Then $\varepsilon(i-1)^2 < p_1 \leq \varepsilon i^2$ for some positive integer i. Assume for now that $\varepsilon i^2 < 1/2$. Choose $q = (\varepsilon i^2, 1-\varepsilon i^2) \in \Lambda_2(\varepsilon)$. Note that we must have $1 - \varepsilon i^2 > 1/2$. Using Lemma 1,

$$\begin{aligned}
D(p\|q) &\leq \frac{(p_1 - q_1)^2}{q_1(1-q_1)} && (2.78)\\
&= \frac{(p_1 - \varepsilon i^2)^2}{(\varepsilon i^2)(1-\varepsilon i^2)} && (2.79)\\
&\leq \frac{(\varepsilon(i-1)^2 - \varepsilon i^2)^2}{(\varepsilon i^2)(1-\varepsilon i^2)} && (2.80)\\
&\leq \varepsilon \frac{(-2i+1)^2}{i^2(1/2)} && (2.81)\\
&\leq \varepsilon \frac{4i^2 - 4i + 1}{i^2/2} && (2.82)\\
&\leq 4\varepsilon && (2.83)
\end{aligned}$$

If $\varepsilon i^2 > 1/2$, we can choose $q = (1/2, 1/2)$. For this case, we can also assume $i > 1$, otherwise one center point, $q = (1/2, 1/2)$ is sufficient for covering the whole simplex. Then

$$\begin{aligned}
D(p\|q) &\leq \frac{(p_1 - 1/2)^2}{(1/2)(1/2)} && (2.84)\\
&\leq \frac{(\varepsilon(i-1)^2 - \varepsilon i^2)^2}{1/4} && (2.85)\\
&\leq 4\varepsilon^2(-2i+1)^2 && (2.86)\\
&\leq \frac{4}{2}\varepsilon \frac{4i^2 - 4i + 1}{(i-1)^2} && (2.87)\\
&\leq 18\varepsilon && (2.88)
\end{aligned}$$

where we used that $\varepsilon < 1/(2(i-1)^2)$. This shows that we can set $\gamma = 18$. By symmetry, $\min_{q\in\Lambda_2(\varepsilon)} D(p\|q) \leq \gamma\varepsilon$ holds for $p_1 > 1/2$ as well.

Suppose in dimension $k-1$, that we have for any $p \in \Delta_{k-2}$,

$$\min_{q\in\Lambda_{k-1}(\varepsilon)} D(p\|q) \leq \gamma\varepsilon \tag{2.89}$$

For each $p = (p_1, ..., p_k) \in \Delta_{k-1}$, we will specify a scalar quantity $\lambda_p \in [0, 1]$. If $p_k < 1/2$, like above, we can find a positive integer i where

$$\frac{\varepsilon}{k}(i-1)^2 \leq p_k \leq \frac{\varepsilon}{k} i^2 \tag{2.90}$$

and set

$$\lambda_p = \min\left\{\frac{\varepsilon}{k} i^2, \frac{1}{2}\right\} \in \Lambda\left(\frac{\varepsilon}{k}\right). \tag{2.91}$$

If $p_k > 1/2$, find i such that

$$1 - \frac{\varepsilon}{k} i^2 < p_k \leq 1 - \frac{\varepsilon}{k}(i-1)^2 \tag{2.92}$$

and set

$$\lambda_p = \max\left\{1 - \frac{\varepsilon}{k} i^2, \frac{1}{2}\right\} \in \Lambda\left(\frac{\varepsilon}{k}\right). \tag{2.93}$$

Define $p'_k = (p_k, 1-p_k)$ and $\lambda'_p = (\lambda_p, 1-\lambda_p)$, then similar to above

$$D(p'_k\|\lambda'_p) \leq \gamma\frac{\varepsilon}{k}. \tag{2.94}$$

$$\begin{aligned}
\min_{q\in\Lambda_k(\varepsilon)} D(p\|q) &\leq \min_{q\in\Lambda_k(\varepsilon):q_k=\lambda_p} p_k \log\frac{p_k}{q_k} + \sum_{i=1}^{k-1} p_i \log\frac{p_i}{q_i} && (2.95)\\
&\leq p_k \log\frac{p_k}{\lambda_p} + \min_{q\in\Lambda_k(\varepsilon):q_k=\lambda_p} \sum_{i=1}^{k-1} p_i \log\frac{p_i}{q_i} && (2.96)\\
&\leq p_k \log\frac{p_k}{\lambda_p} + \min_{q\in\Lambda_k(\varepsilon):q_k=\lambda_p} (1-p_k)\log\frac{1-p_k}{1-\lambda_p} + (1-p_k)\sum_{i=1}^{k-1}\frac{p_i}{1-p_k}\log\frac{p_i/(1-p_k)}{q_i/(1-\lambda_p)} && (2.97)\\
&\leq D(p'_k\|\lambda'_p) + (1-p_k)\min_{q'\in\Lambda_{k-1}\left(\frac{k-1}{k}\varepsilon\right)} \sum_{i=1}^{k-1}\frac{p_i}{1-p_k}\log\frac{p_i/(1-p_k)}{q'_i} && (2.98)\\
&\leq \gamma\frac{\varepsilon}{k} + (1-p_k)\gamma\frac{k-1}{k}\varepsilon && (2.99)\\
&\leq \gamma\varepsilon\,. && (2.100)
\end{aligned}$$

□

Proof of Theorem 1. We will use $\mathcal{Q}_k(\varepsilon)$ to denote the set of centers we need to cover Δ_{k-1} with radius ε.

Let

$$\mathcal{Q}_k(\varepsilon) = \Lambda_k\left(\frac{\varepsilon}{\gamma}\right) \tag{2.101}$$

where γ is the constant in Lemma 2. Then using Lemma 2, for $p \in \Delta_{k-1}$,

$$\min_{q\in\mathcal{Q}_k(\varepsilon)} D(p||q) \leq \varepsilon\,. \tag{2.102}$$

Since, $M(k,\varepsilon) \le |\mathcal{Q}_k(\varepsilon)|$, it remains to count the size of each $\mathcal{Q}_k(\varepsilon)$. We will show its size by induction. First, we have that

$$\left|\Lambda\left(\frac{\varepsilon}{\gamma}\right)\right| \le 2\sqrt{\frac{\gamma}{2\varepsilon}} + 1 = \sqrt{\frac{2\gamma}{\varepsilon}} + 1 \le \frac{\sqrt{2\gamma}+1}{\sqrt{\varepsilon}} \tag{2.103}$$

where the last inequality holds if $\varepsilon \le 1$. Therefore we have for some constant c (we can show $c \le 7$),

$$|\mathcal{Q}_2(\varepsilon)| \le c\frac{1}{\sqrt{\varepsilon}}\,. \tag{2.104}$$

For the inductive case, given alphabet size k and any $\varepsilon \le 1$, we have $|\Lambda_k\left(\frac{\varepsilon}{\gamma}\right)| \le c^{k-1}\left(\frac{k-1}{\varepsilon}\right)^{\frac{k-1}{2}}$.

Now consider the case of alphabet size $k+1$. The set $\Lambda_{k+1}\left(\frac{\varepsilon}{\gamma}\right)$ is defined as a set of points which is a product of sets $\Lambda\left(\frac{1}{k}\frac{\varepsilon}{\gamma}\right)$ and $\Lambda_{k+1}\left(\frac{k-1}{k}\frac{\varepsilon}{\gamma}\right)$. This gives

$$|\mathcal{Q}_{k+1}(\varepsilon)| = \left|\Lambda_{k+1}\left(\frac{\varepsilon}{\gamma}\right)\right| = \left|\Lambda\left(\frac{1}{k}\frac{\varepsilon}{\gamma}\right)\right|\left|\Lambda_k\left(\frac{k-1}{k}\frac{\varepsilon}{\gamma}\right)\right| \tag{2.105}$$

$$\le \left(c\frac{1}{\sqrt{\frac{\varepsilon}{k}}}\right)\left(c^{k-1}\left(\frac{k-1}{\varepsilon\frac{k-1}{k}}\right)^{\frac{k-1}{2}}\right) \tag{2.106}$$

$$= c\frac{\sqrt{k}}{\sqrt{\varepsilon}}c^{k-1}\left(\frac{k}{\varepsilon}\right)^{\frac{k-1}{2}} \tag{2.107}$$

$$= c^k\left(\frac{k}{\varepsilon}\right)^{\frac{k}{2}} \tag{2.108}$$

as the number of centers. □

Notice that our explicit arrangement of centers is very structured. Instead of being on an evenly spaced grid, we have a grid-like structure but the spacing is adjusted. A denser number of centers are required near the edges, where it is "harder" to cover points.

Compared to the lower bound Proposition 2, we do not lose in the exponent, but there is an extra K multiplicative factor raised to the exponent. While our construction is not the best we can do, we do not believe that using an explicit grid-like structure can significantly improve this extra K term.

Our construction does not optimize the number of centers. We can make a small modification to improve the number of centers needed. Consider the "slice" in Δ_{K-1} where $p_K = \lambda$. This slice is a Δ_{K-2} dimensional simplex. We can pick a set of points which cover it at radius $\gamma\frac{K-1}{K}\varepsilon\frac{1}{1-\lambda}$. This term will make (2.100) be closer to $\gamma\varepsilon$ and reduce the number of points. However, such a modification would only change the constant and not the extra K multiplicative factor.

Also, though we did not explain it, the choice to allocate radius $\varepsilon\frac{K-1}{K}$ to the previous dimension and ε/K to next dimension is the optimal choice. We could have chosen any two values that add up to ε, but this choice minimizes the number of centers.

Thus there is no real place to make any improvements in the grid-like structure to remove the extra K term which will be raised to the exponent. This might make sense intuitively. If we ignore the edges of the simplex and stick to center, KL divergence (as seen from the χ^2 upper bound) behaves like the square of Euclidean (or ℓ_2) distance. If we pick centers for Euclidean distance fitted on a grid, then two centers might be separated with a distance $2r$. For each center, the hardest points to cover (the ones farthest from the center) are diagonal points, which in each dimension differ from the center point's value by r. The Euclidean distance from the center to this diagonal point is $r\sqrt{K}$. Hence if we set $r\sqrt{K} \le \sqrt{\varepsilon}$, we need to set $r = \sqrt{\varepsilon/K}$. Then points which differ from the center points in only 1 dimension will be covered by multiple points and thus such a covering will not be efficient. We believe the extra factor of K appears for the same reason in our explicit divergence covering.

Hence, for our next approach, we will only show that a set of centers exists. We will not know exactly where they lie.

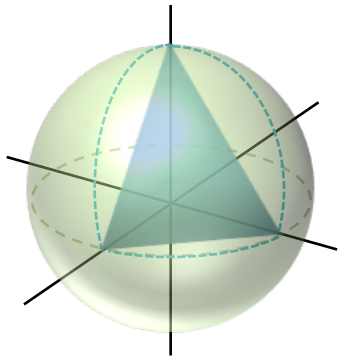


Figure 2.3: Instead of covering the simplex, the using-Hellinger bound first covers the quadrant of the sphere. Point in the quadrant of sphere are mapped back into the simplex. We show the quadrant of the sphere corresponding to $\triangle_2$.

2.6 Using Hellinger: Tighter Upper Bound for All Small Radii

The following bound will be strongest bound we have when the radius ε is small (constant with respect to K).

Theorem 2. *For $\varepsilon \leq \log K$,*

$$M(K, \varepsilon) \leq K \left(C \frac{\log K}{\varepsilon} \right)^{\frac{K-1}{2}} \tag{2.109}$$

Remark 2. *The constant C in the statement of the result can be set to 200. If we are not concerned about the constant C, note that since $K \leq 2^{K-1}$ for positive integers K, we can always remove the first term K in the expression by increasing C to $4C$.*

The main idea of this proof is to bound KL divergence with Hellinger distance. We simply need to make adjustments for how a bound in Hellinger translates into KL. We use the notation S^{K-1} to mean a sphere in K-dimensional space,

$$S^{K-1} = \{\boldsymbol{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K : x_1^2 + \cdots + x_K^2 = 1\}. \tag{2.110}$$

We again use the notation $B_{\ell_2}^K(R)$ as the unit Euclidean ball with radius R.

Proof. First, we consider a set of points which cover $A = \mathbb{R}_{\geq 0}^K \cap S^{K-1}$. We want to cover $A \in \mathbb{R}^K$ with ℓ_2 balls which we denote as $B_{\ell_2}^K\left(\frac{R}{2}\right)$, where R is the radius of the ball.

Using Fact 7, the covering number, denoted as $M(A, \ell_2, R)$, is bounded by

$$M(A, \ell_2, R) \leq \frac{\operatorname{vol}\left(A + B_{\ell_2}^K\left(\frac{R}{2}\right)\right)}{\operatorname{vol}\left(B_{\ell_2}^K\left(\frac{R}{2}\right)\right)} \tag{2.111}$$

$$= \frac{1}{2^K} \frac{\frac{\pi^{K/2}}{\Gamma(K/2+1)} \left((1 + R/2)^K - (1 - R/2)^K\right)}{\frac{\pi^{K/2}}{\Gamma(K/2+1)} R^K} \tag{2.112}$$

$$= c_1^K \frac{\left((1 + R/2)^K - (1 - R/2)^K\right)}{R^K} \tag{2.113}$$

$$(1+R/2)^K-(1-R/2)^K \tag{2.114}$$

$$=\left(1+K\frac{R}{2}+\binom{K}{3}\left(\frac{R}{2}\right)^3+\cdots+\binom{K}{K}\left(\frac{R}{2}\right)^K\right)-\left(1-K\frac{R}{2}+\binom{K}{3}\left(\frac{R}{2}\right)^3+\cdots\pm\binom{K}{K}\left(\frac{R}{2}\right)^K\right) \tag{2.115}$$

$$\leq KR+2\frac{(KR/2)^3}{3!}+\ldots \tag{2.116}$$

$$=2\sinh(KR/2)\,. \tag{2.117}$$

If $Rd<1$, then the sum above is less than c_2dR. Thus, we have for $R\in(0,1/d)$, that$KN(A,\ell_2,R)\leq K(c/R)^{K-1}$.

If $\frac{1}{K}<R<1$, then

$$\frac{\left((1+R/2)^K-(1-R/2)^K\right)}{R^K}\leq\left(\frac{1+R/2}{R}\right)^K \tag{2.118}$$

$$=\left(\frac{1}{R}+\frac{1}{2}\right)^K \tag{2.119}$$

$$\leq\left(\frac{c'}{R}\right)^K \tag{2.120}$$

$$\leq c'K\left(\frac{c'}{R}\right)^{K-1}\,. \tag{2.121}$$

Hence we get for all $R\in(0,1)$, that $N(A,\ell_2,R)\leq K(c/R)^{K-1}$. This means we can find a set of $K(c/R)^{K-1}$ in $\mathbb{R}^K$ which covers the sphere in the positive coordinates with radius R.

We first loosen the requirement so that $A'=\mathbb{R}^K_{>R/\sqrt{K}}\cap S^{K-1}\subset A$ is covered. This requires almost as many points it takes to cover A.

We would like the centers of the cover to be in A'. We will achieve this by mapping each point in our cover to its closest point in A'. This implies that each center will have all its coordinates larger than $R/\sqrt{K}$. Using triangle inequality, these centers are still a cover with of A' if we expand our radius to $2R$.

We currently have a set of points, call the set $\mathcal{Q}_H$, where

- Each point has the form $a=(a_1,...,a_K)$ so that $a_1^2+...+a_K^2=1$
- For each i, $a_i>R/\sqrt{K}$
- $|\mathcal{Q}_H|<K(c/R)^K$

Then for each $b\in\mathbb{R}^K\cap S^{K+1}$, each p is at most distance R from a point in A'. Using triangle inequality again, we have $\|a-b\|_2\leq 3R$.

To make the covering for KL divergence space, we will set

$$\mathcal{Q}_{KL}=\{(a_1^2,...,a_K^2):(a_1,...,a_K)\in\mathcal{Q}_H\}\,. \tag{2.122}$$

Hence if $q\in\mathcal{Q}_{KL}$, then for each $p\in\triangle_{K-1}$, $D_{H^2}(p,q)=\sum_{i=1}^K(\sqrt{p_i}-\sqrt{q_i})^2\leq 9R^2$. We can then use the inequality (2.33) which relates KL divergence to Hellinger distance. This gives

$$D(p||q)\leq 9R^2\frac{1}{(\sqrt{e}-1)^2}\max\left\{1,\max_i\log\frac{p_i}{q_i}\right\}\,. \tag{2.123}$$

Then we will show that

$$\sum_{i=1}^{K}(\sqrt{p_i}-\sqrt{q_i})^2 \le 9R^2 \tag{2.124}$$

$$\implies (\sqrt{p_i}-\sqrt{q_i})^2 \le 9R^2 \tag{2.125}$$

$$\implies |\sqrt{p_i}-\sqrt{q_i}| \le 3R \tag{2.126}$$

$$\implies \sqrt{p_i} \le 3R+\sqrt{q_i} \tag{2.127}$$

$$\implies \frac{\sqrt{p_i}}{\sqrt{q_i}} \le \frac{3R+\sqrt{q_i}}{\sqrt{q_i}} \tag{2.128}$$

$$\implies \frac{\sqrt{p_i}}{\sqrt{q_i}} \le 1+\frac{3R}{R/\sqrt{K}} \tag{2.129}$$

$$\implies \max\left\{1, 2\log\frac{p_i}{q_i}\right\} \le 4\log\left(1+3\sqrt{K}\right). \tag{2.130}$$

Hence we have

$$D(p\|q) \le 9R^2\frac{1}{(\sqrt{e}-1)^2}4\log\left(1+3\sqrt{K}\right) \tag{2.131}$$

$$\le 200R^2\log K\,. \tag{2.132}$$

Setting $R=\sqrt{\frac{\varepsilon}{200\log K}}$ gets the desired result.

□

We could potentially improve the bound if instead of using ℓ_2 distance, we try ℓ_z for larger values of z. However, we cannot get a bound like that in Equation (2.32) to be true. In the proof of Equation (2.32), near $r=1$, the value will explode to infinity. We cannot simply add $\alpha p_i-\beta q_i$ to make the function have a limit near $r=1$.

2.7 Using χ^2: Almost Tight Covering

The goal of this section is to try improve the upper bound on the covering number by drawing random centers according to Jeffreys' prior (or Dirichlet with parameters 1/2). We do not exactly use Jeffreys' prior because it is difficult to show what happens on the edge of the simplex. We instead use a modified Jeffreys' prior. The key will be to use a χ^2 divergence upper bound on KL divergence.

2.7.1 Properties of Ellipsoids

Because KL divergence is not symmetric, we define two different types of divergence balls.

Definition 8 (I-Ball and R-Ball)**.**

$$B_{KL,I}(q,\varepsilon) \triangleq \{p : D(p||q) \le \varepsilon\} \tag{2.133}$$

$$B_{KL,R}(p,\varepsilon) \triangleq \{q : D(p||q) \le \varepsilon\}\,. \tag{2.134}$$

We name I-balls after Information projections. The name R-balls is for reverse I-projection (potentially we will change this name). The R-ball $B_{KL,R}(p,\varepsilon)$ represents all points p in the simplex which can be covered by q at a radius ε.

We will use a χ^2 upper bound for KL divergence in this section. Hence, we also want the notion of a χ^2 ball. These balls will look like ellipsoids in high dimensional space, though the size of the ellipsoid depends on the center point.

Definition 9 (Ellipsoids)**.** *For $a = (a_1, ..., a_K) \in \triangle_{K-1}$ let*

$$\mathcal{E}(a, r^2) \stackrel{\triangle}{=} \left\{ (x_1, ..., x_K) \in \triangle_{K-1} : \sum_{i=1}^{K} \frac{(x_i - a_i)^2}{a_i} \leq r^2 \right\}. \tag{2.135}$$

$$\mathcal{E}_R(a, r^2) \stackrel{\triangle}{=} \left\{ (x_1, ..., x_K) \in \triangle_{K-1} : \sum_{i=1}^{K} \frac{(x_i - a_i)^2}{x_i} \leq r^2 \right\}. \tag{2.136}$$

We could have equivalently defined the sets using notation $D_{\chi^2}(x||a) \leq r^2$ and $D_{\chi^2}(a||x) \leq r^2$. Based on the χ^2 upper bound, every p in $\mathcal{E}(q, \varepsilon)$ will also satisfy the that $D(p\|q) \leq \varepsilon$, so $p \in B_{KL,I}(q, \varepsilon)$. The same thing holds for q in $\mathcal{E}_R(p, \varepsilon)$. This gives

$$\mathcal{E}(q, \varepsilon) \subset B_{KL,I}(q, \varepsilon) \tag{2.137}$$

$$\mathcal{E}_R(p, \varepsilon) \subset B_{KL,R}(p, \varepsilon). \tag{2.138}$$

Lemma 3. *Fix $\varepsilon > 0$. Let $a = (a_1, \ldots, a_K) \in \triangle_{K-1}$ be such that $a_i \geq \varepsilon$ for each i. Then*

$$\mathcal{E}(a, \varepsilon/4) \subset \mathcal{E}_R(a, \varepsilon). \tag{2.139}$$

Instead of $\varepsilon/4$, we can do better by using

$$\frac{\varepsilon}{\frac{3}{2} + \frac{\sqrt{5}}{2}}. \tag{2.140}$$

However we choose to keep the constant simple.

Proof. Let $x \in \mathcal{E}(a, \varepsilon/4)$. Given x, there exists a $\varepsilon_1, \ldots \varepsilon_K$ where each $\varepsilon_i \geq 0$ and $\sum_{i=1}^{K} \varepsilon_i/4 = \varepsilon/4$ such that

$$\frac{(x_i - a_i)^2}{a_i} \leq \frac{\varepsilon_i}{4}. \tag{2.141}$$

This implies that

$$a_i - \sqrt{\frac{\varepsilon_i a_i}{4}} \leq x_i \leq a_i + \sqrt{\frac{\varepsilon_i a_i}{4}}. \tag{2.142}$$

Then

$$\sum_{i=1}^{K} \frac{(a_i - x_i)^2}{x_i} \leq \sum_{i=1}^{K} \frac{1}{4} \frac{\varepsilon_i a_i}{a_i - \sqrt{\frac{\varepsilon_i a_i}{4}}} \tag{2.143}$$

$$\leq \sum_{i=1}^{K} \frac{1}{4} \frac{\varepsilon_i a_i}{a_i - \sqrt{\frac{a_i a_i}{4}}} \tag{2.144}$$

$$= \sum_{i=1}^{K} \frac{1}{4} \frac{\varepsilon_i a_i}{\frac{1}{2} a_i} \tag{2.145}$$

$$= \sum_{i=1}^{K} \frac{1}{2} \varepsilon_i \tag{2.146}$$

$$= \frac{1}{2} \varepsilon \tag{2.147}$$

where we used the condition that $a_i > \varepsilon > \varepsilon_i$ in (2.144). Thus, $x \in \mathcal{E}_R(a, \varepsilon/2) \subset \mathcal{E}_R(a, \varepsilon)$ □

We will need the volume that these ellipsoids cover.

Lemma 4 (Volume of Ellipsoid)**.** *Fix r^2. For $a = (a_1, ..., a_K) \in \triangle_{K-1}$ where $a_i \geq r^2$,*

$$\text{vol}\left(\mathcal{E}(a, r^2)\right) = \frac{\pi^{(K-1)/2}}{\Gamma\left(\frac{K-1}{2} + 1\right)} r^{K-1} \prod_{i=1}^{K} \sqrt{a_i}\,. \tag{2.148}$$

The condition that $a_i \geq r^2$ prevents the ellipsoid from crossing the boundary of the simplex.

Proof. We will integrate over $x_1, ..., x_{K-1} \in \mathbb{R}^{K-1}$ and let $x_K = 1 - x_1 - \cdots - x_{K-1}$.

$$\text{vol}\left(\mathcal{E}(a, r^2)\right) = \int_{x_1} \cdots \int_{x_{K-1}} \mathbb{I}\{(x_1, \ldots, x_K) \in \mathcal{E}(a, r^2)\} dx_1 \ldots dx_{K-1} \tag{2.149}$$

$$= \int_{x_1} \cdots \int_{x_{K-1}} \mathbb{I}\left\{\sum_{i=1}^{K} \frac{(x_i - a_i)^2}{a_i} \leq r^2\right\} \mathbb{I}\left\{\sum_{i=1}^{K} x_i = 1\right\} \prod_{i=1}^{K} \mathbb{I}\{x_i > 0\} dx_1 \ldots dx_{K-1}\,. \tag{2.150}$$

We will do a substitution with $u_i = \frac{x_i}{\sqrt{a_i}} - \sqrt{a_i}$.

$$\text{vol}\left(\mathcal{E}(a, r^2)\right) = \int_{x_1} \cdots \int_{x_{K-1}} \mathbb{I}\left\{\sum_{i=1}^{K} u_i^2 \leq r^2\right\} \mathbb{I}\left\{\sum_{i=1}^{K} u_i \sqrt{a_i} + a_i = 1\right\}$$
$$\prod_{i=1}^{K} \mathbb{I}\{u_i \sqrt{a_i} + a_i > 0\} \sqrt{a_1 \ldots a_K}\, du_1 \ldots du_{K-1} \tag{2.151}$$

$$= \int_{x_1} \cdots \int_{x_{K-1}} \mathbb{I}\left\{\sum_{i=1}^{K} u_i^2 \leq r^2\right\} \mathbb{I}\left\{\sum_{i=1}^{K} u_i \sqrt{a_i} = 0\right\} \prod_{i=1}^{K} \mathbb{I}\{u_i > -\sqrt{a_i}\} \sqrt{a_1 \ldots a_K}\, du_1 \ldots du_{K-1}\,. \tag{2.152}$$

Since $a_i \geq r^2$, the constraint $\mathbb{I}\{u_i > -\sqrt{a_i}\}$ must occur in order for $\mathbb{I}\left\{\sum_{i=1}^{K} u_i^2 \leq r^2\right\}$ to hold. That constraint is redundant and we can remove it.

$$\text{vol}\left(\mathcal{E}(a, r^2)\right) = \sqrt{a_1 \ldots a_K} \int_{x_1} \cdots \int_{x_{K-1}} \mathbb{I}\left\{\sum_{i=1}^{K} u_i^2 \leq r^2\right\} \mathbb{I}\left\{\sum_{i=1}^{K} u_i \sqrt{a_i} = 0\right\} du_1 \ldots du_{K-1}\,. \tag{2.153}$$

The integral is the $K-1$-dimensional volume of the intersection of K-dimensional sphere with radius r and a hyperplane through the origin. By symmetry, this is just the volume of a $K-1$-dimensional sphere with radius r. Therefore,

$$\text{vol}\left(\mathcal{E}(a, r^2)\right) = \sqrt{a_1 \ldots a_K} \frac{\pi^{(K-1)/2}}{\Gamma\left(\frac{K-1}{2} + 1\right)} r^{K-1}\,. \tag{2.154}$$

□

Lemma 5. *Fix $\varepsilon > 0$. Let $a = (a_1, \ldots, a_K) \in \triangle_{K-1}$ where $a_i > \varepsilon$ for each i. Then if*

$$X \sim Dir_K(1/2, 1/2, ..., 1/2) \tag{2.155}$$

then

$$\mathbb{P}(X \in \mathcal{E}(a, \varepsilon)) \geq \sqrt{\frac{2}{K\pi}} \sqrt{\varepsilon}^{K-1}\,. \tag{2.156}$$

Proof. We will use the notation $x = (x_1, \ldots, x_K)$.

For the first step, we use a Taylor expansion on $1/\sqrt{x_i}$ around a_i for each i.

$$\frac{1}{\sqrt{x_i}} = \frac{1}{\sqrt{a_i}} - \frac{x_i - a_i}{2a_i^{3/2}} + \sum_{k=2}^{\infty} (-1)^k \frac{(2k-1)!!}{(2k)!!} \frac{(x_i - a_i)^k}{a_i^{(2k+1)/2}} \tag{2.157}$$

$$= \frac{1}{\sqrt{a_1}} \left(1 - \frac{x_i - a_i}{2a_i} + \sum_{k=2}^{\infty} (-1)^k \frac{(2k-1)!!}{(2k)!!} \frac{(x_i - a_i)^k}{a_i^k} \right) . \tag{2.158}$$

If we constrain $x \in \mathcal{E}(a, \varepsilon)$, then

$$|x_i - a_i| \leq \sqrt{a_i \varepsilon} \tag{2.159}$$

and

$$\left| \frac{x_i - a_i}{a_i} \right| \leq \frac{\sqrt{\varepsilon}}{\sqrt{a_i}} < 1 . \tag{2.160}$$

Since we also know that $(2k-1)!!/(2k)!! < 1$, we can conclude the Taylor series for $1/\sqrt{x_i}$ converges absolutely.

The product of the series for $1/\sqrt{x_i}$ and $1/\sqrt{x_j}$ multipling out all the terms will also converge (see [10, Theorem 3.50]) and this series also converges absolutely (we can just take absolute values of all the terms).

For a given a, we can integrate over the distribution of x, which is Dirichlet, to find the probability X lands in the ellipsoid around a.

$$\mathbb{P}(X \in \mathcal{E}(a, \varepsilon)) = \int_{\mathcal{E}(a,\varepsilon)} \frac{\Gamma(K/2)}{\pi^{K/2}} \frac{1}{\sqrt{x_1 \dots x_K}} \, dx_1 \, \dots \, dx_K \tag{2.161}$$

$$= \int_{\mathcal{E}(a,\varepsilon)} \frac{\Gamma(K/2)}{\pi^{K/2}} \prod_{i=1}^{d} \left(\frac{1}{\sqrt{a_i}} - \frac{x_i - a_i}{2a_i^{3/2}} + \sum_{k=2}^{\infty} (-1)^k \frac{(2k-1)!!}{(2k)!!} \frac{(x_i - a_i)^k}{a_i^{(2k+1)/2}} \right) dx_1 \, \dots \, dx_K \tag{2.162}$$

$$= \int_{\mathcal{E}(a,\varepsilon)} \frac{\Gamma(K/2)}{\pi^{K/2}} \frac{1}{\sqrt{a_1 \dots a_K}} \prod_{i=1}^{K} \left(1 - \frac{x_i - a_i}{2a_i} + \sum_{k=2}^{\infty} (-1)^k \frac{(2k-1)!!}{(2k)!!} \frac{(x_i - a_i)^k}{a_i^k} \right) dx_1 \, \dots \, dx_K . \tag{2.163}$$

For any $x \in \mathcal{E}(a, \varepsilon)$, suppose $x_i = a_i + \delta_i$ for each $i \in [K]$. Let x' be where $x'_i = a_i - \delta_i$ for all i. Then $x' \in \mathcal{E}(a, \varepsilon)$. Hence by symmetry, any term in the integral which is multiplied by an odd power of $(x_i - a_i)$ will cancel out to 0. We can remove these terms in the integration.

$$\mathbb{P}(X \in \mathcal{E}(a, \varepsilon)) = \int_{\mathcal{E}(a,\varepsilon)} \frac{\Gamma(K/2)}{\pi^{K/2}} \frac{1}{\sqrt{a_1 \dots a_K}} \prod_{i=1}^{K} \left(1 + \sum_{k=2, k \text{ is even}}^{\infty} \frac{(2k-1)!!}{(2k)!!} \frac{(x_i - a_i)^k}{a_i^k} \right) dx_1 \, \dots \, dx_K \tag{2.164}$$

$$\geq \frac{\Gamma(K/2)}{\pi^{K/2}} \frac{1}{\sqrt{a_1 \dots a_K}} \int_{\mathcal{E}(a,\varepsilon)} 1 \, dx_1 \, \dots \, dx_K \tag{2.165}$$

$$= \frac{\Gamma(K/2)}{\pi^{K/2}} \frac{1}{\sqrt{a_1 \dots a_K}} \text{vol}(\mathcal{E}(a, \varepsilon)) \tag{2.166}$$

$$= \frac{\Gamma(K/2)}{\pi^{K/2}} \frac{1}{\sqrt{a_1 \dots a_K}} \frac{\pi^{(K-1)/2}}{\Gamma\left(\frac{K-1}{2} + 1\right)} \sqrt{\varepsilon}^{K-1} \prod_{i=1}^{K} \sqrt{a_i} \tag{2.167}$$

$$= \frac{1}{\sqrt{\pi}} \frac{\Gamma(K/2)}{\Gamma\left(\frac{K}{2} + \frac{1}{2}\right)} \sqrt{\varepsilon}^{K-1} \tag{2.168}$$

$$\geq \sqrt{\frac{2}{K\pi}} \sqrt{\varepsilon}^{K-1} . \tag{2.169}$$

We used Lemma 4 to compute $\text{vol}(\mathcal{E}(a,\varepsilon))$, and in the last inequality we used Gautschi's inequality (2.50).

$$\frac{\Gamma(K/2)}{\Gamma\left(\frac{K}{2}+\frac{1}{2}\right)} = \frac{\Gamma(K/2+1)}{\Gamma\left(\frac{K}{2}+\frac{1}{2}\right)}\frac{\Gamma(K/2)}{\Gamma(K/2+1)} \geq (K/2)^{1/2}\frac{1}{K/2} = \sqrt{\frac{2}{K}}\,. \tag{2.170}$$

□

2.7.2 A Modification of Jeffreys' Prior on Simplex

For K and ε, we will define the following modification to the Jeffreys' prior on $\triangle_{K-1}$ (the idea to do this is inspired by [11]). First, we present how this prior, which we denote as $W(K,\varepsilon)$ is generated. Let $x = (x_1, ..., x_K)$ be the probability we are randomly generating. We will need to check that the probabilities we picked are valid (values are between 0 and 1) but we leave that to the end.

First, we define a set S. For each $i \in [K]$, independently with probability $(K+1)/\eta$, let $i \in S$. Then if $i \in S$, with probability $1/(K+1)$, let $x_i = z\frac{\varepsilon}{K}$ for $z \in 0, 1, 2, ..., K$. Thus, we will have with probability $1/\eta$ that $x_i = z\frac{\varepsilon}{K}$ and $i \in S$.

For notation, let $\mathcal{S}[z]$ be the set of all subsets of d elements. Then for each $S \in \mathcal{S}[z]$, let $s = |S|$. We will use S^c to denote the complement of S. Order the indices in S^c as $(j_1, ..., j_{K-s})$.

Next, given what we have specified for each x_i where $i \in S$, to determine the values of each x_i where $i \in S^c$, draw a vector $x' = (x'_1, ..., x'_{K-s})$ using $\text{Dir}_{K-s}(1/2, ...1/2)$. Let $\beta = 1 - \sum_{i\in S} x_i$ and set $x_{j_i} = \beta x'_i$. (Note that it is technically possible for $x_i = z\frac{\varepsilon}{K}$ even if $i \notin S$). This defines all of x. (Because of scaling by β, the values of x does sum up to 1).

Next we need to set η. First, we must have that $(K+1)/\eta < 1$. We will have a stricter condition that $(K+1)/\eta < 1/2$. We set $1/\eta = \sqrt{\varepsilon}$. This gives the condition that

$$(K+1)\sqrt{\varepsilon} \leq \frac{1}{2} \tag{2.171}$$

$$\varepsilon \leq \frac{1}{4(K+1)^2} \tag{2.172}$$

We also need that $\beta < 1$ which will happen if $K\varepsilon \leq 1$, so we have that condition covered.

There is also the caveat that in the way we are generating the distribution. It is possible that all $i \in [K]$ are randomly chosen to be in S. Call this event ζ. Event ζ would not be a valid probability distribution. We need to throw out this possibility from the distribution. Doing so would increase the likelihood of all the other events. Since the next step will be to lower bound the probability of one of these events, it is simpler for our calculation to ignore ζ.

Proposition 3. *Let $\varepsilon \leq \frac{1}{4(K+1)^2}$. Let $p = (p_1, ..., p_K) \in \triangle_{K-1}$ where for each i, either*

- *$p_i > \varepsilon$*
- *or p_i has the form $p_i = k\frac{\varepsilon}{K}$ for $k \in \{0, 1, \ldots, K\}$.*

Then if $q \sim W(K,\varepsilon)$ (as defined above)

$$\mathbb{P}[q \in \mathcal{E}_R(p,\varepsilon)] \geq \sqrt{\frac{1}{d2\pi}}\left(\frac{\sqrt{\varepsilon}}{4}\right)^{K-1} \tag{2.173}$$

Proof. For each $p = (p_1, ..., p_K)$, let S be the set of symbols i where $p_i \leq \varepsilon$. We again use the notation $s = |S|$.

We will use the notation $p|_{S^c}$ and $q|_{S^c}$ to mean the coordinates of p and q respectively where $i \in S^c$. Let $\beta = 1 - \sum_{i\in S} p_i$. We can treat $\frac{p|_{S^c}}{\beta}$ as a probability vector over $K - s$ dimensions.

First, if for $i \in S$, we have $q_i = p_i$, then

$$D_{\chi^2}(p||q) = \sum_i \frac{(p_i - q_i)^2}{q_i} = \sum_{i\in S^c}\frac{(p_i - q_i)^2}{q_i} = \beta\sum_{i\in S^c}\frac{(p_i/\beta - q_i/\beta)^2}{q_i/\beta} = \beta D_{\chi^2}\left(\frac{p|_{S^c}}{\beta}\middle\|\frac{q|_{S^c}}{\beta}\right). \tag{2.174}$$

Then if $\frac{q|_{S^c}}{\beta}$ is such that $\frac{q|_{S^c}}{\beta} \in \mathcal{E}_R\left(\frac{p|_{S^c}}{\beta}, \varepsilon\right)$, this would imply that $D_{\chi^2}(p||q) \leq \beta\varepsilon \leq \varepsilon$. Thus,

$$\mathbb{P}[q \in \mathcal{E}_R(p,\varepsilon)] \geq \mathbb{P}\left[q_i = p_i \text{ for } i \in S \text{ and } D_{\chi^2}\left(\frac{p|_{S^c}}{\beta}\middle\|\frac{q|_{S^c}}{\beta}\right) \leq \varepsilon\right] \tag{2.175}$$

$$= \mathbb{P}\left[q_i = p_i \text{ for } i \in S \text{ and } \frac{q|_{S^c}}{\beta} \in \mathcal{E}_R\left(\frac{p|_{S^c}}{\beta}, \varepsilon\right)\right] \tag{2.176}$$

$$\geq \mathbb{P}\left[q_i = p_i \text{ for } i \in S \text{ and } \frac{q|_{S^c}}{\beta} \in \mathcal{E}\left(\frac{p|_{S^c}}{\beta}, \frac{\varepsilon}{4}\right)\right]. \tag{2.177}$$

For the last inequality, we used Lemma 3. We can check the conditions of Lemma 3. For each $i \in S^c$,

$$\frac{p_i}{\beta} > p_i > \varepsilon\,. \tag{2.178}$$

Given that $p_i = q_i$ for $i \in S$, we know that $\frac{q|_{S^c}}{\beta} \sim \mathrm{Dir}_{K-s}(1/2, ..., 1/2)$. Using Lemma 5 (this requires condition (2.178) again),

$$\mathbb{P}\left[\frac{q|_{S^c}}{\beta} \in \mathcal{E}\left(\frac{p|_{S^c}}{\beta}, \frac{\varepsilon}{4}\right)\middle| q_i = p_i \text{ for } i \in S\right] = \mathbb{P}\left[\frac{q|_{S^c}}{\beta} \in \mathcal{E}\left(\frac{p|_{S^c}}{\beta}, \frac{\varepsilon}{4}\right)\middle| \frac{q|_{S^c}}{\beta} \sim \mathrm{Dir}_{K-s}(1/2, ..., 1/2)\right] \tag{2.179}$$

$$\geq \sqrt{\frac{2}{(K-s)\pi}}\sqrt{\frac{\varepsilon}{4}}^{K-s-1}. \tag{2.180}$$

Putting this all together gives

$$\mathbb{P}[q \in \mathcal{E}_R(p,\varepsilon)] \geq \mathbb{P}\left[q_i = p_i \text{ for } i \in S \text{ and } \frac{q|_{S^c}}{\beta} \in \mathcal{E}\left(\frac{p|_{S^c}}{\beta}, \frac{\varepsilon}{4}\right)\right] \tag{2.181}$$

$$= \mathbb{P}\left[q_i = p_i \text{ for } i \in S\right]\mathbb{P}\left[\frac{q|_{S^c}}{\beta} \in \mathcal{E}\left(\frac{p|_{S^c}}{\beta}, \frac{\varepsilon}{4}\right)\middle| q_i = p_i \text{ for } i \in S\right] \tag{2.182}$$

$$\geq \frac{1}{\eta^s}\left(1 - \frac{K+1}{\eta}\right)^{K-s}\sqrt{\frac{2}{(K-s)\pi}}\sqrt{\frac{\varepsilon}{4}}^{K-s-1} \tag{2.183}$$

$$\geq \sqrt{\varepsilon}^s\left(\frac{1}{2}\right)^{K-s}\sqrt{\frac{2}{(K-s)\pi}}\sqrt{\frac{\varepsilon}{4}}^{K-s-1} \tag{2.184}$$

$$= \sqrt{\varepsilon}^s\sqrt{\frac{1}{(K-s)2\pi}}\left(\frac{\sqrt{\varepsilon}}{4}\right)^{K-s-1} \tag{2.185}$$

$$\geq \sqrt{\frac{1}{d2\pi}}\left(\frac{\sqrt{\varepsilon}}{4}\right)^{K-1}. \tag{2.186}$$

We used that $1/\eta = \sqrt{\varepsilon}$ and that $(K+1)\sqrt{\varepsilon} \leq 1/2$. This completes the proof. □

2.7.3 Expanding Radius of Covering

Since Proposition 3 only applies to probabilities p where if $p_i \leq \varepsilon$, then $p_i = k\frac{\varepsilon}{K}$ for some $k \in \{0, 1, \ldots, K\}$, we need some other means to make sure other values of p where $p_i \leq \varepsilon$ are also covered. To do this, we will show that as long as a certain grid of points in the simplex are covered by a set of centers, expanding the radius will cover the other points in the simplex. We will start with the following definitions to make this precise.

Definition 10. *For $\epsilon = \frac{1}{\kappa}$ where $k \in \mathbb{Z}_+$, let a uniform ϵ-net $\mathcal{N}(\epsilon)$ be a set of points in $\triangle_{K-1}$ which are equally spaced with a distance of ϵ between points. Specifically, $x = (x_1, ..., x_K) \in \mathcal{N}(\epsilon)$ if x_i has the form $\frac{j}{\kappa}$ where $j \in \mathbb{Z}$.*

Let $\mathcal{N}_{int}(\epsilon)$ be the proper subset of $\mathcal{N}(\epsilon)$ where $x = (x_1, ..., x_K) \in \mathcal{N}_{int}(\epsilon)$ if x_i has the form $\frac{j}{\kappa}$ where $j \in \mathbb{Z}$ and $j > 0$.

In other words, $\mathcal{N}_{\text{int}}(\epsilon)$ is the set of points in the ϵ-net which are not on the boundary of the simplex.

Definition 11. *Let $\mathcal{V}_K$ be the set of vectors of the form $(v_1, ..., v_K)$ in K-dimension each v_i has one of the values in $\{-1, 0, 1\}$ and $\sum_{i=1}^K v_i = 0$.*

When we use $\mathcal{N}_{\text{int}}(\epsilon)$, it is possible that ϵ will not necessarily be of the form $1/\kappa$ for $\kappa \in \mathbb{Z}$. In this case, assume that we will pick some $\epsilon' < \epsilon$ where $\epsilon' = 1/\kappa$ for $\kappa \in \mathbb{Z}$ and actually create the net using ϵ'. At most this affects our results by a constant.

Proposition 4. *Consider an $\frac{\varepsilon}{K}$-net $\mathcal{N}(\frac{\varepsilon}{K})$ and a set of $\mathcal{C}$ centers in $\triangle_{K-1}$. Suppose for each point in $x \in \mathcal{N}_{int}(\frac{\varepsilon}{K})$, there exists a $c \in \mathcal{C}$ so that $x \in \mathcal{E}(c, \varepsilon)$. Then for any $v \in \mathcal{V}_K$, the point $y = x + \frac{\varepsilon}{K} v$ is such that*

$$y \in \mathcal{E}(c, \alpha_e \varepsilon) \tag{2.187}$$

where $\alpha_e = 10$.

For this proposition, if c has each coordinate greater than $\frac{\varepsilon}{K}$ (that is $c_i > \frac{\varepsilon}{K}$), then we can show this with one simple inequality which states that $(a+b)^2 \leq 2a^2 + 2b^2$. The fact that $x \in \mathcal{E}(c, \varepsilon)$ implies that

$$\sum_{i=1}^K \frac{(x_i - c_i)^2}{c_i} \leq \varepsilon \tag{2.188}$$

Then we have

$$K \sum_{i=1}^K \frac{(y_i - c_i)^2}{c_i} = \sum_{i=1}^K \frac{(x_i + v_i \frac{\varepsilon}{K} - c_i)^2}{c_i} \tag{2.189}$$

$$\leq 2 \sum_{i=1}^K \frac{(x_i - c_i)^2}{c_i} + 2 \sum_{i=1}^K \frac{\left(v_i \frac{\varepsilon}{K}\right)^2}{c_i} \tag{2.190}$$

$$\leq 2\varepsilon + 2 \sum_{i=1}^K \frac{\varepsilon}{K} \tag{2.191}$$

$$= 4\varepsilon \tag{2.192}$$

For the case where there is some $c_i \leq \frac{\varepsilon}{K}$, we need to use something more precise. Our method will be similar. We will just look at some constraints c must have.

Lemma 6. *Suppose $0 \leq m \leq K$ and that $x_i \geq \frac{\varepsilon}{K}$ for $0 < i \leq m$. Given that $0 < c_i < \varepsilon/K$ for $0 < i \leq m$ and the constraint $\sum_{i=1}^m \frac{(c_i - x_i)^2}{c_i} \leq \varepsilon$, it must be that*

$$\sum_{i=1}^m \frac{1}{c_i} \leq \frac{3}{\varepsilon} K^2 \tag{2.193}$$

Proof. Since we need that $\sum_{i=1}^m \frac{(c_i - x_i)^2}{c_i} \leq \varepsilon$, then there must exist values $a_1, ..., a_m$ such that $a_i > 0$ and $\sum_{i=1}^m a_i \leq 1$ so that

$$\frac{(c_i - x_i)^2}{c_i} = a_i \varepsilon \,. \tag{2.194}$$

We will show that for each i

$$c_i > \frac{\varepsilon}{K^2\left(a_i + \frac{2}{K}\right)}. \tag{2.195}$$

Otherwise,

$$\begin{aligned}
\frac{(c_i - x_i)^2}{c_i} &> \frac{\left(c_i - \frac{\varepsilon}{K}\right)^2}{c_i} && (2.196)\\
&> \frac{\left(\frac{\varepsilon}{K^2\left(a_i+\frac{2}{K}\right)} - \frac{\varepsilon}{K}\right)^2}{\frac{\varepsilon}{K^2\left(a_i+\frac{2}{K}\right)}} && (2.197)\\
&= \frac{K^2\left(a_i + \frac{2}{K}\right)}{\varepsilon}\left(\frac{\varepsilon}{K^2\left(a_i + \frac{2}{K}\right)} - \frac{\varepsilon}{K}\right)^2 && (2.198)\\
&= \frac{K^2\left(a_i + \frac{2}{K}\right)}{\varepsilon}\frac{\varepsilon^2}{K^2}\left(\frac{1 - K\left(a_i + \frac{2}{K}\right)}{K\left(a_i + \frac{2}{K}\right)}\right)^2 && (2.199)\\
&= \left(a_i + \frac{2}{K}\right)\varepsilon\frac{1 - 2d\left(a_i + \frac{2}{K}\right) + K^2\left(a_i + \frac{2}{K}\right)^2}{K^2\left(a_i + \frac{2}{K}\right)^2} && (2.200)\\
&= \varepsilon\frac{1 - 2d\left(a_i + \frac{2}{K}\right) + K^2\left(a_i + \frac{2}{K}\right)^2}{K^2\left(a_i + \frac{2}{K}\right)} && (2.201)\\
&= \varepsilon\left(\left(a_i + \frac{2}{K}\right) - \frac{2}{K} + \frac{1}{K^2\left(a_i + \frac{2}{K}\right)}\right) && (2.202)\\
&> a_i\varepsilon && (2.203)
\end{aligned}$$

which is a contradiction. In (2.197), we used that $x_i > \varepsilon/K$.

Then,

$$\begin{aligned}
\sum_{i=1}^{m}\frac{1}{c_i} &\leq \sum_{i=1}^{m}\frac{1}{\frac{\varepsilon}{K^2\left(a_i+\frac{2}{K}\right)}} && (2.204)\\
&= \frac{K^2}{\varepsilon}\sum_{i=1}^{m}a_i + \frac{2}{K} && (2.205)\\
&\leq 3\frac{K^2}{\varepsilon} && (2.206)
\end{aligned}$$

□

Proof of Proposition 4. Suppose that the coordinates where $c_i \leq \frac{\varepsilon}{K}$ are the first set of coordinates, $c_1, c_2, ..., c_m$, where $0 \leq m \leq K$. Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\sum_{i=1}^{K} \frac{(y_i - c_i)^2}{c_i} = \sum_{i=1}^{K} \frac{(x_i + v_i \frac{\varepsilon}{K} - c_i)^2}{c_i} \tag{2.207}$$

$$\leq 2 \sum_{i=1}^{K} \frac{(x_i - c_i)^2}{c_i} + 2 \sum_{i=1}^{K} \frac{\left(v_i \frac{\varepsilon}{K}\right)^2}{c_i} \tag{2.208}$$

$$\leq 2 \sum_{i=1}^{K} \frac{(x_i - c_i)^2}{c_i} + 2 \sum_{i=1}^{m} \frac{\left(\frac{\varepsilon}{K}\right)^2}{c_i} + 2 \sum_{i=m+1}^{K} \frac{\left(\frac{\varepsilon}{K}\right)^2}{c_i} \tag{2.209}$$

$$\leq 2 \sum_{i=1}^{K} \frac{(x_i - c_i)^2}{c_i} + 2 \left(\frac{\varepsilon}{K}\right)^2 \sum_{i=1}^{m} \frac{1}{c_i} + 2 \sum_{i=m+1}^{K} \left(\frac{\varepsilon}{K}\right) \tag{2.210}$$

$$\leq 10\varepsilon \tag{2.211}$$

The first term (2.210) is less than 2ε since the fact that $x \in \mathcal{E}(c, \varepsilon)$ implies that $\sum_{i=1}^{K} \frac{(x_i - c_i)^2}{c_i} \leq \varepsilon$. The second term in (2.210) is less than 6ε due to Lemma 6. The third term in (2.210) is less than 2ε since it is upperbounded by $2d\frac{\varepsilon}{K}$.

□

2.7.4 A χ^2 Covering of the Simplex

Theorem 3. *For $\varepsilon \leq \frac{1}{4(K+1)^2}$,*

$$M(K, \varepsilon) \leq c_0 K^{3/2} \left(\frac{c_1}{\varepsilon}\right)^{\frac{K-1}{2}} \log \frac{1}{\varepsilon} \tag{2.212}$$

where $c_0 \leq 10$ and $c_1 \leq 4$.

Since $K^{3/2} \leq 5^K$, a simpler way to express the bound in Theorem 3 with explicit constants is

$$M(K, \varepsilon) \leq 10 \left(\frac{20}{\varepsilon}\right)^{\frac{K-1}{2}} \log \frac{1}{\varepsilon} \tag{2.213}$$

Proof. We show the KL divergence covering by given a χ^2 divergence covering. We will find a set of M centers $\mathcal{Q}$ so that for every point $p \in \mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right)$, $p \in \mathcal{E}\left(q, \frac{\varepsilon}{\alpha_e}\right)$ for some q in $\mathcal{Q}$.

For notation in this proof, we will let the result of Proposition 3 be denoted as

$$f(\varepsilon, K) = \sqrt{\frac{1}{K 2\pi}} \left(\frac{\sqrt{\varepsilon}}{4}\right)^{K-1}. \tag{2.214}$$

Using Proposition 3, we know that for a random center drawn from $W(K, \varepsilon)$, any one point $p \in \mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right)$ has probabiltiy greater than $f(\varepsilon/\alpha_e, K)$ of being covered. Suppose M points are drawn with $W(K, \varepsilon/\alpha_e)$. We will use a union bound to upper bound the probability that there exists a p in $\mathcal{N}_{\text{int}}(\frac{\varepsilon}{\alpha_e K})$ which is not covered. Let $\mathcal{Q}(M)$ denote the M randomly selected centers drawn according to $W(K, \varepsilon/\alpha_e)$.

$$\mathbb{P}\left[\exists p \in \mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right), \text{ s.t. } \forall q \in \mathcal{Q}(M), p \notin \mathcal{E}(q, \varepsilon/\alpha_e)\right] \leq \sum_{p \in \mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right)} \mathbb{P}\left[\forall q \in \mathcal{Q}(M), f \notin \mathcal{E}(q, \varepsilon/\alpha_e)\right] \tag{2.215}$$

$$\leq \sum_{p \in \mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right)} (1 - f(\varepsilon/\alpha_e, K))^M \tag{2.216}$$

$$\leq \left|\mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right)\right| (1 - f(\varepsilon/\alpha_e, K))^M. \tag{2.217}$$

$$\left|\mathcal{N}_{\text{int}}\left(\frac{\varepsilon}{\alpha_e K}\right)\right| = \binom{\alpha_e K \frac{1}{\varepsilon} - 1}{K-1} \tag{2.218}$$

$$\leq \left(\frac{e\left(\alpha_e K \frac{1}{\varepsilon} - 1\right)}{K-1}\right)^{K-1} \tag{2.219}$$

$$\leq \left(\frac{\alpha_e e d \frac{1}{\varepsilon}}{2K}\right)^{K-1} \tag{2.220}$$

$$= \left(\frac{(\alpha_e/2)e}{\varepsilon}\right)^{K-1}. \tag{2.221}$$

We want the event in (2.215) to occur with probability less than 1. We can choose $1/2$ for this value. We need to find M so that

$$\left(\frac{(\alpha_e/2)e}{\varepsilon}\right)^{K-1} (1 - f(\varepsilon/\alpha_e, K))^M < \frac{1}{2} \tag{2.222}$$

$$(K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right) + M\log\left(1 - f(\varepsilon/\alpha_e, K)\right) < \log\frac{1}{2} \tag{2.223}$$

$$M\log\left(1 - f(\varepsilon/\alpha_e, K)\right) < \log\frac{1}{2} - (K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right) \tag{2.224}$$

$$M > \frac{\log\frac{1}{2} - (K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right)}{\log\left(1 - f(\varepsilon/\alpha_e, K)\right)} \tag{2.225}$$

$$M > \frac{\log 2 + (K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right)}{-\log\left(1 - f(\varepsilon/\alpha_e, K)\right)}. \tag{2.226}$$

The ensure the condition is met for M, we need an upper bound on the quantity in (2.226). Since $\log(1-y) \leq -y$, which implies that $\frac{1}{-\log(1-y)} \leq \frac{1}{y}$.

We can determine that

$$\frac{\log 2 + (K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right)}{-\log\left(1 - f(\varepsilon/\alpha_e, K)\right)} \leq \frac{\log 2 + (K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right)}{f(\varepsilon/\alpha_e, K)} \tag{2.227}$$

$$\leq \frac{\sqrt{K2\pi}\left(\log 2 + (K-1)\log\left(\frac{(\alpha_e/2)e}{\varepsilon}\right)\right)}{\sqrt{\varepsilon/4}^{K-1}} \tag{2.228}$$

$$\leq c_0 K^{3/2}\left(\log\frac{1}{\varepsilon}\right)\frac{1}{\sqrt{\varepsilon/c_1}^{K-1}} \tag{2.229}$$

Where $c_0 \leq 10$ and $c_1 \leq 4$.

Thus, there exists a random selection of centers where we can cover all the points in $\mathcal{N}_{\text{int}}(\frac{\varepsilon}{\alpha_e K})$. We can then use Proposition 4 to show that by increasing the size from ε/α_e to ε in order to cover all points within the simplex.

The covering in ellipsoids gives a covering in KL divergence using (2.137). □

The using-chi-squared bound presented has some limitations. First, ε must be smaller than roughly $1/K^2$ due to how we have defined the center points. Second, the factor $\log(1/\varepsilon)$ make make this bound must worse than the bound in Theorem 2 if ε is too small.

However, there is a band of ε values where the chi-squared bound is better than the Hellinger bound. We can see this in Figure 2.4. Thus, there are values of ε where the chi-squared bound is the best bound.

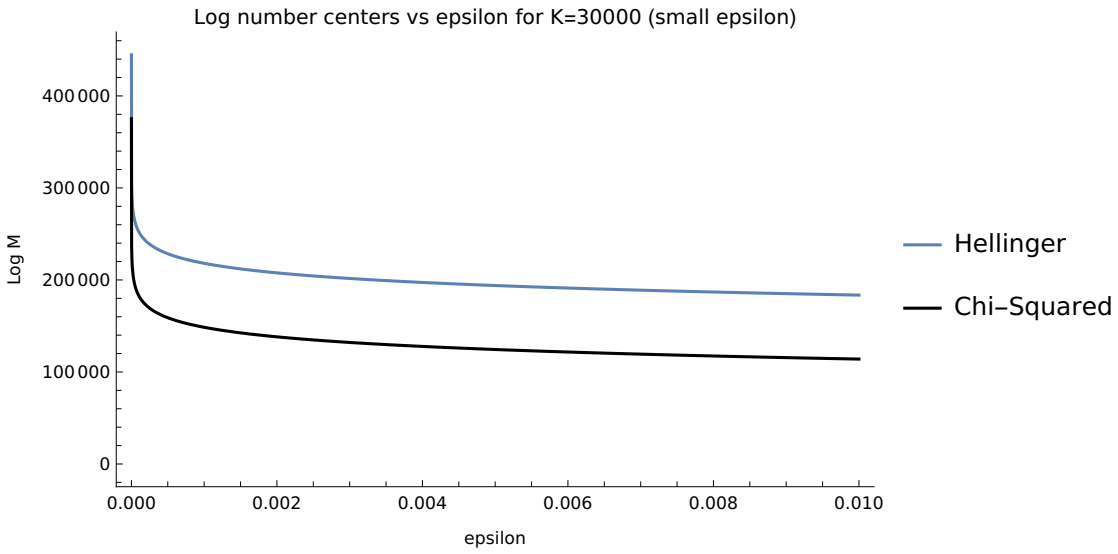


Figure 2.4: Comparing the using-chi-squared bound and the using-Hellinger bound for alphabet size $K = 30000$. We plot the log of the number of centers against values of ε. Only the values of ε where the chi-squared bound is valid is plotted.

We believe the chi-squared bound is likely the technique that will gives us a tight bound. We can possibly improve the analysis and remove the $\log(1/\varepsilon)$ entirely, which is an artifact using the probabilistic method to show all points are covered. There also might be a more clever way modify the Jeffrey's prior and allow ε to take more values.

2.8 Some Discussion

In this chapter, we have given various bounds on finding the (worst-case) divergence covering number. In later chapters, we will show how these bounds can be used to get existing results on redundancy and regret in the setting of universal prediction and use our upper bounds to determine the capacity of the permutation channel.

Some things to consider is how would these bounds differ if instead we are only covering a subset of the simplex. For instance, what would happen if we only needed to cover points on grid in the simplex?

In the next chapter we continue to find more (worst-case) divergence covering bounds, but for the case when ε is larger. We will continue to discuss the bounds in this chapter in the next chapter.

Also, surprisingly, our analysis for average-case divergence covering (spoilers!) gives us an additional bound which is not included in this chapter or the next. We save that discussion for later.

Chapter 3

Divergence Covering Continued: Subexponential Covering

3.1 Introduction

In the previous chapter, our bounds for the divergence covering number $M(K, \varepsilon)$ had a behavior of $1/\varepsilon$ to the power of $(K-1)/2$. This exponent is linear in the dimension (or alphabet size) K and thus the bound is exponential in $1/\varepsilon$. In this chapter we explore the regime where the covering number for KL divergence is subexponential, meaning that as K grows, the number of centers needed grows with an exponent K^c where $c < 1$.

For the subexponential bounds, we want to work with regimes where ε is large. Here is an alternative way of looking at the problem: We know that one center point is sufficient for covering the whole simplex when $\varepsilon = \log K$. If we change ε to $\frac{1}{2}\log K$, how does that change the number of centers needed? This is the key question we are trying to answer. If radius is some fraction of the largest possible radius, can we get exponentially fewer centers?

Summary of Main Results We first present the subexponential bound in an very informal (not quite precise) statement: For constant $\alpha \leq 1$, if the the radius ε is set to $\alpha \log K$,

$$\frac{1}{e^2} K^{1-\alpha} \leq \log M\left(K, \alpha \log K\right) \leq K^{1-\alpha} \log K\,. \tag{3.1}$$

Based on the bound above, if we let $\varepsilon = \frac{1}{2}\log K$, then log of the number of centers should behave as $\sqrt{K}$, ignoring the additional logarithmic factor. In (3.1), the upper and lower bounds are off by a factor of $e^2 \log K$.

Our precise statement of the subexponential bound has one main difference from (3.1). For the upper bound, there is an additional term added to the radius. This term prevents us from just making α smaller to get an arbitrarily small radius. However, this additional term is negligible if α is fixed and K is large. The tightest version of the upper bound is given with the proof. A less tight but accurate bound is the following: For $1/(e \log K) \leq \alpha \leq 1$,

$$M\left(K, \alpha \log K + \log\left(\frac{1}{\alpha} + 1\right)\right) \leq e^{K^{1-\alpha} \log K}\,. \tag{3.2}$$

(Also, in the proofs below, instead of using α in the expression $\alpha \log K$ for the radius, we use $(1/r)\log K$ where $r = 1/\alpha$.)

We also have polynomial bounds, where the number of centers are polynomial in K. This occurs when the radius is $\log K - c$ for some small c.

Implications on Applications Our main motivation for studying the subexponential bounds is for our cloud communication problem (see Section 1.1.2). The small radius regime is for when we want the excess

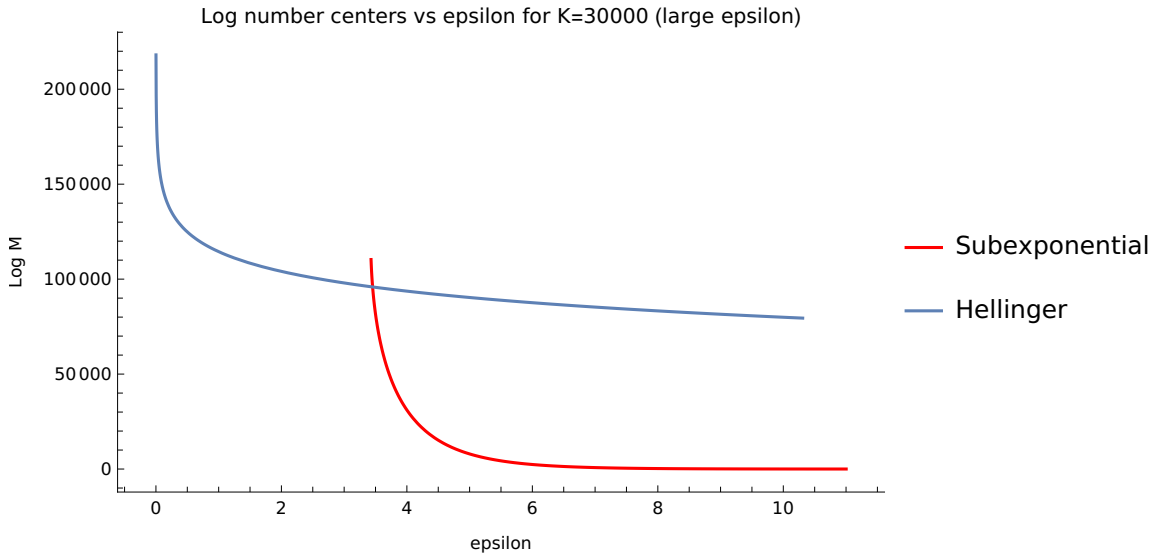


Figure 3.1: When $K = 30,000$, comparison of log of the number of centers needed to cover the simplex plotted against ε. The subexponential bound and the using-Hellinger bounds are plotted.

code length of the subsidiary communication to be within a constant number of bits from optimal. However, given a large alphabet size, this might not be a feasible expectation. If the alphabet is the set of English words, it is typical in language models to use a vocabulary size of around $10,000$ to $50,000$ words (very large vocabulary sizes for language models can go up to $500,000$.) For our example, suppose our system had an alphabet size of $K = 30,000$. Then sending a probability distribution over $30,000$ symbols, so that the excess distortion is within say 1 bit, would require around $173,000$ bits using our Hellinger bound. (The Hellinger bound is best bound for the setting of $K = 30,000$ and $\varepsilon = 1$. The subexponential bound tells us that we would require around $223,000$ bits.) While this is very reasonable (since we are sending information for $30,000$ symbols) it might still be too much. Depending on the application, we may want a trade-off where the probability distribution takes fewer bits to describe and have more distortion.

Our subexponential bound does show that we can get a number of bits sublinear in the size of the alphabet. If instead, we let the excess code length be $\frac{1}{2}\log K \approx 7.5$, then the subexponential bound would tell us that we require around $2,600$ bits, which is a dramatic amount of savings.

Using the Hellinger bound in place of the subexponential bound does not improve the number of savings as much for the large radius setting. (For instance, when $\varepsilon = 7.5$ and $K = 30,000$, the Hellinger bound gives around $129,000$ bits.) We show the comparison in Figure 3.1.

In Figure 3.2 give a table of all our bounds on divergence covering numbers thus far.

Divergence Covering Bounds			
Name	ε Regime	Lower Bound	Upper Bound
using-Hellinger	$0 < \varepsilon < \log K$	$(\frac{c}{\varepsilon})^{\frac{K-1}{2}}$	$(C\frac{\log K}{\varepsilon})^{\frac{K-1}{2}}$
Using-chi-squared	$0 < \varepsilon < \frac{1}{4(K-1)^2}$	$(\frac{c}{\varepsilon})^{\frac{K-1}{2}}$	$(\log\frac{1}{\varepsilon})(\frac{C}{\varepsilon})^{\frac{K-1}{2}}$
Subexponential	$\varepsilon = \frac{1}{r}\log K + \log(r+1), 1 \le r \le e\log K$	$e^{\frac{1}{e^2}K^{1-1/r}}$	$e^{K^{1-1/r}\log K}$
Polynomial	$\varepsilon = \log\frac{K}{\gamma}, \gamma \le \sqrt{\frac{K}{2}}$	$e^{\frac{\gamma}{e}+\log\sqrt{2\pi\gamma/e}}$	$e^{2\gamma\log K}$

Figure 3.2: Table comparing bounds for divergence covering for alphabet size K for different regimes of ε. Value of constant c and C will vary for each bound. The using-Hellinger and using-chi-squared are bounds for small values of ε, whereas the subexponential and polynomial for large values of ε.

Chapter Organization We will first discuss the achievability result for subexponential covering in Section 3.2. This is followed by the converse result in Section 3.3. After the subexponential regime, we explore the polynomial regime in Section 3.4.

In this chapter, it is useful to have the ability to use subscripts on different discrete probability distributions. Hence for $p \in \triangle_{K-1}$, we write the probability for symbol i as $p(i)$. So

$$p = (p(1), \ldots, p(K)) \,. \tag{3.3}$$

3.2 A Subexponential Achieveability

Proposition 5. *For* $1 \le r \le e \log K$,

$$M\left(K, \frac{1}{r}\log K + \log\left(r - (r-1)\frac{1}{K^{1/r}} + \frac{1}{K^{2/r} - K^{1/r}}\right)\right) \le e^{K^{1-\frac{1}{r}} \log K} \,. \tag{3.4}$$

A less tight but easier result to work with is that for $1 \le r \le e \log K$,

$$M\left(K, \frac{1}{r}\log K + \log(r+1)\right) \le e^{K^{1-\frac{1}{r}} \log K} \,. \tag{3.5}$$

We will give an explicit construction to achieve the covering at these distances. The main idea of our construction is to find a cover for all probabilities which are equivalent up to a permutation. The distribution q which will do this will have a tiered structure. Symbols in the smallest tier will have the largest probability. Symbols in the second smallest tier will have a slightly lower probability and this keeps decreasing until we get to a tier with all the symbols.

Proof. We will first define a specific distribution q_e in terms of r.

$$q_e(i) \triangleq \begin{cases} \alpha\frac{1}{K^{1/r}} & \text{for } 1 \le i \le K^{1/r} \\ \alpha\frac{1}{K^{2/r}} & \text{for } K^{1/r} < i \le K^{2/r} \\ \vdots & \\ \alpha\frac{1}{K^{(r-1)/r}} & \text{for } K^{(r-2)/r} < i \le K^{(r-1)/r} \\ \alpha\frac{1}{K} & \text{for } K^{(r-1)/r} < i \le K \end{cases} . \tag{3.6}$$

Note that by bounding $r \le e \log K$, the smallest value of $K^{1/r}$ is $e^{1/e}$. Due to the property that $e^{j/e} \ge j$, this guarantees that $\lfloor K^{t/r} \rfloor < \lfloor K^{(t+1)/r} \rfloor$ for all t. This means the value $\alpha\frac{1}{K^{t/r}}$ for each $t \in \{1, \ldots, r\}$ appears as $q_e(i)$ for some i.

The normalization constant α is given as

$$\frac{1}{\alpha} = \frac{1}{K^{1/r}}\lfloor K^{1/r} \rfloor + \sum_{i=2}^{r} \frac{1}{K^{i/r}}(\lfloor K^{i/r} \rfloor - \lfloor K^{(i-1)/r} \rfloor) \tag{3.7}$$

$$\le \frac{1}{K^{1/r}} K^{1/r} + \sum_{i=2}^{r} \frac{1}{K^{i/r}}(K^{i/r} - K^{(i-1)/r} + 1) \tag{3.8}$$

$$= r - \sum_{i=2}^{r} \frac{1}{K^{1/r}} + \sum_{i=2}^{r} \frac{1}{K^{i/r}} \tag{3.9}$$

$$= r - (r-1)\frac{1}{K^{1/r}} + \frac{1}{K^{2/r} - K^{1/r}} \,. \tag{3.10}$$

For any permutation σ of the K symbols, define distribution q_σ as $q_\sigma(\sigma(i)) = q_e(i)$. Different permutations σ can correspond to the same distribution. Let $\mathcal{Q}_r$ be the set of covering centers, defined as

$$\mathcal{Q}_r \triangleq \bigcup_{\sigma} \{q_\sigma\} \,. \tag{3.11}$$

Let $\bar{p}_k$ for $k = 1, \cdots, K$ be the set of distribution where

$$\bar{p}_k(i) \triangleq \begin{cases} \frac{1}{k} & \text{for } i \leq k \\ 0 & \text{for } i > k \end{cases} . \tag{3.12}$$

For each k, let $t(k)$ be value $t \in \{1, \cdots, r\}$ such that $K^{(t-1)/r} \leq k < K^{t/r}$.

$$D(\bar{p}_k||q_e) = \sum_{i=1}^{k} \frac{1}{k} \log \frac{1/k}{q_e(i)} \tag{3.13}$$

$$\leq \sum_{i=1}^{k} \frac{1}{k} \log \frac{1/k}{\alpha \frac{1}{K^{t(k)/r}}} \tag{3.14}$$

$$\leq \log \frac{\frac{1}{K^{t(k)/r}}}{\alpha \frac{1}{K^{(t(k)+1)/r}}} \tag{3.15}$$

$$= \log \frac{1}{\alpha} + \log \frac{K^{(t(k)+1)/r}}{K^{t(k)/r}} \tag{3.16}$$

$$= \log \frac{1}{\alpha} + \log K^{1/r} . \tag{3.17}$$

For each $p \in \triangle_{K-1}$, there exists a $\tilde{p}$ so that $p(\sigma(i)) = \tilde{p}(i)$ and $\tilde{p}(1) \geq \tilde{p}(2) \geq \cdots \geq \tilde{p}(K)$. The distribution $\tilde{p}$ can be expressed as a linear combination of all the $\bar{p}_k$. Using convexity of KL divergence, this means that

$$D(p||q_\sigma) = D(\tilde{p}||q_e) \leq \log \frac{1}{\alpha} + (1/r) \log K . \tag{3.18}$$

It remains to count the size of $\mathcal{Q}_r$.

$$|\mathcal{Q}_r| = \frac{K!}{(K - K^{(r-1)/r})!(K^{(r-1)/r} - K^{(r-2)/r})! \cdots (K^{2/r} - K^{1/r})! K^{1/r}!} \tag{3.19}$$

$$\leq \frac{K!}{(K - K^{(r-1)/r})!} \tag{3.20}$$

$$\leq K^{K^{(r-1)/r}} . \tag{3.21}$$

□

3.3 A Subexponential Converse

Definition 12. *Let $\mathcal{S}(K, j)$ be the set of all subsets of size j on K symbols.*

Lemma 7. *Let q be a distribution on K symbols, where the probability on the ith symbol is given by $q(i)$. Then for any B where $B > 0$,*

$$\sum_{S \in \mathcal{S}(K,j)} \mathbb{I}\left\{ \frac{1}{B} \leq \left(\prod_{i \in S} q(i) \right)^{1/j} \right\} \leq \frac{B^j}{j!} . \tag{3.22}$$

Note that the expression $(1/B) \leq \left(\prod_{i \in S} q(i)\right)^{1/j}$ is the statement that the geometric mean of the of the probabilities associated with j elements of S have a value greater than $1/B$.

As long as K is sufficiently large, it is clear that the number of subsets which satisfy this geometric mean constraint is not affected by the size of K. For very large K compared to B, most symbols would have probability 0 in order to maximize the number of subsets.

The conjecture is that the distribution which maximizes the number of subsets while satisfying the geometric mean constraint is *roughly* the distribution that gives $1/B$ probability to B symbols and 0 to everything else. It might make sense to conjecture that the optimal distribution should the uniform distribution on a subset of symbols. But the uniform distribution on a subset is not optimal in general, as there are counterexamples that exploit the fact that B might not be an integer. The result of Lemma 7 attempts to achieve the same order as this conjecture would.

Proof. Let q be any distribution. Group the symbols into bins where each bin contains all the symbols which have the same $q(i)$. Let $\mathcal{B}$ be the set of bins.

For each $b \in \mathcal{B}$, let $n(b)$ be the number of elements in bin b. If q_b is the probability of each element in b, let $\alpha(b)$ be such that

$$\left(\frac{1}{B^j}\right)^{\alpha(b)} = q_b \,. \tag{3.23}$$

Define $p(b)$ the total probability (under q) of all elements in b. We can write

$$p(b) = n(b)\left(\frac{1}{B^j}\right)^{\alpha(b)} \,. \tag{3.24}$$

We can think of $p(\cdot)$ as a probability distribution over the bins, where bin b has probability $p(n)$.

Select an ordered sequence (with repeats) of j bins, call this sequence $W = w_1, ..., w_j$. Let $\mathcal{W}$ be the set of all sequences. We can choose one symbol from each bin, to give an ordered (and possibly repeating) set of symbols T.

For each W and its associated T, we have the following equivalence.

$$\frac{1}{B} \le \left(\prod_{i\in T} q(i)\right)^{1/j} \tag{3.25}$$

$$\frac{1}{B^j} \le \prod_{i\in T} q(i) \tag{3.26}$$

$$\frac{1}{B^j} \le \prod_{i=1}^{j} \left(\frac{1}{B^j}\right)^{\alpha(w_i)} \tag{3.27}$$

$$\frac{1}{B^j} \le \left(\frac{1}{B^j}\right)^{\sum_{i=1}^{j}\alpha(w_i)} \tag{3.28}$$

$$(B^j)^{\sum_{i=1}^{j}\alpha(w_i)} \le B^j \,. \tag{3.29}$$

For each W, let $n'(W)$ be the number of symbol sequences where the ith symbol is in bin w_i. Then

$$n'(W) = n(w_1)n(w_2)...n(w_j) \tag{3.30}$$

$$= \prod_{i=1}^{j} p(w_i)(B^j)^{\alpha(w_i)} \tag{3.31}$$

$$= (B^j)^{\sum_{i=1}^{j}\alpha(w_i)} \prod_{i=1}^{j} p(w_i) \,. \tag{3.32}$$

We can upperbound summing over all subsets by summing over all sequences of bins and weighting these

sequences of bins by $n'(W)$. Define $\mathcal{T}(K,j)$ to be the set of ordered lists of j elements from K symbols.

$$\sum_{S\in\mathcal{S}(K,j)} \mathbb{I}\left\{\frac{1}{B} \le \left(\prod_{i\in S} q(i)\right)^{1/j}\right\} = \frac{1}{j!}\sum_{T\in\mathcal{T}(K,j)} \mathbb{I}\left\{\frac{1}{B} \le \left(\prod_{i\in T} q(i)\right)^{1/j}\right\} \tag{3.33}$$

$$\le \frac{1}{j!}\sum_{W\in\mathcal{W}} n'(W)\mathbb{I}\left\{(B^j)^{\sum_{i=1}^j \alpha(w_i)} \le B^j\right\} \tag{3.34}$$

$$= \frac{1}{j!}\sum_{W\in\mathcal{W}} (B^j)^{\sum_{i=1}^j \alpha(w_i)}\left(\prod_{i=1}^j p(w_i)\right)\mathbb{I}\left\{(B^j)^{\sum_{i=1}^j \alpha(w_i)} \le B^j\right\} \tag{3.35}$$

$$\le \frac{B^j}{j!}\sum_{W\in\mathcal{W}} \left(\prod_{i=1}^j p(w_i)\right)\mathbb{I}\left\{(B^j)^{\sum_{i=1}^j \alpha(w_i)} \le B^j\right\} \tag{3.36}$$

$$\le \frac{B^j}{j!}\sum_{W\in\mathcal{W}} \prod_{i=1}^j p(w_i) \tag{3.37}$$

$$= \frac{B^j}{j!}\,. \tag{3.38}$$

We get the last line from the fact that we are summing probabilities of all sequences of bins.

□

Proposition 6.

$$M\left(K, \frac{1}{r}\log K\right) > e^{\frac{1}{e^2}K^{1-\frac{1}{r}}}\,. \tag{3.39}$$

Proof. To show the lowerbound, we will relax the problem to covering only a subset of points in the simplex instead of the whole simplex. Let $\mathcal{P}_j$ be the set of probabilities on $\triangle_{K-1}$. Define distribution $p_S \in \mathcal{P}_j$ for $S \in \mathcal{S}(K,j)$ as,

$$p_S(i) \triangleq \begin{cases} \frac{1}{j} & \text{for } i \in s \\ 0 & \text{for } i \notin s \end{cases}\,. \tag{3.40}$$

For each p_S, the condition that $D(p_S||q) < (1/r)\log K$ is equivalent to the following:

$$D(p_S||q) \le \frac{1}{r}\log K \tag{3.41}$$

$$\sum_{i\in S}\frac{1}{j}\log\frac{1}{jq(i)} \le \frac{1}{r}\log K \tag{3.42}$$

$$\log\frac{1}{j\left(\prod_{i\in S} q(i)\right)^{1/j}} \le \log K^{1/r} \tag{3.43}$$

$$\frac{1}{jK^{1/r}} \le \left(\prod_{i\in S} q(i)\right)^{1/j}\,. \tag{3.44}$$

From Lemma 7,

$$\sum_{S\in\mathcal{S}(K,j)} \mathbb{I}\left\{\frac{1}{jK^{1/r}} \le \left(\prod_{i\in S} q(i)\right)^{1/j}\right\} \le \frac{(jK^{1/r})^j}{j!} \tag{3.45}$$

meaning that at most $(jK^{1/r})^j$ distributions in $\mathcal{P}_j$ can have divergence less than $(1/r)\log K$ to any one distribution q.

Let β be such that $j = K^\beta$. There are

$$\binom{K}{K^\beta} \geq \frac{K^{K^\beta}}{K^{\beta K^\beta}} = K^{(1-\beta)K^\beta} \tag{3.46}$$

total number of distributions in $\mathcal{P}_j$. Hence we need

$$\frac{\binom{K}{j}}{(jK^{1/r})^j/j!} \geq \frac{K^{(1-\beta)K^\beta}}{(K^\beta K^{1/r})^{K^\beta}} K^{\beta K^\beta} e^{-K^\beta} = K^{(1-\beta-\frac{1}{r})K^\beta} e^{-K^\beta} \tag{3.47}$$

number of centers to cover all of them. To compute this, we used that

$$j! \geq \sqrt{2\pi j}\left(\frac{j}{e}\right)^j > K^{\beta K^\beta} e^{-K^\beta}. \tag{3.48}$$

We will write $\beta = (1 - 1/r - \gamma)$, give the result that the number of covers is more than

$$\left(\frac{K^\gamma}{e}\right)^{K^{\left(1-\frac{1}{r}-\gamma\right)}}. \tag{3.49}$$

We can choose $\gamma = \frac{2}{\log K}$ (the value which gives the maximum), so that (3.49) becomes

$$e^{\frac{1}{e^2}K^{1-\frac{1}{r}}}. \tag{3.50}$$

□

Some insights from this proof are that if we are looking for j, the subset size so that $\mathcal{P}_j$ requires the most number of covering centers, for a distance of $\frac{1}{r}\log K$, we should choose $j = K^{1-1/r}/e^2$.

Overall, we are able to get results:

- For a distance of $(1/r)\log K + \log(\log K + 1)$, the number of centers M satisfies $\log M < K^{1-\frac{1}{r}}\log K$.
- For a distance of $(1/r)\log K$, the number of centers M satisfies $\log M > K^{1-\frac{1}{r}}/e^2$.

3.4 Polynomial Region

Using similar techniques as we did for the the subexponential bounds above, we can find upper and lower bounds for when the divergence distance between any point in the simplex and our selected centers for covering balls is less than $\log K - c$ where c is a positive constant. For this region, we can show an upper bound of a polynomial number of covers and a lower bound of a constant number of coverings. These results can be thought of as corollaries of the above results, but we will derive the upper bound separately.

It is more useful to express

$$\log K - c = \log \frac{K}{\gamma} \tag{3.51}$$

where γ is such that $\log \gamma = c$. To distinguish this region from the subexponential region above, we require that $\gamma << K$. To be more exact, we will require that

$$\gamma \leq \sqrt{\frac{K}{2}} \tag{3.52}$$

though for γ close to $\sqrt{K}$ the following result will be the same as Proposition 5.

Corollary 1. *If* $\gamma \leq \sqrt{\frac{K}{2}}$,

$$\log M\left(K, \log \frac{K}{\gamma}\right) \leq 2\gamma \log K\,. \tag{3.53}$$

Proof. For any probability $p \in \triangle_{K-1}$ there are at most 2γ symbols with probability more than $1/2\gamma$. Call these symbols set A. Let

$$q_A(i) \triangleq \begin{cases} \frac{1}{2\gamma} + \frac{1}{2d} & \text{for } i \in A \\ \frac{1}{2d} & \text{otherwise} \end{cases}\,. \tag{3.54}$$

$$\begin{aligned} D(p||q_A) &= \sum_i p(i) \log \frac{p(i)}{q_A(i)} && (3.55)\\ &\leq \sum_{i\in A} p(i) \log \frac{1}{\frac{1}{2\gamma} + \frac{1}{2d}} + \sum_{i \notin A} p(i) \log \frac{\frac{1}{2\gamma}}{\frac{1}{2d}} && (3.56)\\ &= \sum_{i\in A} p(i) \log \frac{2\gamma K}{K+\gamma} + \sum_{i \notin A} p(i) \log \frac{K}{\gamma} && (3.57)\\ &\leq \sum_{i\in A} p(i) \log 2\gamma + \sum_{i \notin A} p(i) \log \frac{K}{\gamma} && (3.58)\\ &\leq \log \frac{K}{\gamma}\,. && (3.59) \end{aligned}$$

where we used that $2\gamma < K/\gamma$ from our assumption (3.52) in the last line.

It remains to count the number of such q_A given by the number of possible sets A.

$$|A| = \binom{K}{2\gamma} \leq \frac{(ed)^{2\gamma}}{(2\gamma)^{2\gamma}} \tag{3.60}$$

and thus

$$\log M\left(K, \log \frac{K}{\gamma}\right) \leq \log |A| = 2\gamma \log K - 2\gamma \log \frac{2\gamma}{e} \leq 2\gamma \log K - c \tag{3.61}$$

where c is positive so long as $2\gamma > e$. □

For the lower bound, we know that there has to be at least a constant number of covers. Using similar techniques as we used to prove Proposition 6, we can at best show that the lower bound is a constant. So nothing is gained from this technique in terms of dependence on K. We show the results to understand the dependence on γ.

Corollary 2.

$$\log M\left(K, \log \frac{K}{\gamma}\right) \geq \frac{\gamma}{e} + \log \sqrt{2\pi\gamma/e}\,. \tag{3.62}$$

Using Proposition 6, we can set

$$\frac{1}{r} = \left(1 - \frac{\log \gamma}{\log K}\right) \tag{3.63}$$

$$\log M\left(K, \log \frac{K}{\gamma}\right) \geq \frac{1}{e^2} K^{\frac{\log \gamma}{\log K}} = \frac{\gamma}{e^2}\,. \tag{3.64}$$

However, we can strengthen this by doing the calculations exactly.

Proof. Using Lemma 7, the number of subset distributions of size m covered by one distribution of the form in (3.54) is at most

$$\left(\frac{ed}{\gamma}\right)^m \frac{1}{2\pi m}\,. \tag{3.65}$$

This gives the number of distributions to cover all subset m distributions as

$$\frac{\binom{K}{m}}{\frac{ed}{\gamma}^m \frac{1}{2\pi m}} = \frac{\gamma^m}{m^m}\sqrt{2\pi m}\,. \tag{3.66}$$

By taking the derivative, the value of m which approximately maximizes this value is $m = \frac{\gamma}{e}$.

This gives the number of centers needed as $e^{\gamma/e}\sqrt{2\pi\gamma/e}$.

□

We believe that the lowerbound is not tight. (Though the upper bound is probably not tight either.) One of the issues with the lowerbound is that while it counts how many subsets are covered, it does not make sure that the subsets are distinct. Working with $m = 2$, assuming that the probabilities which cover subsets of 2 are uniform over some set, we can show using the probabilistic method that the number of probabilities needed is at least $\log K$. If we can show this, it would give the results

$$\log M\left(K, \log\frac{K}{\gamma}\right) \geq \log\log K\,. \tag{3.67}$$

However, we do not know how to prove that uniform probabilities cover subsets best (a strong version of Lemma 7). Hence the above is still a conjecture.

Chapter 4

Application to Redundancy and Regret

4.1 Introduction

From the previous two chapters we explored bounds on the divergence covering number. Now in this chapter, we begin using them. The purpose of the current chapter is to examine how divergence covering bounds, particularly the upper bounds, can be used to find minimax redundancy and minimax regret bounds for universal compression (or universal prediction) problems.

The first part of the chapter is an overview on the setting of universal compression. The second part derives a framework for finding minimax redundancy and minimax regret for the class of iid finite alphabet distributions using our divergence covering results from the previous chapters. We compare the results we get here with existing results. The third part looks at another problem of minimax regret where divergence covering applies, but it requires us to consider divergence coverings of a particular subset of the simplex.

Summary of Results Our goal is to develop and analyze the framework for how divergence covering bounds can be transformed into bounds for minimax redundancy (denoted by R_n^+) and minimax regret (denoted by r_n^+). (Here, n is the length of sequence we are predicting.) Getting upper bounds on minimax redundancy from divergence covering upper bounds is a straight-forward application of Yang-Barron [6]. However, getting minimax regret from divergence covering results requires some additional analysis. This is presented in Theorem 4, where divergence covering numbers are connected to minimax regret if the maximum likelihood distribution and empirical distribution match.

We first compare our bounds using divergence covering and Yang-Barron for finding minimax redundancy for the class of iid large alphabet (n-fold product) distributions, denoted by I_K^n. For redundancy, our method using covering gets

$$R_n^+(I_K^n) \leq \frac{K-1}{2}\log n + O(K) + O(\log\log n) \tag{4.1}$$

for large n compared to K. whereas the (well-known and famous) result is

$$R_n^+(I_K^n) = \frac{K-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma\left(1/2\right)^K}{\Gamma\left(K/2\right)} + o(1)\,. \tag{4.2}$$

Our result (4.1) is not nearly as precise as (4.2), but we are able to capture the first order term exactly. We are not as tight by an additive factor of $O(\log\log n)$, but this term is small since applying log twice drastically reduces the value. The advantages of our bound is though we are hiding the constants in the $O(K)$ and $O(\log\log n)$ terms, our bound is not asymptotic. Unlike (4.2), we do not have a $o(1)$ term that that depends on K being fixed.

For regret, we give the comparison between using our method and the known results in Table 4.1. The bounds using our method is very close to the known result. We are only larger by a factor of $K\log\log K$ in each of the bounds.

Regime	Known Regret	Regret Using Divergence Covering
$n = o(K)$	$r_n^+(\mathcal{I}_K^n) \sim n \log \frac{K}{n}$	$r_n^+(\mathcal{I}_K^n) \leq n \log \frac{K}{n} + O(n \log\log K)$
$K = \Theta(n)$	$r_n^+(\mathcal{I}_K^n) = \Theta(n)$	$r_n^+(\mathcal{I}_K^n) = O(n \log\log n)$
$K = o(n)$	$r_n^+(\mathcal{I}_K^n) \sim \frac{K-1}{2} \log \frac{n}{K}$	$r_n^+(\mathcal{I}_K^n) \leq \frac{K-1}{2} \log \frac{n}{K-1} + O(K \log\log K)$

Table 4.1: The regime is only the approximate values of n and K where our bounds apply. Exact results are given in Proposition 7.

In addition to the iid finite alphabet results, we also have results when the source is an order r Markov model. Typically, the known results for redundancy and regret were discovered using methods which require specific details about the class, such as applying Stirling's approximation for iid distributions. In our framework, for several different problems, once divergence covering bounds are known, the same machinery can be applied to get the minimax regret (and redundancy) bounds.

The last part of this chapter is to apply our framework to get minimax regret over patterns. For the class of patterns, denoted by $\mathcal{I}_\Psi^n$, we can show using our methods that

$$r_n^+(\mathcal{I}_\Psi^n) \leq c n^{1/3} \log^{4/3} n \tag{4.3}$$

for some constant c. Here we are able to improve the existing result, which is

$$r_n^+(\mathcal{I}_\Psi^n) \leq c n^{1/3} \log^4 n \tag{4.4}$$

for some (different) constant c. For our result, we need to show a new covering bound (based on a result known for finding redundancy) and a more intricate connection between divergence covering to regret.

Chapter Organization Background on the universal compression begins in Section 4.2. Specifically in Section 4.2.4 we discuss Yang-Barron [6] which will connect covering results to redundancy. In Section 4.3, we start by showing how divergence covering can be used to upper bound minimax regret. Following this, we show how divergence covering results from the previous chapters can be used to bound regret for iid product distributions sources and Markov sources. In Section 4.4, we give the background on patterns and show how to obtain a result for regret on patterns.

Notational Notes The distributions P and Q can be distribution on many variables. For instance, if random variables X, Y, Z have some joint distribution under P, we can use P_{XYZ} to discrete the joint distribution. We can use $P_{XYZ}(x, y, z)$ to mean the probability that $X = x, Y = y$ and $Z = z$. We can also discuss marginal distributions using notation like P_X or conditional distributions using notation like $P_{Y|Z}$. We use the same notation for parameterized distributions: the distribution on Y parameterized by θ can be written as $P_{Y|\theta}(y|\theta)$ for each y.

4.2 Universal Compression Overview

4.2.1 Introduction to Universal Compression

The core problem of universal compression is the problem of compressing an iid sequence of symbols. The main difficultly is not knowing the probability distribution which generates the sequence, yet the goal is to still be able to compress almost as well as if you had known the probability distribution. We elaborate on this:

Suppose the *estimator* has a sequence X^n, which has n symbols on an alphabet of size K. It is the estimator's job to compress this sequence. Classical Shannon Theory has told us that if the symbols in X^n are generated iid from a distribution P_X, (where $X_i \sim P_X$ for each i), then the estimator should asymptotically achieve a compression rate of $R = H(P_X)$. This means that if the original length of the

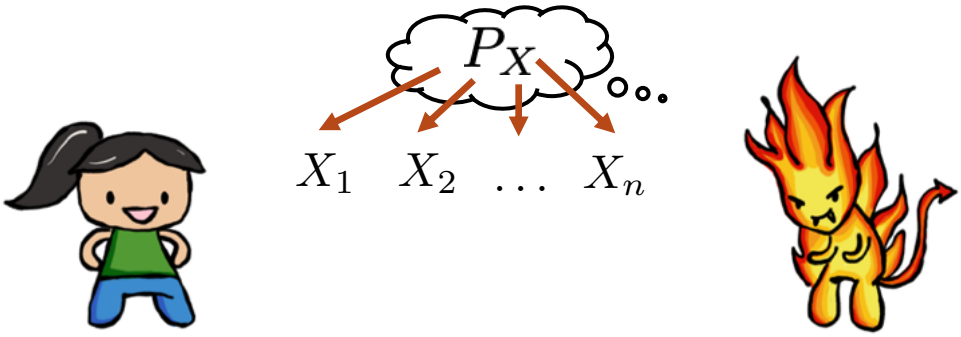


Figure 4.1: Illustration of the prediction game (for redundancy) between the estimator and the devil. The estimator is trying to predict the probability of the next symbol based on past symbols. The symbols are generated with some true distribution P_X.

sequence is n, we expect the compressed sequence to have length nR as $n \to \infty$. To achieves this, it is typical to use a technique like Huffman coding or arithmetic coding.

In the universal compression setting, the estimator is given the same problem, except that the estimator does not know P_X. Perhaps in such a case, one would guess a distribution Q_X and compress with that in place of P_X. Then, the rate of the compression on the sequence would be $R = D(P_X \| Q_X) + H(P_X)$. Compared to if we knew P_X, our rate R is worse by an additional term $D(P_X \| Q_X)$.

However, the estimator is allowed to compress the sequence in an *online* manner. This means that the sequence X^n is revealed to the estimator one at a time. At each time step, estimator must determine their prediction for P_X to use for compressing the next symbol, and this prediction can be informed by all the past symbols they have seen. The question of universal compression is as the estimator repeatedly uses their best prediction of P_X to predict the next symbol, can the estimator eventually achieve a rate which is equal to $H(P_X)$?

At this point, we no longer worry about the mechanics of data compression. The estimator is only evaluated based on how well they predict P_X at each time step. Issues of needing to encode a block of certain length and integer constraints that arise in compression are neglected. We can restate the problem replacing the goal of shortest compression with the goal of getting the lowest loss with the log-loss criteria, making it a universal prediction problem. Both problems are fundamentally the same.

In the universal prediction problem, at each time, the estimator gives the prediction of the next symbol, in the form of a probability distribution, to minimize the log-loss of your probability distribution and the true next symbol. Can this log-loss asymptotically to do a rate of $H(P_X)$?

As we see below, this question has been answered and it turns out that asymptotically the estimator can compress as well as if they knew P_X. We begin to discuss this formally. Our two characters are 1) the *estimator* which we have introduced and 2) the *devil*, a character who chooses the distribution P_X to give the estimator the most trouble.

They are different ways we can define this estimation game between the estimator and the devil. We focus on expected log-loss (giving a quantity called *redundancy*) and individual sequence with log-loss (giving a quantity called *regret*).

4.2.2 Log-Loss and Expected Redundancy

Let n be the total number of symbols (or sequence length) the estimator has to predict. At each time step t, the estimator's task is to produce a function f which minimizes the loss measured by a function $L(f, x)$, where x is the true value revealed at the next time step.

The expected loss at time step t is

$$\min_f \mathbb{E}[L(f, X_t) | X_1^{t-1} = x^{t-1}]\,. \tag{4.5}$$

This notation captures that the fact that the function f can depend on the symbols x_1^{t-1}, which are symbols seen in the past, at steps 1 through $t-1$.

While L can be many different loss functions, we will choose our loss function as the *log-loss* distortion. For log-loss, the estimator's function f needs to be given as a probability distribution over the possible symbols.

Definition 13 (Log-Loss)**.** *Let Q_X be a probability over space $\mathcal{X}$ and $x \in \mathcal{X}$, then the* log-loss *is given by*

$$L(Q_X, x) = \log \frac{1}{Q_X(x)}\,. \tag{4.6}$$

For example: if $\mathcal{X} = [K]$ and $K = 3$, the estimator's f would be a probability Q_X over the symbols $1, 2$ and 3. If say the estimator chooses $Q_X(1) = 1/2, Q_X(2) = 1/4, Q_X(3) = 1/4$ and then afterwards, the true symbols is revealed to be 2, the estimator suffers a loss of

$$L(Q_X, 2) = \log \frac{1}{Q_X(2)} = \log 4\,. \tag{4.7}$$

Redundancy is the loss of the estimator's prediction Q_X against the the true distribution P_X, which generates the symbol X. The estimator's prediction Q_X is evaluated on their *average* log-loss (which is averaged over P_X). The following definitions are from [12]. The first deals with the redundancy of one symbol:

Definition 14 (Redundancy)**.**

$$R(Q_X, P_X) = \mathbb{E}_{X \sim P_X}\left[\log \frac{1}{Q_X(X)} - \log \frac{1}{P_X(X)}\right] = D(P_X || Q_X) \tag{4.8}$$

In the online learning setting, the estimator gives a prediction at each step. This prediction is based on symbols seen in the past. The principled way to denote this prediction is

$$Q_{X_t|X^{t-1}}(x|x^{t-1}) \tag{4.9}$$

where x is the symbol we want the probability for and x^{t-1} are the past $t-1$ symbols. We purposefully write this as a conditional probability distribution, since the rules of conditional probability dictate that for any x^n

$$Q_{X^n}(x^n) = \prod_{t=1}^{n} Q_{X_t|X^{t-1}}(x|x^{t-1})\,. \tag{4.10}$$

This quantity $Q_{X^n}(x^n)$ is relevant if we add up the log-loss taken over all n steps for sequence x^n:

$$\sum_{t=1}^{n} L(Q_{X_t|X^{t-1}}, x_t) = \sum_{t=1}^{n} \log \frac{1}{Q_{X_t|X^{t-1}}(x|x^{t-1})} \tag{4.11}$$

$$= \log \prod_{t=1}^{n} \frac{1}{Q_{X_t|X^{t-1}}(x|x^{t-1})} \tag{4.12}$$

$$= \log \frac{1}{Q_{X^n}(x^n)}\,. \tag{4.13}$$

This development shows that when the estimator gives probabilities as predictions in each step as part of an online process, this is the same as giving a joint probability distribution as a prediction over the whole sequence. Cumulative loss for the estimator's prediction under log-loss can is equal to the log-loss over the whole joint distribution.

If we want to find the cumulative redundancy, we can sum up the log-loss of estimator's prediction with the log-loss of the true distribution at each time step. For this next definition, we don't assume that X^n is

generated iid. We let $X^n \sim P_{X^n}$ for any joint distribution P_{X^n}. If P_{X^n} is known, given X^{t-1}, the symbol X_t is generated as $P_{X_t|X^{t-1}}$.

$$\mathbb{E}_{X^n \sim P_{X^n}} \left[\sum_{t=1}^{n} \log \frac{1}{Q_{X_t|X^{t-1}}(X_t|X^{t-1})} - \log \frac{1}{P_{X_t|X^{t-1}}(X_t|X^{t-1})} \right] \tag{4.14}$$

$$= \mathbb{E}_{X^n \sim P_X^n} \left[\log \frac{1}{Q_{X^n}(X^n)} - \frac{1}{P_{X^n}(X^n)} \right]. \tag{4.15}$$

Definition 15 (Cumulative Redundancy)**.**

$$R_n(Q_{X^n}, P_{X^n}) = \mathbb{E}_{X^n \sim P_X^n} \left[\log \frac{1}{Q_{X^n}(X^n)} - \log \frac{1}{P_{X^n}(X^n)} \right] = D(P_{X^n}||Q_{X^n}). \tag{4.16}$$

If P_{X^n} is an n-fold product distribution, then we can set $P_{X^n} = P_X^n(x^n)$, where use the notation $P_X^n(x^n) = \prod_{t=1}^n P_X(x_t)$ to mean an iid product (or n-fold product) distribution. This will be the case we discuss in this chapter.

So far in the development, our protagonist the estimator has only been faced with some fixed P_X. However, the estimator's prediction, whatever it is, could work better for some P_X's and do terribly for others. What should the estimator's predictions be good against? We use our antagonist, the *devil*, to define this precisely. The devil's purpose is the make things harder for the estimator. By harder, we mean, the devil would like to choose P_X so that the estimator has the largest possible redundancy.

There are different notions of how the devil can make things difficult for the estimator. If in this game, the estimator is forced to pick a prediction strategy first (i.e. choose Q_{X^n} first), then the devil can, based on his knowledge of the estimator's strategy, pick the worst possible P_X^n. The resulting redundancy here is called the *minimax redundancy*. We will use Θ to denote the set of all P_{X^n} the devil can pick. When X^n has an iid product distribution, one distribution P_X specifies P_{X^n}. When this occurs, we will write $P_X = p_\theta$ where $\theta \in \Theta$ is used as the devil's choice.

Something else that can happen, is that the devil must first reveal his strategy. Assume we are only working with iid product distributions. If the devil picks any single P_X, then of course the estimator can always just use $Q_{X^n} = P_X^n$ and get zero redundancy. To make things more interesting, when we say the devil goes first, we let the devil choose a probability distribution on all possible P_X. We call this the *prior*. Then one P_X is chosen randomly according to this prior and this chosen distribution is what the estimator must play their strategy against. Here, we want to average redundancy over the prior. This gives what is called the *maximin redundancy*. Like above, the space of probability distribution at the devil's disposal is given by Θ. The probability the devil can pick over on the space Θ will be denoted by w. This means probability distribution p_θ occurs randomly as the true distribution with probability $w(\theta)$.

While for this work we are focusing on the case where the devil's secret probability is an n-fold product distribution, in general for redundancy, this does not need to be the case. Hence we will use the notation $P_{X^n|\theta}$ to mean a distribution on x^n parameterized by $\theta \in \Theta$. If $P_{X^n|\theta}$ is an n-fold product distribution, then we write $P_{X^n|\theta} = P_{X|\theta}^n$ or p_θ^n as stated above.

The formal definitions of minimax redundancy and maximin redundancy are given here:

Definition 16 (Minimax and Maximin Redundancy)**.** *The minimax redundancy is*

$$R_n^+(\Theta) = \inf_{Q_{X^n}} \sup_{\theta \in \Theta} D(P_{X^n|\theta}||Q_{X^n}). \tag{4.17}$$

For some prior w on Θ define

$$R_n(Q_{X^n}, w) = \mathbb{E}_{\theta \sim w}[D(P_{X^n|\theta}||Q_{X^n})]. \tag{4.18}$$

The maximin redundancy is

$$R_n^-(\Theta) = \sup_w \inf_{Q_{X^n}} R_n(Q_{X^n}, w)\,. \tag{4.19}$$

The setting of maximin redundancy can be understood in Bayesian terms. The quantity $w(\theta)$ is, as its name suggests, a Bayesian prior. If the estimator wants to minimize log-loss, then the best choice for the estimator is to choose Q_{X^n} which is the estimator's belief of the distribution of x^n. This is clear since if we want to choose Q_{X^n} to minimize

$$\mathbb{E}_{X^n \sim P_{X^n}}\left[\log \frac{1}{Q_{X^n}(x)}\right] = D(P_{X^n} \| Q_{X^n}) + H(P_{X^n}) \tag{4.20}$$

picking $Q_{X^n} = P_{X^n}$ is the best choice since it minimizes the KL divergence. Thus, if the estimator knows the prior w, the estimator's best Q_{X^n} is to choose the posterior probability on X^n. This what we call the weighted average estimator and the redundancy we get from using it is called the Bayes Risk.

Definition 17 (Weighted Average Estimator and Bayes Risk)**.** *For any prior w, the weighted average estimator is*

$$m_n^w(x^n) = \int_\Theta P_{X^n|\theta}(x^n) w(\theta) d\theta\,. \tag{4.21}$$

The Bayes risk for prior w is

$$R_n(w) = R_n(m_n^w, w) \tag{4.22}$$

It turns out that the Bayes risk is exactly a mutual information.

$$\inf_{Q_{X^n}} R_n(Q_{X^n}, w) = \inf_{Q_{X^n}} \mathbb{E}_{\theta\sim w}\left[\mathbb{E}_{X^n \sim P_{X^n|\theta}} \log \frac{P_{X^n|\theta}(X^n)}{Q_{X^n}(X^n)}\right] \tag{4.23}$$

$$= \inf_{Q_{X^n}} \int_\theta \sum_{x^n} P_{X^n|\theta}(x^n) w(\theta) \log \frac{P_{X^n|\theta}(x^n)}{Q_{X^n}(x^n)} d\theta \tag{4.24}$$

$$= \int_\theta \sum_{x^n} P_{X^n|\theta}(x^n) w(\theta) \log \frac{P_{X^n|\theta}(x^n)}{m_n^w(x^n)} d\theta \tag{4.25}$$

$$= I(\theta; X^n) \tag{4.26}$$

If we describe minimax redundancy in classical information theory language, we get the following[1]:

$$\inf_{Q_{X^n}} R_n(Q_{X^n}, w) = I(\theta; X^n) \tag{4.27}$$

$$\sup_w \inf_Q R_n(Q, w) = \sup_w I(\theta; X^n) \tag{4.28}$$

$$R_n^-(\Theta) = C_n(\Theta) \tag{4.29}$$

where $C_n(\Theta)$ is the capacity for a channel defined on variables θ and X^n. The optimal w^* which achieves the capacity is called the capacity-achieving prior or the *least-favorable prior*. We will also refer to it as the maximin prior.

The redundancy-capacity theorem says that

$$C_n(\Theta) = R_n^-(\Theta) = R_n^+(\Theta) \tag{4.30}$$

Thus, calculating maximin redundancy will give minimax redundancy.

We will focus on the problem of finding $C_n(\Theta)$ when Θ is the set of n-fold product distributions on a certain fixed alphabet size. We will use $\Theta = \mathcal{I}_K^n$ to denote this.

A long line of work was done in efforts to compute $C_n(\mathcal{I}_K^n)$. In the next section, we go over some of the this history.

[1]The connection to capacity is due to Gallager in unpublished lecture notes [13].

4.2.3 History of Minimax Redundancy

Kolmogorov mentioned the idea of universal compression in [14], but this work was very expository. Fitingof in [15, 16] continued the ideas of [14] with more details. A line of work proceeded from these ideas, summarized in [17]. Among these, an interesting early application of universal coding is the Rice Machine, which was used for transmitting images from the Voyager spacecraft [18]. This line of work eventually lead way to finding $R_n^+(\mathcal{I}_K^n)$ (or $C_n(\mathcal{I}_K^n)$) as a key problem.

The first order term in the expression for $R_n^+(\mathcal{I}_K^n)$ is credited to Krichevsky and Trofimov [19], where they determined that

$$R_n^+(\mathcal{I}_K^n) \sim \frac{K-1}{2} \log n \,. \tag{4.31}$$

Our analysis on using divergence covering for finding $R_n^+(\mathcal{I}_K^n)$ will be about recovering this first order term. However, it is interesting to go over the history of the lower order terms in the expression for $R_n^+(\mathcal{I}_K^n)$.

Authors Clarke and Barron in [20] found that under appropriate conditions, for a given prior w on Θ, and $\theta_0 \in \Theta$,

$$D(p_{\theta_0}^n || m_n^w) = \frac{d}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log \det I(\theta_0) + \log \frac{1}{w(\theta_0)} + o(1) \tag{4.32}$$

where d is the dimension of parameter space Θ. (If $\Theta = \mathcal{I}_K^n$ then $d = K-1$.) They showed this bound by using Laplace's method to approximate $m_n^w(x^n)/p_{\theta_0}^n(x^n)$. Three events need to occur with small probability for the approximation to be accurate. The first is that outside a ball N_δ (with distance measured with the Fisher information) of δ, the integral over $p_\theta(x^n)w(\theta)$ is less than ϵ of that within the ball N_δ. The second is that the empirical Fisher information is close to the theoretical one. The third is that a particular average score function they defined has norm near zero. These events have probabilities that go to zero, but this is assuming that d is fixed.

Continuing down this line, the same authors [21] determined that Jeffreys' prior is asymptotically the worst-case, i.e. the least-favorable prior. (For Jeffreys' prior on the probability simplex, see (2.48).) For each finite n, the prior which is worst-case is a discrete prior (so not Jeffreys'). The minimax risk (under certain assumptions) is then

$$R_n^+(\Theta) = \inf_{q^n} \sup_{\theta \in \Theta} D(p_\theta^n || q^n) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \int_\Theta \sqrt{\det I(\theta)} d\theta + o(1) \,. \tag{4.33}$$

Here, $I(\theta)$ is the Fisher information associated with p_θ^n.

To show this, the important first theorem is that the authors prove on a compact set Θ_0 in the interior of Θ, the Bayes risk converges uniformly. In particular, they showed that

$$\lim_{n \to \infty} \sup_{\theta \in \Theta_0} \left| D(p_\theta^n \| m_n^w) - \left(\frac{d}{2} \log \frac{n}{2\pi e} + \log \frac{\sqrt{\det I(\theta)}}{w(\theta)} \right) \right| = 0 \tag{4.34}$$

The key here is the uniformity. To show this, the authors use Laplace's method, which uses the same methods as in [20]. The expression for this has the empirical Fisher information as the covariance term in the Gaussian approximation. Restricting to a neighborhood around θ and arguing that the empirical Fisher information can be replaced with the theoretical Fisher information gives the desired result.

In order to show the error on this approximation goes to zero, the authors determine upper and lower bounds on the remainder term. To compute the remainder term bounds, the authors need to bound the probability of a few events which are similar to those in [20], and will ultimately show that the remainder terms are $O(1/n)$. Bounding this probability is most of the proof. However, the important thing to notice is that the authors treat d as constant in these bounds. Bounds like $d/(n\varepsilon)$ are $O(1/n)$ but only if d does not grow like n.

To show the result that Jeffreys' prior is maximin, the authors have to argue that they can extend the result on compact subsets in the interior to all of Θ. Then by the fact that the minimax risk and maximin risk are the same and observing that the Jeffreys' prior can be substituted to both equations to get the same

value, the authors conclude that the Jeffreys' prior is the maximin prior and thus give the values for the both the asymptotic minimax and maximin risk.

In Xie and Barron [22], the authors show that for $\Theta = \mathcal{I}_K^n$, both the minimax and maximin redundancy subtracting $\frac{K-1}{2}\log n$ converges to

$$\log \int_{I_K} \sqrt{\det I(\theta)} d\theta = \log \frac{\Gamma(1/2)^K}{\Gamma(K/2)} \tag{4.35}$$

as n goes to infinity. While the Jeffreys' prior is asymptotically least favorable, meaning it is asymptotically the maximin prior, it is not the minimax worst prior. The minimax procedure requires a modification (this modification inspired the modification of Jeffreys' prior in the proof of Theorem 3).

This gives that the final minimax redundancy results as

$$R_n^+(I_K) = \frac{K-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma\left(1/2\right)^K}{\Gamma\left(K/2\right)} + o(1)\,. \tag{4.36}$$

The $o(1)$ term is as $n \to \infty$ for a fixed K.

4.2.4 Yang-Barron: Minimax Bounds Using Covering

Next, we go over the main inequality that lets us connect divergence covering to finding mutual information (which is equal to minimax redundancy).

Authors Yang and Barron in [6] use KL divergence covering to find minimax rates of convergence for density estimation problems. For distortions $d^2(g, f)$ such as squared-error, KL divergence or (squared) Hellinger divergence (see Definition 6), the central result of [6] is that

$$\min_f \max_{g \in G} \mathbb{E}_f d^2(f, g) \asymp \nu_n^2 \tag{4.37}$$

where ν_n^2 is a critical radius related to the packing number. This characterizes the minimax bound in terms of the metric entropy. Their work focuses on the case of densities, but they do use their result in the discrete alphabet case. However, they do not compute the covering numbers necessary; they only determine the rates if the covering number were to have a specific form.

Authors Haussler and Opper [23], similar to [6], use metric entropy to bound the Bayes risk. They show that the Bayes risk is upper bounded by KL divergence covering number and that the Bayes risk is lower bounded by Renyi entropy. For computing the covering numbers, the authors opt for using Hellinger covering. They have theorems where given a dimension of the space (this is defined in terms of the metric entropy), the Bayes risks will be of a certain form. They also do not give covering numbers for finite alphabet spaces.

We will go over one of the inequalities used in [6] which is key to our ideas. This will be used to give bounds for minimax redundancy. Let $\mathcal{Q}$ be a set which covers the probability space in KL divergence with radius ε. (For each $q \in \mathcal{Q}$, we use q^n to be the distribution on the n-fold product). Set

$$Q_{X^n}(x^n) = \frac{1}{|\mathcal{Q}|}\sum_{q \in \mathcal{Q}} q^n(x^n)\,. \tag{4.38}$$

Then for any prior w on θ,

$$I(\theta, X^n) \leq \sup_\theta D(p_\theta^n || Q_{X^n}) \tag{4.39}$$

for any Q_{X^n} (due to property of mutual information. Then,

$$I(\theta, X^n) \leq \sup_\theta \mathbb{E}\left[\log \frac{p_\theta^n(X^n)}{\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} q^n(X^n)}\right] \tag{4.40}$$

$$= \sup_\theta \sum_{x^n} p_\theta^n(x^n) \log \frac{p_\theta^n(x^n)}{\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} q^n(x^n)} \tag{4.41}$$

$$= \log|\mathcal{Q}| + \sup_\theta \sum_{x^n} p_\theta^n(x^n) \log \frac{p_\theta^n(x^n)}{\sum_{q\in\mathcal{Q}} q^n(x^n)} \tag{4.42}$$

$$\leq \log|\mathcal{Q}| + \sup_\theta \sum_{x^n} p_\theta^n(x^n) \log \frac{p_\theta^n(x^n)}{q^n(x^n)} \tag{4.43}$$

$$\leq \log|\mathcal{Q}| + \sup_\theta \inf_q D(p_\theta^n||q^n)\,. \tag{4.44}$$

If $\mathcal{Q}$ is a divergence overing of the space of p_θ at radius ε, then for any p_θ, $\min_{q\in\mathcal{Q}} D(p_\theta\|q) \leq \varepsilon$. Also, since p_θ^n is an iid product distribution, we can relate

$$\max_\theta \min_{q\in\mathcal{Q}} D(p_\theta^n||q^n) = \sup_\theta \inf_{q\in\mathcal{Q}} nD(p_\theta||q) \leq n\varepsilon\,. \tag{4.45}$$

Thus,

$$I(\theta, X^n) \leq \log|\mathcal{Q}| + n\varepsilon \tag{4.46}$$

To find the best bound on the mutual information for each n, we need to find the right trade off between ε and $|\mathcal{Q}|$ by considering different choices of $\mathcal{Q}$. If we are working with p_θ which are over the simplex, then $|\mathcal{Q}| = M(K,\varepsilon)$.

4.2.5 Applying Our Covering Results

In this section we take a break from overviewing literature and apply our divergence covering bounds from Chapter 2 to determine an upper bound on $R_n^+(\mathcal{I}_K^n)$ using (4.44). We know the correct result we should get is (4.36).

Recall that our divergence covering results upper bound $M(K,\varepsilon)$ such that if $m = M(K,\varepsilon)$, then

$$\exists\{Q^{(1)},\ldots,Q^{(m)} : Q^{(i)} \in \triangle_{K-1}\}\, s.t. \sup_{P\in\triangle_{K-1}} D(P||Q^{(i)}) \leq \varepsilon\,. \tag{4.47}$$

We choose to use Theorem 3 for this demonstration. For this bound, we need to assume that $4(K-1)^2 < n$ since we will be setting $\varepsilon = 1/n$. Applying the divergence covering result in Theorem 3, we have that

$$M(K, 1/n) = \left(\frac{c}{1/n}\right)^{\frac{K-1}{2}} \log\frac{1}{1/n}\,. \tag{4.48}$$

Let

$$Q_{X^n}(x^n) = \frac{1}{m}\sum_{t=1}^{m}\prod_{i=1}^{n} Q^{(i)}(x_t)\,. \tag{4.49}$$

Then using (4.44) we have

$$R_n^+(\mathcal{I}_K^n) \le \max_{\theta \in I_K} D(p_\theta^n \| Q_{X^n}) \tag{4.50}$$

$$\le \log m + \max_{\theta \in I_K} \min_{Q^{(i)}} D\left(p_\theta^n \middle\| \prod_{i=1}^n Q^{(i)}\right) \tag{4.51}$$

$$\le \log\left(\left(\frac{c}{1/n}\right)^{\frac{K-1}{2}} \log \frac{1}{1/n}\right) + \max_{\theta \in I_K} \min_{Q^{(i)}} nD(p_\theta \| Q^{(i)}) \tag{4.52}$$

$$\le \frac{K-1}{2} \log(cn) + \log\log n + n\frac{1}{n} \tag{4.53}$$

$$\le \frac{K-1}{2} \log n + O(K) + O(\log\log n) \tag{4.54}$$

We are able to recover the first order term in (4.36). While we do not have the correct constants, our $O(K)$ term is analogous to the second order term in (4.36). Our bound is not tight because we have an additional $O(\log\log n)$ which is not present in (4.36).

However, the advantage of our bound is that it is explicit. For every K and n allowed, we give a value for which the minimax redundancy is definitely less than. There is no $o(1)$ term like in (4.36) which hides possibly large terms for finite n.

Remark 3. *Since we have known results for redundancy, we can translate these known results to lower bounds on divergence covering. For instance, since we know that $\frac{K-1}{2}\log n$ is the correct minimax redundancy, it implies that $M(K,\varepsilon) \ge (1/\varepsilon)^{\frac{K-1}{2}}$. Otherwise we have a contradiction. (We already know this to be a lower bound, but this is a good sanity check.)*

We can also apply our other upper bounds in Chapter 2 or Chapter 3 to get some different bounds on the minimax redundancy. However, instead of doing this, we begin our analysis of regret, a quantity different than though similar to redundancy. It turns out that upper bounds on regret translate to upper bounds on redundancy. We will apply our other bounds on the divergence covering after the next section.

4.2.6 Worst-Case Regret

We now go back to the story of the estimator and the devil. This time, we have them play a slightly different game than before. Instead of having the length n sequence of symbols be generated iid from some distribution P_X, we let the devil choose any length n sequence. Since there is no generating distribution, there is no averaging. The estimator's predictions are evaluated as the log-loss against the devil's single sequence. This is compared against the log-loss of the best possible probability that could have been used for the devil's sequence. Essentially, the *regret*[2] is

$$\inf_Q \sup_{x^n} \left[\log \frac{1}{Q(x^n)} - \inf_P \log \frac{1}{P(x^n)}\right] = \inf_Q \sup_P \sup_{x^n} \log \frac{P(x^n)}{Q(x^n)} \tag{4.55}$$

$$\tag{4.56}$$

where the estimator chooses Q first, and then devil gets to choose both P and x^n. It might seem that it is impossible for the estimator to have a chance of getting good log-loss against a devil who can pick both P and x^n. However, what makes this work is that P is restricted to a particular set. The devil can only choose P where $P = P_{X^n|\theta}(x^n)$ for some $\theta \in \Theta$.

We give the formal definitions below.

Definition 18. *Regret for an estimator Q_{X^n}, a class Θ, and sequence x^n is*

$$r_n(Q_{X^n}, x^n, \Theta) = \sup_{\theta \in \Theta} \log \frac{P_{X^n|\theta}(x^n)}{Q_{X^n}(x^n)}\,. \tag{4.57}$$

[2]Instead of the term *regret*, sometimes the term maximum redundancy is used. In such a case, our notion of redundancy is called average or mean redundancy.

The minimax regret is

$$r_n^+(\Theta) = \inf_{Q_{X^n}} \sup_{x^n} r_n(Q_{X^n}, x^n, \Theta)\,. \tag{4.58}$$

The maximin regret is

$$r_n^-(\Theta) = \sup_{\theta\in\Theta} \inf_{Q_{X^n}} \sum_{x^n} P_{X^n|\theta}(x^n) r_n(Q_{X^n}, x^n, \Theta)\,. \tag{4.59}$$

As the estimator, the optimal prediction Q_{X^n} is the Shtarkov estimator. The Shtarkov estimator is given by

$$Q_n^*(x_1^n) = \frac{1}{\gamma_n(\Theta)} \sup_{\theta\in\Theta} P_{X^n|\theta}(x^n|\theta) \tag{4.60}$$

where

$$\gamma_n(\Theta) = \sum_{x^n} \sup_{\theta\in\Theta} P_{X^n|\theta}(x^n|\theta)\,. \tag{4.61}$$

Similar to capacity redundancy theory, we have for regret that

$$r_n^-(\Theta) = r_n^+(\Theta) = \log\gamma_n\,. \tag{4.62}$$

The Shtarkov estimator is the unique minimax strategy and the unique least favorable distribution for the maximin. This follows from computation similar to that discussed in Section 4.2.4. Compared to the redundancy setting, we have the following important fact:

Fact 9.

$$\log\gamma_n(\Theta) \geq C_n(\Theta) \tag{4.63}$$

Proof. We will fix a specific Q_{X^n}.

$$\begin{aligned}
C_n(\Theta) &\leq \sup_{\theta\in\Theta} \mathbb{E}_{x^n\sim P_{X^n|\theta}} \left[\log\frac{1}{Q_{X^n}(x^n)} - \log\frac{1}{P_{X^n|\theta}(x^n|\theta)}\right] && (4.64)\\
&\leq \sup_{\theta\in\Theta} \mathbb{E}_{x^n\sim P_{X^n|\theta}} \left[\log\frac{1}{Q_{X^n}(x^n)} - \inf_{\theta'\in\Theta}\log\frac{1}{P_{X^n|\theta}(x^n|\theta')}\right] && (4.65)\\
&= \sup_{\theta\in\Theta} \sum_{x^n} P_{X^n|\theta}(x^n|\theta) \left[\log\frac{1}{Q_{X^n}(x^n)} - \inf_{\theta'\in\Theta}\log\frac{1}{P_{X^n|\theta}(x^n|\theta')}\right] && (4.66)\\
&\leq \sup_{x^n} \left[\log\frac{1}{Q_{X^n}(x^n)} - \inf_{\theta'\in\Theta}\log\frac{1}{P_{X^n|\theta'}(x^n|\theta')}\right] && (4.67)
\end{aligned}$$

If we choose Q_{X^n} to minimize (4.67), then (4.67) is equal to $\log\gamma_n$. □

Authors Xie and Barron [11] find that for finite alphabet of size K, the minimax regret is

$$r_n(\mathcal{I}_K^n) = \frac{K-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(1/2)^K}{\Gamma(K/2)} + o(1) \tag{4.68}$$

The Dirichlet prior with all parameters 1/2 (Jeffreys' prior!) is asymptotically maximin for the regret. Compared to minimax redundancy, the difference is an additive factor of $\frac{K-1}{2}\log e$.

4.2.7 Large Alphabet Redundancy and Regret

The formulas we give above for the redundancy (4.36) and regret (4.68) is for when when the alphabet size K is fixed and the sequence length n is very large (accurate when n goes to infinity). They do not give satisfying enough solutions when the alphabet size is large compared to the number of samples. In other words, the $o(1)$ term that appears in (4.36) and (4.68) depends on K. If we disregard this fact and use the formula anyways, then if we let the alphabet size $K = n$, this results in a minimax redundancy (or regret) of order $n \log n$. This value is really large! It is the regret that the estimator is guaranteed to achieve if the estimator predicted that each symbol was equally likely at every step. We expect that perhaps we can do better. We start the discussion of these better results for large alphabet size with some history.

Davisson in his 1973 paper [17] on noiseless universal coding defines that a universal code as possible if the redundancy goes to zero asymptotically when normalized by n (the number of samples). Davisson includes multiple definitions of what it means to be universal, such as weighted universal codes, weakly minimax universal and minimax universal. Weakly minimax universal corresponds to our definition. For the universal code to exist, the per letter minimax redundancy has to go to zero (per letter redundancy is the cumulative redundancy divided by n). Minimax universal is a stronger case where the convergence of the redundancy is uniform over the set of all possible parameters. He observed that for a stationary ergodic process over a finite alphabet, weakly minimax universal codes exists (and if the source satisfies certain entropy requirements, minimax universal codes exists.) The agrees with our background results on minimax redundancy from above. However, for infinite alphabets however, additional conditions are required to show that weakly minimax universal codes exists.

Following this, Keiffer in [24] clarified what the conditions for the existence of weakly minimax universal codes are. His paper also showed that for an iid source with an infinite alphabet, there is indeed no way to get a weakly universal minimax code. We can infer from this that for large alphabet sizes (and small n), we will have per letter redundancy which is fixed above some constant. Then the cumulative redundancy should at least be a constant times the sample size n.

In [25], the authors determine the regret for large alphabet iid distributions. They compute the regret for sample lengths n and alphabet sizes K for the cases where $K = o(n)$, $K = \Theta(n)$ and $n = o(K)$. Their results are summarized below. The upper bound results heavily utilize results from [26].

Regime	Regret
$K = o(n)$	$r_n^+(\mathcal{I}_K^n) \sim \frac{K-1}{2} \log \frac{n}{K}$
$K = \Theta(n)$	$r_n^+(\mathcal{I}_K^n) = \Theta(n)$
$n = o(K)$	$r_n^+(\mathcal{I}_K^n) \sim n \log \frac{K}{n}$

Table 4.2: Table with regrets for standard iid distribution. Results are given for asymptotic n.

Because there is no diminishing per-symbol redundancy in the large alphabet case, another direction in the study of redundancy for large alphabet is to consider *shapes* and *patterns* of infinite alphabets sequences. These are explored in [25] and we discuss them in Section 4.4.

4.3 Connection to Minimax Regret

Yang-Barron's technique (4.44) directly connects divergence covering to minimax redundancy. This makes sense since minimax redundancy is a mutual information and (4.44) is a bound on mutual information. What is less obvious is what divergence covering can show for minimax regret, as minimax regret is not expressed as mutual information. Yet, it is possible to finagle minimax regret so that it is also bounded by (4.44).

The connection between divergence covering and regret requires the following definitions.

Definition 19. *Define $\hat{\theta}(x^n)$ to be the parameter such that*

$$\hat{\theta}(x^n) = \arg\max_{\theta\in\Theta} P_{X^n|\theta}(x^n|\theta)\,. \tag{4.69}$$

In other words, $\hat{\theta}(x^n)$ is the parameter that gives the maximum likelihood estimator for x^n.

Definition 20. *Define $\hat{p}(x;x^n)$ to be the empirical distribution for x^n, i.e.*

$$\hat{p}(x;x^n) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{x_t = x\}\,. \tag{4.70}$$

Theorem 4. *If for each x^n and $x \in \mathcal{X}$ we have that*

$$\hat{p}(x;x^n) = P_{X|\theta}(x|\hat{\theta}(x^n)) \tag{4.71}$$

and we have a divergence covering with set of centers $\mathcal{Q}_n$ on the distributions parameterized by $\theta \in \Theta$ at radius ε, then

$$r_n^+(\Theta) \le \log|\mathcal{Q}_n| + n\varepsilon \tag{4.72}$$

To find the best possible bound, we would need to determine the optimal ε.

Proof. For any ε, the conditions of the theorem states that there exists a divergence covering with centers $\mathcal{Q}_n = Q_{(1)}, ..., Q_{(m)}$ where $m = |\mathcal{Q}_n|$. Recall this implies that for any $\theta \in \Theta$ we have $\min_j D(P_{X|\theta}\|Q_{(j)}) \le \varepsilon$.

Let $j(\theta) = \arg\min_j D(P_{X|\theta}\|Q_{(j)})$ and let

$$\bar{Q}^n(x^n) = \frac{1}{m}\sum_{j=1}^{m}\prod_{t=1}^{n} Q_{(j)}(x_t) \tag{4.73}$$

$$\begin{aligned}
r_n^+(\Theta) &\le \max_{x^n} r_n(\bar{Q}^n, x^n, \Theta) && (4.74)\\
&= \max_{x^n} \log \frac{P_{X^n|\theta}(x^n|\hat{\theta}(x^n))}{\bar{Q}^n(x^n)} && (4.75)\\
&= \max_{x^n} \log \frac{P_{X^n|\theta}(x^n|\hat{\theta}(x^n))}{\frac{1}{m}\sum_{j=1}^{m}\prod_{t=1}^{n} Q_{(j)}(x_t)} && (4.76)\\
&\le \log m + \max_{x^n} \log \frac{P_{X^n|\theta}(x^n|\hat{\theta}(x^n))}{\prod_{t=1}^{n} Q_{(j(\hat{\theta}(x^n)))}(x_t)} && (4.77)\\
&= \log m + \max_{x^n} \log \frac{\prod_{t=1}^{n} P_{X|\theta}(x_t|\hat{\theta}(x^n))}{\prod_{t=1}^{n} Q_{(j(\hat{\theta}(x^n)))}(x_t)} && (4.78)\\
&= \log m + \max_{x^n} \sum_{t=1}^{n} \log \frac{P_{X|\theta}(x_t|\hat{\theta}(x^n))}{Q_{(j(\hat{\theta}(x^n)))}(x_t)} && (4.79)\\
&= \log m + \max_{x^n} \sum_{x\in\mathcal{X}} n\hat{p}(x;x^n) \log \frac{P_{X|\theta}(x|\hat{\theta}(x^n))}{Q_{(j(\hat{\theta}(x^n)))}(x)} && (4.80)\\
&= \log m + n\max_{x^n} \sum_{x\in\mathcal{X}} \hat{P}_{X|\theta}(x|\hat{\theta}(x^n)) \log \frac{P_{X|\theta}(x|\hat{\theta}(x^n))}{Q_{(j(\hat{\theta}(x^n)))}(x)} && (4.81)\\
&= \log m + n\max_{x^n} D(\hat{P}_{X|\theta}(\cdot|\hat{\theta}(x^n))\|Q_{(j(\hat{\theta}(x^n)))}(\cdot)) && (4.82)\\
&= \log m + n\varepsilon && (4.83)
\end{aligned}$$

□

The key to this proof relies on the fact that for regret, the devil will always choose the maximum likelihood estimator for his chosen sequence x^n. If the maximum likelihood estimator reflects the empirical distribution of x^n, then taking log of the maximum likelihood estimator will give a quantity that resembles a KL divergence. This allows us to connect regret for these problems to divergence covering.

Since minimax regret is always larger than minimax redundancy, bounding the regret automatically gives a bound on the redundancy. Our divergence covering results for finite dimensional simplices corresponds to regret problems on iid product distributions. We find these exact bounds given by this below.

4.3.1 Regret for IID Distributions Using Divergence Covering

In this section we will use Theorem 4 and our covering results to find upper bounds on minimax regret for the standard iid product distribution on any alphabet size relative to the number of samples.

For each divergence covering bound we have, we can use Theorem 4 and then optimize for the best value of ε to use. We will start with the subexponential bound in Proposition 5. We will use the simpler version where

$$\log M\left(K, \frac{1}{r}\log K + \log(r+1)\right) \leq K^{1-\frac{1}{r}} \log K \tag{4.84}$$

for $1 < r < e \log K$. We have that for any alphabet size K and n samples, that

$$r_n^+(\mathcal{I}_K^n) \leq K^{1-\frac{1}{r}} \log K + n\left(\frac{1}{r}\log K + \log(r+1)\right) \tag{4.85}$$

$$= e^{-\alpha \log K} K \log K + n\left(\alpha \log K + \log\left(\frac{1}{\alpha}+1\right)\right). \tag{4.86}$$

Using derivatives, we could solve for the value of α which makes the above the smallest. This gets a bit messy, so instead we ignore the the $\log\left(\frac{1}{\alpha}+1\right)$ and just optimize the other terms. The best value of α is at

$$e^{-\alpha \log K} = \frac{n}{K \log K} \tag{4.87}$$

$$\implies \alpha = \frac{\log \frac{K \log K}{n}}{\log K} \tag{4.88}$$

We need to satisfy the condition that $1 \leq 1/\alpha \leq e \log K$. This forces the constraints that $\log K \leq n \leq e^{-1/e} K \log K$.

This gives

$$r_n^+(\mathcal{I}_K^n) \leq \frac{n}{K \log K} K \log K + n \frac{\log \frac{K \log K}{n}}{\log K} \log K + n \log\left(\frac{\log K}{\log \frac{K \log K}{n}} + 1\right) \tag{4.89}$$

$$\leq n + n \log \frac{K}{n} + 2n \log(\log K + 1). \tag{4.90}$$

(we used that $\log \frac{K \log K}{n} \geq 1$). This gives us a bound which is the same order as the bound given in Table 4.2 for when $K = o(n)$. We will skip using the polynomial covering in Corollary 1 since it gives the same result as the subexponential bound (up to constants on the lower order terms).

Next, we will do the same for the divergence covering result using Hellinger (in Theorem 2) where

$$\log M(K, \varepsilon) = \log\left(K\left(\frac{C \log K}{\varepsilon}\right)^{\frac{K-1}{2}}\right) \tag{4.91}$$

(where the constant $C \leq 200$).

This gives

$$r_n^+(\mathcal{I}_K^n) \leq \frac{K-1}{2} \log\left(\frac{C \log K}{\varepsilon}\right) + \log K + n\varepsilon \tag{4.92}$$

Optimizing for ε gives $\varepsilon = \frac{K-1}{2n}$, which gives

$$r_n^+(\mathcal{I}_K^n) \leq \frac{K-1}{2} \log\left(\frac{C 2n \log K}{K-1}\right) + \log K + n\frac{K-1}{2n} \tag{4.93}$$

$$= \frac{K-1}{2} \log \frac{n}{K-1} + \frac{K-1}{2} (\log\log K + \log 2Ce) + \log K\,. \tag{4.94}$$

We technically need $\varepsilon \leq \log K$, so $n > \frac{K-1}{2 \log K}$.

When $K - 1 = n$, we see that the dominate term is $O(K \log\log K)$. This is not as tight as the result for when $K = \Theta(n)$ in Table 4.2. However, when $K = o(n)$, we do get the correct order of growth.

Lastly, we will apply our covering bound which uses χ^2 for the regime of small ε. This bound will be restricted only to the case when K is at least some polynomial factor smaller than n. The bound is

$$M(K, \varepsilon) \leq K^{3/2} \left(\frac{c}{\varepsilon}\right)^{\frac{K-1}{2}} \log \frac{1}{\varepsilon} \tag{4.95}$$

Instead of optimizing, we will choose $\varepsilon = \frac{K-1}{2n}$ but we need $\varepsilon \leq \frac{1}{4(K+1)^2}$. This implies that roughly $n > 4d^3$ for this bound to hold. We get

$$r_n^+(\mathcal{I}_K^n) \leq \frac{K-1}{2} \log\left(\frac{2cn}{K-1}\right) + \log\log \frac{2n}{K-1} + \frac{3}{2} \log K + n\frac{K-1}{2n} \tag{4.96}$$

$$\leq \frac{K-1}{2} \log\left(\frac{n}{K-1}\right) + \frac{K-1}{2} \log 2ce + \frac{3}{2} \log K + \log\log \frac{2n}{K-1} \tag{4.97}$$

$$\tag{4.98}$$

We will summarize these results in the following proposition:

Proposition 7 (Divergence Covering Upper Bounds on Regret)**.** *For iid distributions on alphabet of size K and n samples, the following all bound the minimax regret:*

- *For $\log K \leq n \leq e^{-1/e} K \log K$,*

$$r_n^+(I_K^n) \leq n \log \frac{K}{n} + n + 2n \log(\log K + 1)\,. \tag{4.99}$$

- *For $n > (K-1)/(2 \log K)$,*

$$r_n^+(I_K^n) \leq \frac{K-1}{2} \log \frac{n}{K-1} + \frac{K-1}{2} (\log\log K + \log 2Ce) + \log K \tag{4.100}$$

  where C is a constant.

- *For $n > 4(K+1)^3$,*

$$r_n^+(I_K^n) \leq \frac{K-1}{2} \log\left(c \frac{n}{K-1}\right) + \frac{3}{2} \log K + \log\log \frac{2n}{K-1} \tag{4.101}$$

  where c is a constant.

We summarize the same results in the following table:

Comparing to the regrets calculated in [25], it seems our upper bounds have an additional $O(n \log\log K)$ or $O(K \log\log K)$ term. For instance, in the case where $n = o(K)$, it seems the bound $n \log \frac{K}{n} + n$ is close to correct. The $O(n \log\log K)$ is the only additional term.

Regime	Regret
$n \le e^{-1/e} K \log K$	$r_n^+(I_K^n) \le n \log \frac{K}{n} + O(n \log\log K)$
$n > (K-1)/(2\log K)$	$r_n^+(I_K^n) \le \frac{K-1}{2} \log \frac{n}{K-1} + O(K \log\log K)$
$n > 4(K+1)^3$	$r_n^+(I_K^n) \le \frac{K-1}{2} \log \frac{n}{K-1} + O(K)$

This could be due to looseness in our divergence covering bounds. However, our bounds are still very close. We are correct up to the first order term, except in the case where $K = \Theta(n)$. Our results are also very clearly non-asymptotic. We did not need to use any approximations that rely on n being large.

Though we do not beat the minimax regret bounds calculated specifically for iid distributions, our technique for covering is still more versatile. With the same techniques, we can apply our bounds to other problems. The next section shows this for Markov models.

4.3.2 Markov Models

In this section we will look at how divergence covering can be applied to Markov models. (In [27], the author is also able to extend regret results for iid sources to Markov models; hence we believe that we should have a parallel connection.) For notation, we will use $\mathcal{I}_{K,r}^n$ to mean the class of probability distributions on x^n which are Markov models of order r over alphabet K.

For the first example, suppose we have an alphabet of size K and an order $r = 1$ Markov model. If sequence x^n is generated using probability P, this means that there is a $p_0(\cdot)$ and $p_w(\cdot|\cdot)$ so that

$$P_{X^n}(x^n) = p_0(x_1) \prod_{t=1}^{n-1} p_w(x_{t+1}|x_t)\,. \tag{4.102}$$

Similarly, for another distribution Q, we can alter the notation to have

$$Q_{X^n}(x^n) = q_0(x_1) \prod_{t=1}^{n-1} q_w(x_{t+1}|x_t)\,. \tag{4.103}$$

Let $p_t(x_t)$ be the marginal probability of x_t under P and similarly let $q_t(x_t)$ be the marginal probability of x_t under Q. If we only consider length n sequences x^n, we have

$$D(P_{X^n} \| Q_{X^n}) = \sum_{x^n} p_0(x_1) \prod_{t=1}^{n-1} p_w(x_{t+1}|x_t) \log \frac{p_0(x_1) \prod_{t=1}^{n-1} p_w(x_{t+1}|x_t)}{q_0(x_1) \prod_{t=1}^{n-1} q_w(x_{t+1}|x_t)} \tag{4.104}$$

$$= \sum_{x_1} p_0(x_1) \log \frac{p_0(x_1)}{q_0(x_1)} + \sum_{t}^{n-1} \sum_{x_t} p_t(x_t) \sum_{x_{t+1}} p_w(x_{t+1}|x_t) \log \frac{p_w(x_{t+1}|x_t)}{q_w(x_{t+1}|x_t)} \tag{4.105}$$

To cover all distributions P_{X^n}, one solution is to cover all possible $p_0(\cdot)$ and all possible $p_w(\cdot|x)$ for each $x \in [K]$. Each of these can be thought of as separate iid distribution on an alphabet of size K. Thus, one way to cover P is to take the product of the covers at radius ε for the distribution $p_0(\cdot)$ and each $p_w(\cdot|x)$. This gives the total number of covering points as

$$M = M(K, \varepsilon)^{K+1} \tag{4.106}$$

and the divergence will be bounded by $n\varepsilon$.

This can give a bound for a redundancy for a Markov model of order 1 on a fixed length sequence. We can also consider higher order Markov models. The analysis is the same for Markov models of order r, but we would need a covering on the simplex for each K^r possible previous sequences. To covering the first r starting symbols, we actually need $(K+1)^r$ different coverings. (The value is a slight overestimate. We need $K+1$ instead of K since we need the conditional distributions for the start of the sequence. One way of understanding this is to consider "positions" before the start of the sequence as a new $K+1$-th symbol.)

Like we did with iid distributions, we also want to make connections to minimax regret. Given an arbitrary sequence x^n, to calculate the regret, we need the best order 1 Markov model that bests fits x^n. This is given by setting

$$p_0^{(x^n)}(a) = \begin{cases} 1 & \text{if } x_1 = a \\ 0 & \text{else} \end{cases} \tag{4.107}$$

and

$$p_w^{(x^n)}(a|b) = \frac{|t : \{x_{t+1} = a \text{ and } x_t = b\}|}{|\{t : x_t = b, t < n\}|}. \tag{4.108}$$

We will also use

$$p^{(x^n)}(b) = \frac{|\{t : x_t = b, t < n\}|}{n}. \tag{4.109}$$

(For $r > 1$, we will need b to be r consecutive symbols.)

Then

$$r_n^+(Q, x^n, \mathcal{I}_{K,1}^n) = \max_P \log \frac{p_0^{(x^n)}(x_1) \prod_{t=1}^{n-1} p_w^{(x^n)}(x_{t+1}|x_t)}{q_0(x_1) \prod_{t=1}^{n-1} q_w(x_{t+1}|x_t)} \tag{4.110}$$

$$= \log \frac{p_0^{(x^n)}(x_1) \prod_{t=1}^{n-1} p_w^{(x^n)}(x_{t+1}|x_t)}{q_0(x_1) \prod_{t=1}^{n-1} q_w(x_{t+1}|x_t)} \tag{4.111}$$

$$= \log \frac{p_0^{(x^n)}(x_1)}{q_0(x_1)} + \log \frac{\prod_{a,b} p_w^{(x^n)}(a|b)^{np^{(x^n)}(b)p_w^{(x^n)}(a|b)}}{\prod_{a,b} q_w(a|b)^{np^{(x^n)}(b)p_w^{(x^n)}(a|b)}} \tag{4.112}$$

$$= \log \frac{p_0^{(x^n)}(x_1)}{q_0(x_1)} + \sum_{a,b} np^{(x^n)}(b) p_w^{(x^n)}(a|b) \log \frac{p_w^{(x^n)}(a|b)}{q_w(a|b)} \tag{4.113}$$

$$= \log \frac{p_0^{(x^n)}(x_1)}{q_0(x_1)} + n \sum_b p^{(x^n)}(b) \sum_a p_w^{(x^n)}(a|b) \log \frac{p_w^{(x^n)}(a|b)}{q_w(a|b)} \tag{4.114}$$

$$\tag{4.115}$$

In this formulation, the value of $p_0^{(x^n)}$ is always zero or one. We can let $\mathcal{Q}_0$ be the set of all possible p_0. Then $|\mathcal{Q}_0| = K$. Fix ε, for each b, we want a set $\mathcal{Q}(b)$ so that

$$\min_{q_w(\cdot|b) \in \mathcal{Q}(b)} D(p_w^*(\cdot|b) \| q_w^*(\cdot|b)) \leq \varepsilon \tag{4.116}$$

we need the set of $\{q_w^*(\cdot|b)\}$ to have size $M(K, \varepsilon)$. Then $|\mathcal{Q}(b)| = M(K, \varepsilon)$.

The overall distribution Q on Markov models can be chosen from $\mathcal{Q} = \mathcal{Q}_0 \times \mathcal{Q}(1) \times ... \times \mathcal{Q}(K)$, and thus $|\mathcal{Q}| = K \cdot M(K, \varepsilon)^K$. Let $\bar{Q}_{X^n}(x^n) = \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} Q_{X^n}(x^n)$.

Then let $q^*(a|b)$ be the distribution in $\mathcal{Q}(b)$ which is the arg min of (4.116) given $p^{(x^n)}$. Let q_0^* be equal to $p_0^{(x^n)}$ (take can take one of K values). Let $Q_{X^n}^*$ be the overall distribution created from $q^*(a|b)$ and q_0^*.

Then

$$r_n^+(\mathcal{I}_{K,1}^n) =\leq \max_{x^n} r_n(\bar{Q}_{X^n}, x^n, \mathcal{I}_{K,1}^n) \tag{4.117}$$

$$= \max_{x^n} \log \frac{p_0^{(x^n)}(x_1) \prod_{t=1}^{n-1} p_w^{(x^n)}(x_{t+1}|x_t)}{\bar{Q}_{X^n}(x^n)} \tag{4.118}$$

$$= \max_{x^n} \log \frac{p_0^{(x^n)}(x_1) \prod_{t=1}^{n-1} p_w^{(x^n)}(x_{t+1}|x_t)}{\frac{1}{|\mathcal{Q}|}\sum_{Q_{X^n} \in \mathcal{Q}} Q_{X^n}(x^n)} \tag{4.119}$$

$$= \log|\mathcal{Q}| + \max_{x^n} \log \frac{p_0^{(x^n)}(x_1) \prod_{t=1}^{n-1} p_w^{(x^n)}(x_{t+1}|x_t)}{Q^*_{X^n}(x^n)} \tag{4.120}$$

$$= \log|\mathcal{Q}| + \max_{x^n} \log \frac{p_0^{(x^n)}(x_1)}{q_0^*(x_1)} + n\sum_b p^{(x^n)}(b) \sum_a p_w^{(x^n)}(a|b) \log \frac{p_w^{(x^n)}(a|b)}{q_w^*(a|b)} \tag{4.121}$$

$$\leq \log|\mathcal{Q}| + n\varepsilon \tag{4.122}$$

$$= \log K + K \log M(K, \varepsilon) + n\varepsilon \tag{4.123}$$

The above is similar to Theorem 4 except specifically tailored to Markov models. We can generalize to other order Markov models as well, getting the result $\log K^r + K^r \log M(K, \varepsilon) + n\varepsilon$.

Proposition 8. *For order r Markov models on K symbols, the minimax regret on sequences of length n can be bounded by*

$$r_n^+(\mathcal{I}_{M,k}^n) \leq \log K^r + K^r \log M(K, \varepsilon) + n\varepsilon \tag{4.124}$$

for any ε.

This bound matches the result for Markov models to the first term given in [27] (as claimed by [28]). A more precise regret is given in [29].

4.3.3 Poisson Distributions

It is very straightforward to connect distributions on iid samples to Poisson distributions. Suppose we have two probabilities on K symbols, $p = (p_1, \ldots, p_K)$ and $q = (q_1, \ldots, q_K)$. For some n, let $\lambda_i = np_i$ and $\mu_i = nq_i$. Let $\mathcal{I}_K^{\text{poi}}$ be the class of Poisson Distributions over K symbols.

For this section, let $\boldsymbol{v} = (v_1, \ldots, v_K)$ be a type on j symbols (so v_i is the number of times symbol i appears). We use the notation $\text{poi}(\lambda, \boldsymbol{v})$ to mean to mean the probability of $\boldsymbol{v}$ under the poisson process with parameters $\lambda = (\lambda_1, \ldots, \lambda_K)$. Specifically,

$$\text{poi}(\lambda, \boldsymbol{v}) = \prod_{i=1} \frac{e^{-\lambda_i} \lambda_i^{v_i}}{v_i!} \tag{4.125}$$

We have that

$$D(\text{poi}(\lambda, \cdot) \| \text{poi}(\mu, \cdot)) = \sum_{\boldsymbol{v}} \text{poi}(\lambda, \boldsymbol{v}) \log \frac{\prod_i \frac{e^{-\lambda_i}\lambda_i^{v_i}}{v_i!}}{\prod_i \frac{e^{-\mu_i}\mu_i^{v_i}}{v_i!}} \tag{4.126}$$

$$= \sum_v \text{poi}(\lambda, \boldsymbol{v}) \left(\sum_i \mu_i - \lambda_i + v_i \log \frac{\lambda_i}{\mu_i} \right) \tag{4.127}$$

$$= \sum_i \lambda_i \log \frac{\lambda_i}{\mu_i} \tag{4.128}$$

$$= nD(p\|q) \tag{4.129}$$

Hence any divergence covering over the simplex is also a divergence covering for poisson distributions over types.

Suppose we are also interested in the regret of a type under all possible Poisson parameters. For a given μ, suppose we want

$$r_n(\mu, \boldsymbol{v}, \mathcal{I}_K^{\text{poi}}) = \max_{\lambda} \log \frac{\text{poi}(\lambda, \boldsymbol{v})}{\text{poi}(\mu, \boldsymbol{v})}\,. \tag{4.130}$$

Given the observation $\boldsymbol{v}$ the most likely value of λ is $\lambda = \boldsymbol{v}$. Suppose that $\boldsymbol{v}$ is a type over n symbols and p is such that $p_i = v_i/n$. We will also let $q_i = \mu_i/n$. Then,

$$\begin{aligned} r_n(\mu, \boldsymbol{v}, \mathcal{I}_K^{\text{poi}}) &= \max_{\lambda} \left(\sum_i \mu_i - \lambda_i + v_i \log \frac{\lambda_i}{\mu_i} \right) && (4.131)\\ &= \max_{\lambda} \sum_i v_i \log \frac{\lambda_i}{\mu_i} && (4.132)\\ &= \sum_i v_i \log \frac{v_i}{\mu_i} && (4.133)\\ &= n \sum_i p_i \log \frac{np_i}{nq_i} && (4.134)\\ &= nD(p\|q) && (4.135) \end{aligned}$$

So for any μ, both for redundancy and regret, the log-loss of poisson distribution over types is given by the KL divergence.

4.4 Regret on Patterns

In this section, we look at regret on the class of patterns, which we denote by $\mathcal{I}_\Psi^n$. We discuss what patterns are in Section 4.4.1. Our main results are that we can show the following using divergence covering on a particular subset of the simplex.

Theorem 5 (Regret on Patterns)**.**

$$r_n^+(\mathcal{I}_\Psi^n) \leq cn^{1/3} \log^{4/3}(n) \tag{4.136}$$

for some constant c.

Compared to the current known regret upper bound for patterns (see Section 4.4.1), our result improves the exponent on the logarithmic term.

Our procedure for finding this result is to first find a way to connect regret to a divergence covering, using the Yang-Barron technique (4.44). This is non-trivial since the class of patterns is very much not iid. (Our previous result connecting divergence covering to regret requires iid.) However, we manage this with help of the log-sum inequality. Once this connection is made, it determines a particular subset of the simplex which we need to cover. We then find a set of centers which covers it, and we get our bound.

4.4.1 Patterns Background

Because the per-symbol redundancy of the class of large-alphabet iid distributions does not go to zero, there is a line of work developed to study redundancy on patterns. Patterns are discussed in [25, 30]; we give a brief exposition here in this section. Essentially, a pattern only considers the number of times each symbol occurs, and not the actual label on the symbol. To describe what a pattern is, we first discuss what a shape is.

A *shape* of sequence of symbols is what happens when symbols in a sequence are relabeled based on when they first appear. The first symbol that occurs is labeled a 1. If the same symbol appears again, it is similarly labeled a 1. The second symbol that appears (which is different than the symbol associated with 1) is labeled a 2 and so on. This relabeling keeps all the new symbols in a length n sequence between 1 and

n, which nicely keeps the alphabet finite. However, even with this relabeling, regret for shapes is still linear in the length of the sequence, meaning it does not have diminishing per-symbol redundancy. Thus we need to apply an additional simplification in order to get vanishing per-symbol regret. This is why *patterns* are studied. A pattern is similar to a shape except that it is only the multiset of multiplicities of symbol counts which matter. Any one multiset of multiplicities is called a *profile*.

For instance, a profile could be $\varphi = \{3, 2, 2\}$. This means that two symbols appear twice and one symbols appears three times. Depending on the alphabet size, this profile corresponds with any type with these numbers of symbols appearing (we need at least alphabet size of 3). If for example the alphabet size is 4, the types which map to profile $\varphi = \{3, 2, 2\}$ are

$$(3, 2, 2, 0), (3, 2, 0, 2), (3, 0, 2, 2), \ldots, (0, 2, 2, 3)\,. \tag{4.137}$$

Each type corresponds with some (disjoint) set of sequences. For each sequence x^n, we can find its corresponding profile, denoted as $\varphi(x^n)$.

Let $P = (p_1, \ldots, p_K)$ be a discrete probability where p_i is the probability of symbol i occurring. For each P, we use P_{X^n} to be the n-fold product distribution on sequence x^n. Let the probability of profile φ occurring under P be denoted as P_φ where

$$P_\varphi(\varphi) = \sum_{x^n : \varphi(x^n) = \varphi} P_{x^n}(x^n)\,. \tag{4.138}$$

We let the alphabet size of P be equal to n, so that all possible patterns, including the one where $\varphi = \{1, 1, \ldots, 1\}$, has some P where it occurs with positive probability.

Finding the distribution over profiles can be very complicated. In order to find redundancy and regret on profiles, authors in [25, 30, 2] work with a Poissonized process on profiles. For distribution P, let $\lambda_i = np_i$. The poissonized version of P with average length n is given by parameters $\Lambda = (\lambda_1, \ldots, \lambda_n)$. For profile φ, let φ_μ be the number of elements in the multiset φ which is equal to μ. Then the probability a profile φ is generated is equal to

$$\Lambda(\varphi) = \frac{1}{\prod_{\mu=0}^{\infty} \varphi_\mu!} \sum_\sigma \prod_i^n \operatorname{poi}(\lambda_{\sigma(i)}, \mu_i)\,. \tag{4.139}$$

where we use $\Lambda(\varphi)$ to denote the probability of φ in this Poisson process with parameters given by Λ. This characterization is used to determine bounds for redundancy and regret of profiles. The class for profiles is defined as follows:

Definition 21. *Let the class of probabilities for patterns be*

$$\mathcal{I}_\Psi^n = \{P_\varphi(\cdot) : P \text{ is a discrete iid distribution } \}\,. \tag{4.140}$$

The tightest results in the literature for redundancy and regret on patterns, given in [2], are

$$0.3 \cdot n^{1/3} \le R_n^+(\mathcal{I}_\Psi^n) \le n^{1/3}(\log n)^{4/3} \tag{4.141}$$

$$\left(\frac{3}{2\log 2}\right) n^{1/3} \le r_n^+(\mathcal{I}_\Psi^n) \le n^{1/3}(\log n)^4\,. \tag{4.142}$$

Comparing our result Theorem 5 to the upper bound in (4.142), we see that we improve the exponent on the logarithm from 4 to 4/3, matching the exponent on the upper bound for redundancy of patterns.

4.4.2 Regret on Mixtures of Multinomials

We can see from (4.138) that the distribution P_φ is a mixture of iid distributions. We use this idea to show how divergence covering can be used for getting bounds on redundancy.

Because of (4.138), if P_1 and P_2 are iid probabilities which are equivalent to another up to a permutation of the symbols, then $(P_1)_\varphi$ and $(P_2)_\varphi$ are the exact same distribution on every profiles. Thus, to specify all the distributions on patterns of length n, it is enough to define the following class of probabilities:

$$\mathcal{P}_\searrow^n = \{ \text{ discrete probability } P = (p_1, \ldots, p_n) : p_1 \ge p_2 \ge \cdots \ge p_n \} \tag{4.143}$$

There is a bijective map between the class $\mathcal{I}_\Psi^n$ and $\mathcal{P}_\searrow^n$.

We can bound regret with the following:

$$r_n^+(\mathcal{I}_\Psi^n) = \inf_{Q_\varphi} \sup_{\varphi} \sup_{P_\varphi \in \mathcal{I}_\Psi^n} \log \frac{P_\varphi(\varphi)}{Q_\varphi(\varphi)} \tag{4.144}$$

$$= \inf_{Q_\varphi} \sup_{\varphi} \sup_{P_\varphi \in \mathcal{I}_\Psi^n} \frac{1}{P_\varphi(\varphi)} P_\varphi(\varphi) \log \frac{P_\varphi(\varphi)}{Q_\varphi(\varphi)} \tag{4.145}$$

$$\tag{4.146}$$

For an upperbound, we assume that $Q_\varphi(\varphi) = \sum_{x^n:\varphi(x^n)=\varphi} Q_{X^n}(x^n)$ for some $Q \in \mathcal{P}_\searrow^n$. We choose this so that we can apply log-sum inequality next.

$$r_n^+(\mathcal{I}_\Psi^n) \le \sup_{\varphi} \sup_{P \in \mathcal{P}_\searrow^n} \frac{1}{P_\varphi(\varphi)} \left(\sum_{x^n:\varphi(x^n)=\varphi} P_{X^n}(x^n) \right) \log \frac{\sum_{x^n:\varphi(x^n)=\varphi} P_{X^n}(x^n)}{\sum_{x^n:\varphi(x^n)=\varphi} Q_{X^n}(x^n)} \tag{4.147}$$

$$\le \sup_{\varphi} \sup_{P \in \mathcal{P}_\searrow^n} \frac{1}{P_\varphi(\varphi)} \sum_{x^n:\varphi(x^n)=\varphi} P_{X^n}(x^n) \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \tag{4.148}$$

$$\le \sup_{\varphi} \sup_{P \in \mathcal{P}_\searrow^n} \frac{1}{P_\varphi(\varphi)} \sum_{x^n:\varphi(x^n)=\varphi} P_{X^n}(x^n) \left(\sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \right) \tag{4.149}$$

$$= \sup_{P \in \mathcal{P}_\searrow^n} \sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \,. \tag{4.150}$$

Thus, finding $r_n^+(\mathcal{P}_\searrow^n)$ gives an upper bound to $r_n^+(I_\Psi^n)$.

Let the type of x^n be $\boldsymbol{v} = (v_1, \ldots, v_n)$. (So v_i represents the count of symbol i.) If $v_1 \ge v_2 \ge \cdots \ge v_n$, then we know the P which maximizes $P_{X^n}(x^n)$ is where $P = \boldsymbol{v}/n$. However, when $\boldsymbol{v}$ does not have this sorted form, we need to determine which $P \in \mathcal{P}_\searrow^n$ is the maximum likelihood estimator for $\boldsymbol{v}$.

Define

$$P_{ML}^\searrow(x^n) = P_{ML}^\searrow(\boldsymbol{v}) = \max_{P=(p_1,\ldots,p_n)\in\mathcal{P}_\searrow^n} \prod_{i=1}^n p_i^{v_i} \,. \tag{4.151}$$

Finding $P_{ML}^\searrow(x^n)$ is a well known problem. The solution is given by the Grenander estimator [31]. The Grenander estimator is typically used for the case of continuous densities on an interval. The question it answers is what density f, constrained so that $f(x) \ge f(y)$ if $x > y$, is the maximum likelihood estimator of the a given set of data points. The solution is that f is the the left derivative of the least concave majorant of the empirical cumulative distribution of the data points. This implies that f is a vector of local averages over a partition. We capture this fact when the data points are counts over $[K]$ in the following:

Lemma 8. *Fix $\boldsymbol{v}$ to be the type of x^n (e.g. $\sum_{i=1}^n v_i = n$). There exists a vector $\boldsymbol{u} = (u_1, \ldots, u_n)$ where*

- $u_1 \ge u_2 \ge ... \ge u_n$
- $\sum_{i=1}^n u_i = n$
- *Let $\mathcal{J}$ be all induces i where $u_i = u_j$ for some u_j. Then $u_i = \frac{\sum_{i\in\mathcal{J}} v_i}{|\mathcal{J}|}$.*

so that

$$P_{ML}^\searrow(\boldsymbol{v}) = P_{ML}^\searrow(\boldsymbol{u}) \,. \tag{4.152}$$

The lemma merely states the local average property just in terms of a vector $\boldsymbol{u}$ of decreasing counts that sum up to n. Because of the local average property we know that

$$u_i = \frac{z}{r} \tag{4.153}$$

for $z \in \mathbb{Z}_{\geq 0}$ and $r \in [n-1]$. This is because u_i is an average of integers and in fact it is an average of less than n integers. If it were an average of all n integers, then $u_i = 1$, so a denominator of n is not necessary.

We get (4.152) since

$$P_{ML}^{\searrow}(\boldsymbol{v}) = \prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{v_i} = \prod_{u}^{n} \left(\frac{u}{n}\right)^{\sum_{i:v_i=u} v_i} = \prod_{u}^{n} \left(\frac{u}{n}\right)^{\sum_{i:v_i=u} u} = \prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{u_i} = P_{ML}^{\searrow}(\boldsymbol{u})\,. \tag{4.154}$$

There are two important points we need from Lemma 8. First, is the form of $\boldsymbol{u}$. For any x^n, let $\boldsymbol{v}$ be the type of x^n. Let $\boldsymbol{u}$ be the vector given by Lemma 8 for $\boldsymbol{v}$.

$$\log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \leq \log \frac{P_{ML}^{\searrow}(\boldsymbol{v})}{Q_{X^n}(x^n)} \tag{4.155}$$

$$\leq \log \frac{P_{ML}^{\searrow}(\boldsymbol{u})}{Q_{X^n}(x^n)} \tag{4.156}$$

$$\leq \log \frac{\prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{u_i}}{Q_{X^n}(x^n)} \tag{4.157}$$

$$\tag{4.158}$$

Now suppose that there is a set $\mathcal{Q}$ which is a divergence covering of $\mathcal{P}_{\searrow}^{n}$ with radius ε. We will add an attentional constraint on $Q = (q_1, \ldots, q_n) \in \mathcal{Q}$ which covers $P = (p_1, \ldots, p_n)$: For indices where $p_i = p_{i+1}$, we require that $q_i = q_{i+1}$. Let $\bar{Q}$ be such that

$$\bar{Q}_{X^n}(x^n) = \frac{1}{|\mathcal{Q}|} \sum_{Q_{X^n} \in \mathcal{Q}} Q_{X^n}(x^n)\,. \tag{4.159}$$

Then,

$$\log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \leq \log \frac{\prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{u_i}}{Q_{X^n}(x^n)} \tag{4.160}$$

$$= \log \frac{\prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{u_i}}{\frac{1}{|\mathcal{Q}|} \sum_{Q_{X^n} \in \mathcal{Q}} Q_{X^n}(x^n)} \tag{4.161}$$

$$\leq \log |\mathcal{Q}| + \log \frac{\prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{u_i}}{\prod_{i=1}^{n} q_i^{v_i}} \tag{4.162}$$

$$= \log |\mathcal{Q}| + \log \frac{\prod_{i=1}^{n} \left(\frac{u_i}{n}\right)^{u_i}}{\prod_{i=1}^{n} q_i^{u_i}} \tag{4.163}$$

In the last equality, we chose q to cover $\boldsymbol{u}/n$. This means when $u_i = u_j$, we have that $q_i = q_j$. For the set of indices $\mathcal{J}$ where u_i are the same, we have that $\sum_{i \in \mathcal{J}} u_i = \sum_{i \in \mathcal{J}} v_i$. Thus, $\prod_{i \in \mathcal{J}} q_i^{v_i} = \prod_{i \in \mathcal{J}} q_i^{u_i}$.

$$\log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} = \log |\mathcal{Q}| + \sum_{i=1}^{n} u_i \log \frac{\left(\frac{u_i}{n}\right)}{q_i} \tag{4.164}$$

$$= \log |\mathcal{Q}| + n \sum_{i=1}^{n} \frac{u_i}{n} \log \frac{\left(\frac{u_i}{n}\right)}{q_i} \tag{4.165}$$

$$= \log |\mathcal{Q}| + n D(\boldsymbol{u}/n \| Q) \tag{4.166}$$

$$= \log |\mathcal{Q}| + n\varepsilon \tag{4.167}$$

and thus

$$r_n^+(I_\Psi^n) = \sup_{P \in \mathcal{P}_{\searrow}^{n}} \sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \tag{4.168}$$

$$\leq \sup_{P \in \mathcal{P}_{\searrow}^{n}} \sup_{x^n} (\log |\mathcal{Q}| + n\varepsilon) \tag{4.169}$$

$$\leq \log |\mathcal{Q}| + n\varepsilon \tag{4.170}$$

This gives a bound on regret for patterns. It remains to find a covering of the space $\mathcal{P}^n_{\searrow}$ with the additional constraint that if Q covers P, Q is equal on indices for each P has equal values on. The second fact we use from Lemma 8 is that the minimum non-zero value of u_i/n is some positive integer divided by $n(n-1)$. If p_i is not zero, then $p_i \geq 1/n^2$. Our covering only needs to look at P of this type. We formally define this class of distributions which is a subset of the class $\mathcal{P}^n_{\searrow}$ but with a lower bound on the minimum non-zero value.

Definition 22 (Sorted Class with Minimum)**.** *Let*

$$\mathcal{P}^n_{\searrow}\{L\} = \{\textit{discrete probability } P = (p_1, \ldots, p_n) : p_1 \geq p_2 \geq \cdots \geq p_n \textit{ and } \forall i, p_i \geq L \textit{ or } p_i = 0\} \quad (4.171)$$

4.4.3 Divergence Covering of Sorted Distributions with a Minimum

For the next lemma. we will look at covering the subset $\mathcal{P}^n_{\searrow}\{1/n^2\}$ in the simplex (recall Definition 4).

Lemma 9. *For the set of probabilities on sorted step distributions, we have that*

$$M\left(n, c_0 \frac{\log^{2/3} n}{n^{2/3}}, \mathcal{P}^n_{\searrow}\{1/n^2\}\right) = n^{c_1 n^{1/3} \log^{1/3} n} \quad (4.172)$$

for some absolute contants c_0 and c_1. This covering also satisfies the condition that if $P = (p_1, \ldots, p_n)$ is covered by some $Q = (q_1, \ldots, q_n)$ and $p_i = p_{i+1}$, then $q_i = q_{i+1}$.

This proof is based on some of the ideas of the proof in [32].

Proof. Let $\mathcal{Q}$ be a set of centers for divergence covering of radius $\varepsilon = \frac{n^{1/3}}{n} \log^{1/3} n$ on $\mathcal{P}^n_{\searrow}\{1/n^2\}$. We will determine the set of centers $\mathcal{Q}$ needed by first dividing all possible probabilities in $\mathcal{P}^n_{\searrow}\{1/n^2\}$ into groups. Let $\mathcal{G}$ be the set of all groups. We will assign one $q^G \in \mathcal{Q}$ to each $G \in \mathcal{G}$. We will show that $\forall p \in G$, the q^G assigned is such that

$$D(p||q^G) \leq 30 \frac{\log^{4/3} n}{n^{2/3}}. \quad (4.173)$$

Before defining groups, we will first define tiers. Each $p \in \mathcal{P}^n_{\searrow}\{1/n^2\}$ will be associated with a set of n tiers. Each symbol $x \in [n]$ will be assigned to one tier. The tiers are $T_1, \ldots, T_t$ and T_{zero}, for some integer t (defined very soon), The tier for x is determined by the value of $p(x)$.

$$x \in T_1 \text{ if } p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right), 1\right] \quad (4.174)$$

$$x \in T_2 \text{ if } p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^2, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)\right] \quad (4.175)$$

$$\vdots$$

$$x \in T_i \text{ if } p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right] \quad (4.176)$$

$$\vdots$$

$$x \in T_t \text{ if } p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^t, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{t-1}\right] \quad (4.177)$$

$$x \in T_{\text{zero}} \text{ if } p(x) = 0 \quad (4.178)$$

In order so that all x belong to some tier (recall either $p(x) > 1/n^2$ or $p(x) = 0$), it is sufficient that the last tier t is such that

$$\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^t \leq \frac{1}{n^2} \tag{4.179}$$

If we let (ignore integer constants, since it differs by a constant),

$$t = \frac{n^{1/3} \log(n^2)}{\log^{2/3} n} = 2n^{1/3} \log^{1/3} n \tag{4.180}$$

then

$$\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{\frac{n^{1/3} \log(n^2)}{\log^{2/3} n}} \leq (e^{-1})^{\log(n^2)} \tag{4.181}$$

$$\leq \frac{1}{n^2} \tag{4.182}$$

For each $p \in \mathcal{P}^n_{\searrow}\{1/n^2\}$, we define the *tier fingerprint* of p as $T(p) = (a_1, \ldots, a_t, a_{\text{zero}})$ where $a_j = |\{x : x \in T_j\}|$. Because p is sorted by value, knowing $T(p)$ specifies exactly which x belongs to each tier. If p_1 and p_2 have the same tier fingerprint, then they must have the same values of x in each teir (because $p_1(x)$ and $p_2(x)$ are sorted). We place p_1 and p_2 in the same group if $T(p_1) = T(p_2)$.

Fix some p. For each tier T_i, for $i \in \{[t] \cup \text{zero}\}$, define

$$\beta_i = \sum_{x \in T_i} p(x)\,. \tag{4.183}$$

Naturally $\sum \beta_i = 1$.

Let n_i be the number of symbols in T_i. Then for $i \in [t]$, we can bound

$$n_i \leq \frac{\beta_i}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i}\,. \tag{4.184}$$

This uses the fact that the smallest possible probability of $p(x)$ where $x \in T_i$ is given by $\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i$.

We now define the probability q^G for each group G, which we do by picking any $p \in G$ and defining

$$q^G(x) = \frac{1}{|T_j|} \sum_{y \in T_j} p(y) \text{ if } x \in T_j \tag{4.185}$$

for all $x \in G$. Since for every $p \in G$, the same symbols x are in the same tiers, if $x \in T_i$, then for all $p \in G$

$$p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right]\,. \tag{4.186}$$

and thus

$$q^G(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right]\,. \tag{4.187}$$

(So $p(x)$ and $q(x)$ should be pretty close.) For $x \in T_{\text{zero}}$, $q^G(x) = p(x) = 0$ for every $p \in G$. For every $p \in G$, if $p(x) = p(y)$, then x and y will be in the same tier. If x and y are in the same tier, then $q^G(x) = q^G(y)$.

For each q^G, we can compute the divergence distance to $p \in G$.

$$D(p||q) \leq \sum_x \frac{(p(x) - q^G(x))^2}{q^G(x)} \tag{4.188}$$

$$\leq \sum_{i\in[t]} n_i \frac{\left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i - \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right)^2}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i} \tag{4.189}$$

$$\leq \sum_{i\in[t]} \frac{\beta_i}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i} \frac{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{2i-2} \left(1 - \frac{\log^{2/3} n}{n^{1/3}} - 1\right)^2}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i} \tag{4.190}$$

$$= \sum_{i\in[t]} \beta_i \frac{\left(-\frac{\log^{2/3} n}{n^{1/3}}\right)^2}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^2} \tag{4.191}$$

$$= \left(\sum_{i\in[t]} \beta_i\right) \left(\frac{\left(-\frac{\log^{2/3} n}{n^{1/3}}\right)^2}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^2}\right) \tag{4.192}$$

$$= \frac{\log^{4/3} n}{n^{2/3}\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^2} \tag{4.193}$$

The quantity $1/\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^2$ has a maximum occurring at $n = 7$, so

$$\frac{1}{\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^2} \leq 29.1542\,. \tag{4.194}$$

This shows (4.173).

Next, we want to count the total number of groups, i.e., determine $|\mathcal{G}|$. We do this by counting the number of possible tier fingerprints. There are $t+1$ different tier levels. While it is true that tiers corresponding to larger values of $p(x)$ cannot possibly contain too many x's, for the purposes of getting an upper bound, we assume the values a_i in the tier fingerprint can be of any value between 0 and n. Then

$$|\mathcal{G}| = \binom{n+1+t+1}{t+1} \tag{4.195}$$

$$\leq \left(\frac{e(n+1+t+1)}{t+1}\right)^{t+1} \tag{4.196}$$

$$\leq \left(\frac{e(n+2n^{1/3}\log^{1/3} n + 2)}{2n^{1/3}\log^{1/3} n + 1}\right)^{2n^{1/3}\log^{1/3} n+1} \tag{4.197}$$

$$\leq (e2n)^{2n^{1/3}\log^{1/3} n+1} \tag{4.198}$$

$$\leq n^{8n^{1/3}\log^{1/3} n+1} \tag{4.199}$$

Finally, $|\mathcal{Q}| = |\mathcal{G}|$.

□

Our bound is given by

$$r_n^+(\mathcal{I}_\Psi^n) \leq \log|\mathcal{Q}| + \min_{Q\in\mathcal{Q}} \max_{P\in\mathcal{P}_\searrow^n\{1/n^2\}} nD(P||Q) \tag{4.200}$$

$$= \log\left(n^{8n^{1/3}\log^{1/3} n}\right) + n\cdot 30\frac{\log^{2/3} n}{n^{2/3}} \tag{4.201}$$

$$= 8n^{1/3}\log^{4/3}(n) + 30n^{1/3}\log^{2/3} n \tag{4.202}$$

$$\leq 38n^{1/3}\log^{4/3}(n)\,. \tag{4.203}$$

This show Theorem 5.

4.5 Some Discussion

We developed the framework for connecting divergence covering to minimax redundancy and minimax regret in this chapter. Where our framework is likely the most useful for finding new results is for problems where we need a divergence covering of some subset of the simplex. We did exactly this to get our result for regret on patterns. Even for computing minimax regret on discrete iid distributions, it is only necessary to cover all possible empirical distributions. We can perhaps improve our results by using this fact.

Chapter 5

Application to Permutation Channels

The noisy permutation channel is an information theory problem where divergence covering is needed in order to compute the capacity. We go over the motivation, definitions, results and proofs of the problem in this chapter. While divergence covering is the central argument, there are other results, unrelated to divergence covering, which also need to be proved. A particularly interesting one is a result on the KL divergence between drawing balls from an urn with replacement and without replacement. Another additional result included here is about divergence covering of a smaller dimensional linear subspace within the simplex.

Chapter Introduction We start the next section with the problem statement and a summary of the main results. We discuss the motivations for studying the noisy permutation channel in Section 5.1.2. In Section 5.2, we discuss how covering is used to determine the capacity of the noisy permutation channel. In Section 5.2.1, we give our result on covering a subspace of the probability simplex. We begin the proof of our main theorem in Section 5.3. Other theorems are proved afterwards.

5.1 Problem Statement and Main Results

The noisy permutation channel, as formally introduced in [3], is a communication model in which an n-letter input undergoes a concatenation of a discrete memoryless channel (DMC) and a uniform permutation of the n letters. Since the receiver observes a uniformly permuted output, the order of symbols conveys no information. More formally, the channel $P_{Y^n|X^n}$ can be described by the following Markov chain:

$$X^n \to Z^n \to Y^n \,. \tag{5.1}$$

Here the channel input X^n is a length n sequence where each position takes a value in $\mathcal{X} = [q]$ (where $[q] = \{1, 2, \ldots, q\}$). The sequence X^n goes through the DMC which operates independently and identically on each symbol. This results in a sequence Z^n where each position takes a value in $\mathcal{Y} = [K]$. The DMC transition probabilities can be represented as a $q \times K$ matrix $P_{Z|X}$. After the DMC, the sequence Z^n goes through the permutation part of the channel and results in sequence Y^n which is a uniformly random permutation of symbols on Z^n.

Let f_n and g_n be the channel encoder and decoder respectively. For each message $W \in [M]$, the input to the channel is $X^n = f_n(W)$. The output is Y^n, which the decoder decodes as $\hat{W} = g_n(Y^n)$. The probability of error is given by $P_{\mathsf{error}}^{(n)} \triangleq \mathbb{P}[W \neq \hat{W}]$. The rate for the encoder-decoder pair (f_n, g_n) is defined as

$$R \triangleq \frac{\log M}{\log n} \,. \tag{5.2}$$

A rate R is *achievable* if there is a sequence of encoder-decoder pairs (f_n, g_n) with rate R such that $\lim_{n\to\infty} P_{\mathsf{error}}^{(n)} = 0$. The capacity for the noisy permutation channel with DMC $P_{Z|X}$ is $C_{\mathsf{perm}}(P_{Z|X}) \triangleq \sup\{R \geq 0 : R \text{ is achievable}\}$.

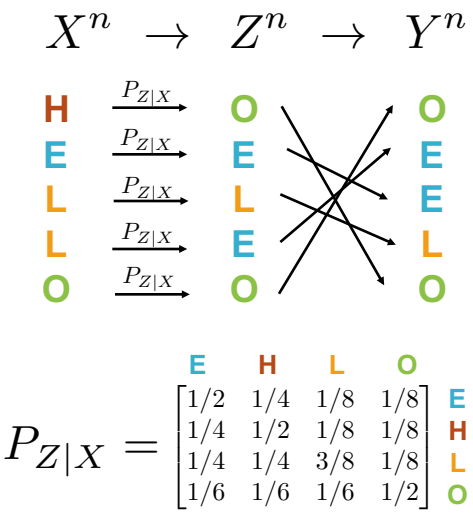


Figure 5.1: Illustration of the noisy permutation channel procedure. The vertical line of colored letters represents a sequence of $n = 5$ symbols. A iid noisy process, represented by DMC matrix $P_{Z|X}$, is applied to transform X^n to Z^n. Then, the sequence of symbols in Z^n is permuted uniformly to get Y^n.

In [3], the author determined that the noisy permutation channel capacity[1] for DMC $P_{Z|X}$ is bounded by

$$C_{\mathsf{perm}}(P_{Z|X}) \geq \frac{\mathsf{rank}(P_{Z|X}) - 1}{2}. \tag{5.3}$$

For strictly positive matrices $P_{Z|X}$ (meaning all the transition probabilities are greater than 0), the author shows a converse bound

$$C_{\mathsf{perm}}(P_{Z|X}) \leq \frac{|\mathcal{Y}| - 1}{2}. \tag{5.4}$$

The author also gives a second converse bound: $C_{\mathsf{perm}}(P_{Z|X}) \leq (\mathsf{ext}(P_{Z|X}) - 1)/2$, where $\mathsf{ext}(P)$ is the number of extreme points of the convex hull of the rows of P. For the case of strictly positive DMC $P_{Z|X}$, these upper and lower bounds do not necessarily match if the rank of matrix $P_{Z|X}$ does not equal to $|\mathcal{Y}|$ or $\mathsf{ext}(P_{Z|X})$.

5.1.1 Main Results

Our main result is establishing tightness of the lower bound (5.3), resolving Conjecture 1 of [3].

Theorem 6 (Strictly Positive DMC)**.** *For strictly positive $P_{Z|X}$,*

$$C_{\mathit{perm}}(P_{Z|X}) = \frac{\mathit{rank}(P_{Z|X}) - 1}{2}. \tag{5.5}$$

In order to reduce to the covering question, we first need another result that is, perhaps, of separate interest as well. We discuss some notation for this result first.

We let $\mathcal{P}_n$ be the set of n-types (probabilities which can be written with denominator n). For $P \in \mathcal{P}_n$, let $T_n(P)$ be the set of sequences of length n in the type class of P. For instance, when working with 3 symbols, the sequence 1233 is in $T_4(P)$ where $P = (1/4, 1/4, 1/2)$ and P is a 4-type. The notation Q_Y means a distribution on random variable Y. For any distribution Q_Y, we will use Q_Y^n to mean the product distribution $Q_Y^n(y^n) = \prod_{t=1}^n Q_Y(y_t)$.

[1] While it might seem that the noisy permutation channel capacity should be a continuous function of the values in $P_{Z|X}$, note that this is not the case due to how capacity is defined. Changing values in $P_{Z|X}$ by a small δ could change the rank of $P_{Z|X}$ by 1, but no matter how small δ is, there exists an n large enough so its effects can make a difference.

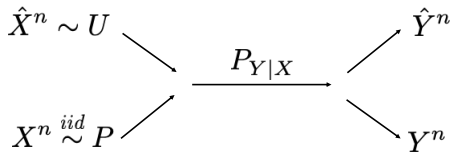


Figure 5.2: This diagram illustrates the special case of Theorem 7 where $Q_Y = P_Y$. It shows how $\hat{Y}^n$ relates to Y^n.

Theorem 7. *Fix channel $P_{Y|X}$ which is strictly positive. Then there exists a constant $c = c(P_{Y|X})$ such that the following holds: For any n-type P, let U be uniform on $T_n(P)$. For all Q_Y we have*

$$nD(P_Y\|Q_Y) \le D(P_{Y|X}^n \circ U\|Q_Y^n) \le nD(P_Y\|Q_Y) + c \tag{5.6}$$

where P_Y is the marginal distribution of Y under $(P \times P_{Y|X})$.

Remark 4. *It can be shown that the constant c in Theorem 7 is*

$$c \le \frac{q-1}{2}\log\frac{2\pi\alpha^2}{c_*} + \frac{q}{12n} \le \frac{q-1}{2}\log\frac{2\pi\alpha^2}{c_*} + \frac{q}{12}\,. \tag{5.7}$$

where α is a universal constant defined in Theorem 9 (see Section 5.3.3) and if p_{ij} denote the values in matrix $P_{Y|X}$,

$$c_* = \min_i \frac{\min_j p_{ij}}{\max_j p_{ij}}\,. \tag{5.8}$$

Theorem 7 deals with the following scenario: Select some $P \in \mathcal{P}_n$ and suppose we have two sequences, X^n and $\hat{X}^n$. The sequence X^n is generated iid using the probability P. On the other hand, $\hat{X}^n$ has uniform probability over all sequences in the type $T_n(P)$. Both sequences X^n and $\hat{X}^n$ undergo the transition $P_{Y|X}$ applied independently on each symbol and respectively results in Y^n and $\hat{Y}^n$. How different are the distributions of Y^n and $\hat{Y}^n$ under KL divergence? See Figure 5.2 for a diagram. Another interpretation of this scenario is if there are n balls of q colors in an urn. The sequence X^n are n draws from the urn with replacement and $\hat{X}^n$ are n draws without replacement (all the balls are drawn). These observations then both go through the same noisy process to produce Y^n and $\hat{Y}^n$.

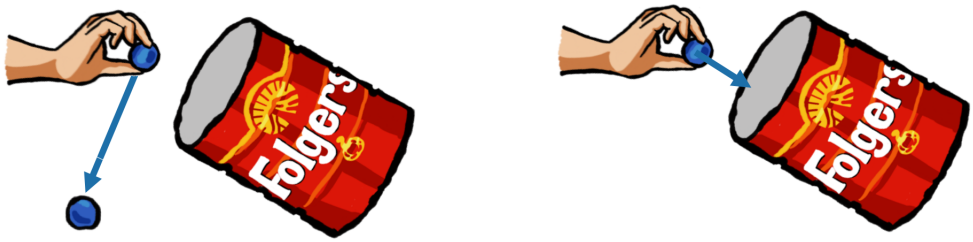
It turns out that if $P_{Y|X}$ is strictly positive, then regardless of the sequence length n,

$$D(P_{\hat{Y}^n}\|P_{Y^n}) \le c \tag{5.9}$$

where c is a constant that only depends on $P_{Y|X}$. Theorem 7 actually shows something more general. The sequence X^n can be generated iid with another distribution Q, and the KL divergence can still be bounded by constant c plus another term which is the KL divergence of the marginals on Y generated by P and Q. In other words, the divergence of (a complicated distribution) $P_{\hat{Y}^n}$ to any iid distribution Q_Y^n can be approximated with $nD(P_Y\|Q_Y)$ and this approximation will only be off by a constant.

Remark 5. *Note that $D(P_{\hat{X}^m}\|P_X^m)$ describes the difference between sampling m balls from an n-urn with and without replacement. This is a classical question studied in [33]. Our setting studies this question for the particular case when $n = m$ and when the observations are noisy. Bounds for the noiseless case $D(P_{\hat{X}^m}\|P_X^m)$ can still be an upper bound for the noisy case if we apply the data processing inequality. This shows that $D(P_{Y|X}^n \circ U\|P_Y^n) \le D(P_{\hat{X}^n}\|P_X^n) \le \frac{K-1}{2}(\log n + c)$, where the second inequality is shown using Stirling's approximation. Our result removes the $\log n$ term in this bound, but only under the assumption of a strictly positive $P_{Y|X}$. See Section 5.3.5 for more details on comparing our bound to that of [33] when $m < n$. We also note that results of [33] as shown in [34] imply the finitary case of de Finetti's theorem.*

Other contributions of this work use similar techniques to get converse results in other settings which do not have strictly positive DMC matrices.



$P_{Y|X}^n \circ U$ $\qquad$ P_Y^n

Figure 5.3: Illustration of drawing balls from an urn with and without replacement.

Theorem 8. *Other channel results:*

1. *Suppose $P_{Z|X}$ can be written as a block diagonal matrix with β blocks where each block is strictly positive. Then,*

$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{\mathsf{rank}(P_{Z|X}) + \beta - 2}{2}. \tag{5.10}$$

2. *For DMC $P_{Z|X}$ which is a q-ary erasure channel for $q \geq 2$ (assuming non-zero transition probabilities), then*

$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{q-1}{2}. \tag{5.11}$$

3. *For DMC $P_{Z|X}$ which is a Z-channel (assuming non-zero probabilities on the edges), then*

$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{1}{2}. \tag{5.12}$$

The first result in Theorem 8 applies to DMC $P_{Z|X}$ which are block diagonal matrices where each block is strictly positive. As eq. (5.10) of Theorem 8 implies, we are able to show both the achievability and converse results for block diagonal DMC matrices. We will prove both these results in Section 5.4.

The second result is for binary erasure channels and q-ary erasure channels. The work in [3] determines the capacity for when the DMC matrix is the binary symmetric channel, but leaves the binary and q-ary erasure channels as open problems. Item 2 of Theorem 8 resolves Conjecture 2 presented in [3] regarding the capacity of binary erasure channels and the conjecture regarding q-ary erasure channels[2]. These results use (5.3) as the tight achievability.

The third result in Theorem 8 deals with DMC which is the Z-channel. While this is tight, for a q-ary Z-channel (if this exists) or what we call a "zigzag" channel, we are not able to find tight results with our current covering technique. The erasure channels, Z-channels, and a brief analysis on the zigzag channel are discussed in Section 5.5

All of these results also use the method of covering. Using covering as a technique to determine the capacity for the noisy permutation channel is reasonable because the covering centers can intuitively be equated with messages that can be distinguishable. When the messages correspond to two distributions q_1 and q_2 which are far in KL divergence, it is unlikely that noisy versions of q_1 will be close to noisy versions

[2]Note that while binary erasure channels and q-ary erasure channels usually have the same probabilities for each symbol getting erased, item 2 of Theorem 8 is still true even if these probabilities are different. The only requirement is that the probability of getting erased is not 0 or 1. The capacity when the transition probabilities are 0 or 1 is discussed in [3].

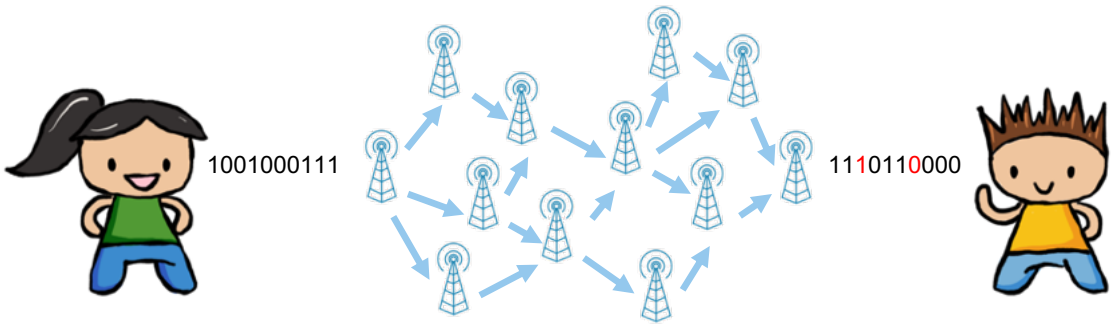


Figure 5.4: Illustration of communication over a multipath routed network. The sender transmits bits which gets shuffled at the receiver end due to the latency of different paths in the network. The received bits can also be noisy.

of q_2. If two distributions are close in KL divergence, their noisy versions are likely to be confused. For a converse bound, covering is useful because we know that no more center points can be added which are far from all existing center points. This roughly corresponds to not being able to add another decodable message.

5.1.2 Motivation

The motivation for studying the permutation channel is that it captures a setting where codewords get reordered. This occurs in applications such as communication networks and biological storage systems. We briefly describe some of these applications. More details on these applications and other relevant work can be found in [3].

Communication Networks Suppose we have a point-to-point communication network where the information is transmitted through a multipath routed network. Different packets are transmitted through different routes in the network, and each route has its own amount of latency, causing packets traveling on different routes to arrive at different times. The order in which the sender transmits packets is no longer preserved at the receiver end. Such a scenario is studied in [35] where the authors are primarily concerned with reducing delay in their channel. Unlike our work, they do not consider noisy symbols. Another line of work on packet-switched networks deals with the permutation channel along with errors such as insertions, deletions, and substitutions of symbols [36, 37]. Their work primarily focuses on building minimum distance codes and perfect codes for the permutation channel.

DNA Storage Systems DNA-based storage systems are an attractive option for data storage due to its ability to withstand time and encode a very high-density of information [38, 39]. The state-of-the-art technology for storing information on DNA uses nucleotides with relatively small lengths (few hundreds) [40]. Each of these DNA molecules are stored in a pool without any regard to order. The different molecule types can be treated as symbols in the setting of the permutation channel. Noise in this channel models any error that can occur, whether it is in synthesizing the DNA molecules or in reading the molecules. DNA storage is also the motivation for studying the permutation channel in [41, 42].

As typical in information theory, a question of fundamental interest is to determine the capacity of channels. We determine the capacity of the noisy permutation channel in the strictly positive case, settling the problem introduced in [3]. This setting differs from some of the models studied in the works described in the motivations, as it looks at the problem from a purely information theoretic standpoint and does not include assumptions which might be specific to the application.

Among the works relevant to the motivations described, those that have some information theoretical flavors include [41], which deals with asymptotic bounds on rate, but for a fixed number of errors rather

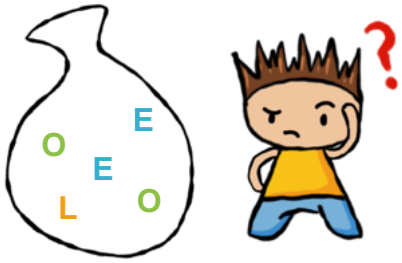


Figure 5.5: Order of received symbols Y^n do not matter since all the permutations are uniform.

than probabilitistic errors. The work in [40] finds the capacity when the symbols are sampled randomly then read, something relevant to DNA models, but not to general permutation channels. The results in [42] are specifically for when the permuted objects are a string of symbols and the noisy process is applied to symbols on a string; the set of strings are permuted but symbols in each string are not.

Our results for when the DMC is the erasure channel are particularly interesting to DNA storage applications since the erased symbol can model deletion errors. Permutation channels with deletions are central to the work in [41]. In our work, we find the probabilistic analogue of their bounds.

5.1.3 Notation

The set of all probability distributions on q symbols is defined as the probability simplex

$$\Delta_{q-1} \triangleq \left\{ (\pi_1, ..., \pi_q) : \sum_{i=1}^{q} \pi_i = 1, 0 \leq \pi_i \leq 1 \right\}. \tag{5.13}$$

For a $q \times K$ DMC matrix $P_{Z|X}$, we can express the individual transitions as

$$P_{Z|X} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1d} \\ p_{21} & p_{22} & \cdots & p_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ p_{q1} & p_{q2} & \cdots & p_{qd} \end{bmatrix}. \tag{5.14}$$

We assume that the values in each row of the matrix sums up to 1 (i.e, the matrix is stochastic). Symbol $i \in \mathcal{X}$ has probability p_{ij} of becoming symbol $j \in \mathcal{Y}$. We can also write this probability as $P_{Z|X}(j|i)$. We say that the DMC matrix (or a submatrix) is *strictly positive* if $p_{ij} > 0$ for all i and j in the matrix (or submatrix).

Though different from how we described it in the introduction, it is convenient to describe the Markov chain of the noisy permutation channel as

$$\pi \to X^n \to Z^n \to Y^n. \tag{5.15}$$

Each $\pi = (\pi_1, .., \pi_q) \in \Delta_{q-1}$ corresponds to a possible channel input. For each n, we will restrict π to be in $\mathcal{P}_n$. The value of π_i represents the proportion of positions in sequence X^n which have symbol i. We can also understand π_i to be the proportion of the n symbols which undergo the transition given in row i of $P_{Z|X}$.

The order of the symbols in Y^n does not matter for decoding the messages. The only relevant statistic the receiver would use from Y^n is which type class Y^n belongs to. Note that it is entirely equivalent to perform the permutation on the sequence X^n first and then apply the DMC. In this case, we no longer need the random variable Z^n. Because of this, we will also use $P_{Y|X}$ to specify the transition matrix, where $P_{Y|X} = P_{Z|X}$.

Since we will need to specify multiple distributions on Y, we will distinguish them using the parameter π. We use the notation $Q_{Y|\mu}$ for $\mu = (\mu_1, ..., \mu_k) \in \Delta_{K-1}$ to mean

$$Q_{Y|\mu}(j) = \mu_j. \tag{5.16}$$

The distribution $P_{Y^n|\pi}$ refers the the distribution on sequences Y^n when $\pi \in \mathcal{P}_n$ is the input to the noisy permutation channel as described in (5.15). However, we also use the following notation

$$P_{Y^n|\pi} = P_{Y|X}^n \circ U \tag{5.17}$$

where U is a uniform distribution on $T_n(\pi)$, as seen in Theorem 7. Equation (5.17) can be understood as permuting the input symbols, which gives a sequence in the support of U, and then passing each permuted symbol through the transition probabilities $P_{Y|X}$ independently.

When it is clear what π is, we use P_Y to mean the marginal distribution for each Y_t in the sequence $Y^n \sim P_{Y|X}^n \circ U$. This distribution does not depend on the index t since U is uniform on all permutations.

Define for any $\pi \in \Delta_{q-1}$

$$\mu^M(\pi) \stackrel{\triangle}{=} \left(\sum_i \pi_i p_{i1}, ..., \sum_i \pi_i p_{ik} \right) . \tag{5.18}$$

The vector $\mu^M(\pi)$ is the mean of the distribution $P_{Y^n|\pi} = P_{Y|X}^n \circ U$. Note that $P_Y(j) = \sum_i \pi_i p_{ij}$. Also if $\mu = \mu^M(\pi)$, we can write $P_Y = Q_{Y|\mu}$.

5.2 Covering Converse

The main concept of our proof uses covering ideas similar to [6, Theorem 1] in order to upper bound the mutual information $I(\pi; Y^n)$. In summary, we need a find a set of covering centers which are close in KL divergence to all the possible distributions on Y^n that can occur as outputs of the noisy permutation channel. Our set of centers need not be possible distributions over Y^n generated by the channel. We will opt for using iid distributions as our set of covering centers.

Let $\mathcal{N}_n$ be a discrete set in Δ_{K-1} which we will specify (later) for each n (these will be the covering centers). Mutual information has the property that

$$I(\pi; Y^n) \leq \max_{\pi} D(P_{Y^n|\pi} \| \tilde{Q}_{Y^n}) . \tag{5.19}$$

The above holds for any $\tilde{Q}_{Y^n}$, thus we can choose

$$\tilde{Q}_{Y^n}(y^n) = \sum_{\mu \in \mathcal{N}_n} Q_{Y|\mu}^n(y^n) = \sum_{\mu \in \mathcal{N}_n} \prod_{t=1}^n Q_{Y|\mu}(y_t) . \tag{5.20}$$

The following proposition the main work-horse of all our converse results.

Proposition 9 (Covering for Noisy Permutation Channels)**.** *Suppose that for the noisy permutation channel with DMC $P_{Y|X}$, we have that for any $\pi \in \mathcal{P}_n$,*

$$D(P_{Y|X}^n \circ U \| Q_Y^n) \leq c + nD(P_Y \| Q_Y) + f(n) \tag{5.21}$$

where U is uniform on the type $T_n(\pi)$, P_Y is the marginal distribution of $P_{Y|X}^n \circ U$, c is a constant (only a function of $P_{Y|X}$) and f is only a function of n (and possibly $P_{Y|X}$) . Then

$$C_{\mathsf{perm}}(P_{Y|X}) \leq \frac{\mathsf{rank}(P_{Y|X}) - 1}{2} + \lim_{n \to \infty} \frac{f(n)}{\log n} . \tag{5.22}$$

Proof. Following techniques used in the proof of Theorem 1 from [6], we can upper bound the mutual information given in (5.19) by

$$I(\pi; Y^n) \leq \log |\mathcal{N}_n| + \max_{\pi \in \mathcal{P}_n} \min_{\bar{\mu} \in \mathcal{N}_n} D(P_{Y^n|\pi} \| Q_{Y|\bar{\mu}}^n) . \tag{5.23}$$

To specify $\mathcal{N}_n$, first define

$$\mathcal{L}(P_{Y|X}) = \bigcup_{\pi \in \Delta_{K-1}} \mu^M(\pi)\,. \tag{5.24}$$

This is the space of all possible marginals P_Y.

Let $\mathcal{N}_n$ be a covering of $\mathcal{L}(P_{Y|X})$ under KL divergence with covering radius $1/n$. In other words, $\mathcal{N}_n = \{\bar{\mu}^{(1)}, ..., \bar{\mu}^{(m)}\}$ so that

$$\max_{\mu \in \mathcal{L}(P_{Y|X})} \min_{\bar{\mu} \in \mathcal{N}_n} D(Q_{Y|\mu} \| Q_{Y|\bar{\mu}}) \leq \frac{1}{n}\,. \tag{5.25}$$

Let ℓ be the dimension of $\mathcal{L}(P_{Z|X})$. Using Theorem 1 and the result specifically about covering an ℓ-dimensional subspace (see next part Section 5.2.1),

$$|\mathcal{N}_n| \leq C(q, \ell) \left(\frac{\ell}{1/n}\right)^{\frac{\ell}{2}} \tag{5.26}$$

where $C(q, \ell)$ depends on q and ℓ but not n.

Using assumption (5.21) and putting this into (5.23), gives

$$\begin{aligned}
I(\pi; Y^n) &\leq \log |\mathcal{N}_n| + \max_{\pi \in \mathcal{P}_n} \min_{\bar{\mu} \in \mathcal{N}_n} D(P_{Y^n|\pi} \| Q^n_{Y|\bar{\mu}}) && (5.27)\\
&\leq \log \left(C(q, \ell) \left(\frac{\ell}{1/n}\right)^{\frac{\ell}{2}} \right) + c + f(n) + \max_{\pi \in \mathcal{P}_n} \min_{\bar{\mu} \in \mathcal{N}_n} nD(P_Y \| Q_{Y|\bar{\mu}}) && (5.28)\\
&\leq \frac{\ell}{2} \log n + \log C(q, \ell) + \frac{\ell}{2} \log \ell + c + f(n) + n\frac{1}{n} && (5.29)\\
&\leq \frac{\ell}{2} \log n + c' + f(n) && (5.30)
\end{aligned}$$

where c' is a constant which does not depend on n.

For the noisy permutation channel, recall that the rate is defined as (5.2). Since asymptotically $\log M \leq I(\pi, Y^n) \leq \frac{\ell}{2} \log n + c' + f(n)$, we have

$$R \leq \frac{\ell}{2} + \frac{c'}{\log n} + \frac{f(n)}{\log n} \to \frac{\ell}{2} + \lim_{n \to \infty} \frac{f(n)}{\log n}\,. \tag{5.31}$$

Since $\ell = \mathsf{rank}(P_{Z|X}) - 1$, we have an upper bound for the capacity of the noisy permutation channel.

(To see that $\ell = \mathsf{rank}(P_{Z|X}) - 1$, let $r = \mathsf{rank}(P_{Z|X})$. When the domain is any vector, the image space of $P_{Z|X}$ is r dimensional. But since we are restricting the domain and image to probability vectors, this adds an additional constraint to the image space and reduces the dimension by 1.)

□

5.2.1 Subspace Covering

To use our covering result for noisy permutation channels, we actually need to cover a lower dimensional subspace of a $(K-1)$-dimensional simplex. We will use Definition 4 (see Chapter 1) for this.

Proposition 10. *For linear space $\mathcal{B} \subset \Delta_{K-1}$, suppose there is a stochastic matrix F which maps Δ_{q-1} onto $\mathcal{B}$. Suppose that $\mathcal{B}$ is a space of dimension $\ell - 1$ (or likewise, F has rank ℓ). Then,*

$$M(K, \varepsilon, \mathcal{B}) \leq \binom{K}{\ell} M(\ell, \varepsilon)\,. \tag{5.32}$$

First, we show the following lemma.

Lemma 10. *For $\mathcal{B} \subset \Delta_{K-1}$, suppose there is a stochastic matrix F which maps $\Delta_{\ell-1}$ onto $\mathcal{B}$. Then,*

$$M(K, \varepsilon, \mathcal{B}) \leq M(\ell, \varepsilon)\,. \tag{5.33}$$

Proof. Let $\mathcal{N}_c(\ell, \varepsilon)$ be the set of points which are centers for a divergence covering of $\Delta_{\ell-1}$ with covering radius ε. For each $b \in \mathcal{B}$, there exists a $p \in \Delta_{\ell-1}$ such that $pF = b$. For this p, let $r \in \mathcal{N}_c(\ell, \varepsilon)$ be such that $D(p||r) \leq \varepsilon$. Let $b^* = rF$. By data processing inequality [43, Theorem 2.2],

$$D(b||b^*) \leq D(p||r) \leq \varepsilon\,. \tag{5.34}$$

Hence the image of the set of centers in $\mathcal{N}_c(\ell, \varepsilon)$ mapped using F, becomes the set of centers for a divergence covering on $\mathcal{B}$ with radius ε.

□

Proof of Proposition 10. The key to this proof is to divide the space $\mathcal{B}$ into simplices of dimension $\ell - 1$.

We will upper bound the number of simplices needed for a partition of $\mathcal{B}$. The image of F is a convex hull of at most q points (recall q is the size of the input symbols). We will call these corner points. Consider all possible choices of ℓ of these q corner points. Let this set of all combinations be S, where

$$|S| = \binom{q}{\ell}\,. \tag{5.35}$$

For each $s \in S$, let $\mathcal{B}_s$ be the simplex which is the convex hull of the ℓ corner points in set s.

For each point x in the image of F, since F has rank ℓ, there exists some linear combination of ℓ corner points which results in x. If s is this set of ℓ points, then $x \in \mathcal{B}_s$. Thus for all $x \in \mathcal{B}$, there exists some $s \in S$, so that $x \in \mathcal{B}_s$.

Label each of these simplices as $\mathcal{B}_1, ..., \mathcal{B}_{|S|}$. There exists a stochastic matrix F_i which maps from space $\Delta_{\ell-1}$ onto the space $\mathcal{B}_i$. In particular, we can find this map F_i by mapping each of the ℓ corners of $\Delta_{\ell-1}$ into one of the ℓ corner points of $\mathcal{B}_i$. This map will cover all of $\mathcal{B}_i$ by linearity.

Hence using Lemma 10, we can find a divergence covering of size $M(\ell, \varepsilon)$ for each $\mathcal{B}_i$. Combining these covering centers together for all i, we get a covering of size

$$\binom{q}{\ell} M(\ell, \varepsilon)\,. \tag{5.36}$$

□

We are most assuredly over counting the number of simplices $\mathcal{B}$ has to be divided up into. However, this number does not depend on ε, which is sufficient for our application to noisy permutation channels.

5.3 Divergence Under Fixed Types

For computing our converse bounds, we need to determine the expression (5.21) for our DMC matrices. This is where we need Theorem 7 which gives the divergence between transition on a fixed type compared to an iid distribution.

We will prove Theorem 7 by showing some relevant intermediate results. The techniques in these intermediate results are also useful for when $P_{Z|X}$ is not strictly positive and therefore relevant for getting results in Theorem 8. Before doing so, we will briefly discuss the tightness of Theorem 7.

5.3.1 Tightness of Theorem 7

Here we show a proof sketch that the constant c in Theorem 7 is sharp (cannot be improved to $o(1)$). One tool we will need is the following theorem by Marton [44] shown below.

Lemma 11 (Marton's Transportation Inequality)**.** *Let $X^n \sim \prod_{t=1}^n P_{X_t}$ and $\hat{X}^n \sim P_{\hat{X}^n}$. Then there exists a joint probability measure $P_{X^n,\hat{X}^n}$ with these given marginals such that*

$$\frac{1}{n}\mathbb{E}[d(X^n,\hat{X}^n)] = \frac{1}{n}\sum_{t=1}^n \mathbb{P}[X_t \neq \hat{X}_t] \leq \left(\frac{1}{n} D\left(P_{\hat{X}^n} \middle\| \prod_{t=1}^n P_{X_t}\right)\right)^{1/2} \tag{5.37}$$

where $d(X,Y)$ is the Hamming distance.

Suppose that we are only working with 2 symbols, $\{0,1\}$, for the space of X and Y. Using the notation in Theorem 7, let $Y^n \sim P_{Y|X}^n \circ U$ and $\hat{Y}^n \sim Q_Y^n$. Choose $P = (n/2, n/2)$ and $P_{Y|X}$ to be that of a binary symmetric channel with crossover probability ε. Choose $Q_Y = P_Y$. Distribution Q_Y will be uniform on the two symbols.

We can choose ε to be non-zero but also small enough (for instance $\varepsilon = 1/n$) so that all sequences likely to occur under $P_{Y|X}^n \circ U$ have a type differing by only a constant number of changes from the type $(n/2, n/2)$. On the other hand, under Q_Y^n, with constant probability, we expect sequences to have types that differ by $\sqrt{n}$ changes from the type $(n/2, n/2)$. Hence no matter the coupling chosen, the expected Hamming distance is

$$\mathbb{E}[d(Y^n,\hat{Y}^n)] \approx \sqrt{n}\,. \tag{5.38}$$

We use Lemma 11 to get the lower bound,

$$\frac{1}{\sqrt{n}} \approx \frac{1}{n}\mathbb{E}[d(Y^n,\hat{Y}^n)] \leq \left(\frac{1}{n} D(P_{Y|X}^n \circ U \| Q_Y^n)\right)^{1/2} \leq \left(\frac{1}{n}(nD(P_Y\|Q_Y) + c)\right)^{1/2} = \frac{\sqrt{c}}{\sqrt{n}}\,. \tag{5.39}$$

Improving c to be $o(1)$ in n would violate this lower bound.

Intuitively, think of what happens when we let $P_{Y|X}$ be the identity matrix where $Y = X$. In such a case, Theorem 7 is not true (in order to get a true statement, the constant c should be replaced with a value that grows logarithmically with n, see Section 5.3.2). This is the same setting as the example given above but with $\varepsilon = 0$. It is clear here that Y^n will have $\sqrt{n}$ deviations in the number of 0's from the mean whereas any sequence $\hat{Y}^n$ will have exactly the type $(n/2, n/2)$. This creates an expected Hamming distance of $\sqrt{n}$. Slightly increasing ε above zero will not change the Hamming distance by much but make $P_{Z|X}$ strictly positive.

5.3.2 Expression for Divergence Under Fixed Types

We will use the following to show Theorem 7. This proposition can be used for any $P_{Y|X}$, not just those which are strictly positive.

Proposition 11. *Let U be uniform on the type $T_n(P)$ and $(X,Y)^n$ be iid from $(P \times P_{Y|X})$. Let P_Y be the marginal distribution of Y under $(P \times P_{Y|X})$.*

Then for all Q_Y,

$$D(P_{Y|X}^n \circ U \| Q_Y^n) = nD(P_Y\|Q_Y) + \mathbb{E}\left[\log \frac{\mathbb{P}[A=1|Y^n]}{\mathbb{P}[A=1]} \middle| A = 1\right] \tag{5.40}$$

where $A = \mathbb{I}\{X^n \in T_n(P)\}$.

For notation, we use $\mathbb{P}$ to mean under the probability where $(X,Y)^n$ is iid from $(P \times P_{Y|X})$.

Proof. Note that $(P_{Y|X}^n \circ U)(y^n) = \mathbb{P}[Y^n = y^n | A = 1]$.

$$D(P_{Y|X}^n \circ U \| Q_Y^n) = \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[Y^n = y^n | A = 1]}{Q_Y^n(y^n)} \tag{5.41}$$

$$= \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[A = 1 | Y^n = y^n]\mathbb{P}[Y^n = y^n]}{\mathbb{P}[A = 1] Q_Y^n(y^n)} \tag{5.42}$$

$$= \mathbb{E}\left[\log \frac{\mathbb{P}[Y^n]}{Q_Y^n(Y^n)} \middle| A = 1\right] + \mathbb{E}\left[\log \frac{\mathbb{P}[A = 1|Y^n]}{\mathbb{P}[A = 1]} \middle| A = 1\right] . \tag{5.43}$$

The marginal distribution $P_Y(a)$ is also the probability that any position t in sequence Y^n takes the value a, i.e. $P_Y(a) = \mathbb{P}[Y_t = a|A = 1]$. This occurs since U is uniform on all permutations of type $T_n(P)$. We get for the first term in the sum,

$$\mathbb{E}\left[\log \frac{\mathbb{P}[Y^n]}{Q_Y^n(Y^n)} \middle| A = 1\right] = \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{P_Y^n(y^n)}{Q_Y^n(y^n)} \tag{5.44}$$

$$= \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \sum_a n \frac{|\{t : y_t = a\}|}{n} \log \frac{P_Y(a)}{Q_Y(a)} \tag{5.45}$$

$$= n \sum_a P_Y(a) \log \frac{P_Y(a)}{Q_Y(a)} \tag{5.46}$$

$$= nD(P_Y \| Q_Y) . \tag{5.47}$$

This give the result (5.40). □

Lemma 12. *Let $P = (p_1, ..., p_q) \in \mathcal{P}_n$ and let $A = \mathbb{I}\{X^n \in T_n(P)\}$. Then if $(X, Y)^n$ is drawn iid from $(P \times P_{Y|X})$,*

$$\log \frac{1}{\mathbb{P}[A = 1]} \leq -\frac{1}{2} \log n + \sum_{i:p_i>0} \frac{1}{2} \log p_i n + \frac{q-1}{2} \log 2\pi + \frac{1}{12n} . \tag{5.48}$$

For this proof, we will use a Stirling approximation type bound from [45]: For positive integers n,

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}} . \tag{5.49}$$

Proof. We will assume that all $p_i > 0$ since we can always reduce P to a shorter vector and decrease q.

The probability a specific type occurs is given by the multinomial distribution.

$$-\log \mathbb{P}[A = 1] = -\log \left(\frac{n!}{\prod_{i=1}^q (p_i n)!} \prod_{i=1}^q p_i^{p_i n} \right) \tag{5.50}$$

$$= -\log \left(\frac{n!}{n^n}\right) - \log \left(\prod_{i=1}^q \frac{(p_i n)^{p_i n}}{(p_i n)!} \right) \tag{5.51}$$

$$\leq n - \frac{1}{2} \log n - \frac{1}{2} \log 2\pi + \sum_{i=1}^q \left(-p_i n + \frac{1}{2} \log p_i n + \frac{1}{2} \log 2\pi + \frac{q}{12n} \right) . \tag{5.52}$$

We used (5.49) last inequality (we can do this since each $p_i n$ is an integer greater than 0). Combining terms gives the result. □

Theorem 7 Without Strictly Positive

We briefly mentioned above that if we remove the strictly positive requirement for $P_{Y|X}$ in Theorem 7, in the worst case, the constant c would need to be replaced with a logarithmic term. To be exact, if $P_{Y|X}$ is

not strictly positive, c needs to be replaced with a poly-logarithmic term in n. Using Proposition 11 and Lemma 12, we can get an upperbound on c with

$$c = \mathbb{E}\left[\log \frac{\mathbb{P}[A=1|Y^n]}{\mathbb{P}[A=1]}\middle| A=1\right] \tag{5.53}$$

$$\leq \frac{q-1}{2}\log n + c' + \mathbb{E}\left[\log \mathbb{P}[A=1|Y^n]\middle| A=1\right] \tag{5.54}$$

$$\leq \frac{q-1}{2}\log n + c'\,. \tag{5.55}$$

We used the fact that the largest value $\mathbb{P}[A=1|Y^n]$ can take is 1 since it is a probability. The inequality is tight when $P_{Y|X}$ is the identity matrix (when $Y=X$). We also discuss this in later sections where we look at DMC matrices that are not strictly positive.

5.3.3 Concentration of Sums of Independent Variables

The concentration function $Q(X;\lambda)$ of random variable X is defined by

$$Q(X;\lambda) = \sup_x \mathbb{P}[x \leq X \leq x+\lambda] \tag{5.56}$$

for every $\lambda \geq 0$ [46]. Let $S_n = \sum_{i=1}^n X_i$ where the X_i are independent.

Theorem 9 (Petrov [46])**.** *Let the numbers a_i and b_i be such that*

$$\mathbb{P}\left[X_i - a_i \leq -\frac{\lambda_i}{2}\right] \geq b_i \tag{5.57}$$

$$\mathbb{P}\left[X_i - a_i \geq \frac{\lambda_i}{2}\right] \geq b_i \tag{5.58}$$

for $i = 1, ..., n$. Then there exists a universal constant α so that

$$Q(S_n;\lambda) \leq \alpha\lambda\left(\sum_{i=1}^n \lambda_i^2 b_i\right)^{-1/2} \tag{5.59}$$

for every positive $\lambda_1, ..., \lambda_n$ none of which exceeds λ.

To apply Theorem 9 to our problem, each X_i is a Bernoulli random variable where the probability $X_i = 1$ is p_i. Let $b_i = \min\{p_i, 1-p_i\}$. We can fix $a_i = 1/2$. We can also fix $\lambda_i = 1/2$ and $\lambda = 1/2$, though this exact value does not matter so long as $\lambda < 1 - \varepsilon$ for a small $\varepsilon > 0$.

This gives that for some integer w

$$\mathbb{P}[S_n = w] \leq Q(S_n; 1/2) \tag{5.60}$$

$$\leq \frac{\alpha(1/2)}{\sqrt{\sum_{i=1}^n (1/2)^2 \min\{p_i, 1-p_i\}}} \tag{5.61}$$

$$\leq \frac{\alpha}{\sqrt{\sum_{i=1}^n \min\{p_i, 1-p_i\}}}\,. \tag{5.62}$$

We will use this in the next lemma which is key to computing the second term in (5.40).

Lemma 13. *There are n balls thrown in q bins independently, so that for the i-th ball, the relative probability of landing in bin b is $p_{i,b}$.*

Let N_b be the ball count of the b-th bin. Then if $\pi_b > 0$ for all b and $\sum_b \pi_b = 1$, we have

$$\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \leq \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{B}} \tag{5.63}$$

where

$$B = c_*^{q-1} \frac{\prod_b \pi_b}{\pi_{max}} \tag{5.64}$$

$$c_* = \min_i \frac{c_-(i)}{c_+(i)} \tag{5.65}$$

$$c_-(i) = \min_b \frac{p_{i,b}}{\pi_b} \tag{5.66}$$

$$c_+(i) = \max_b \frac{p_{i,b}}{\pi_b} \tag{5.67}$$

$$\pi_{max} = \max_b \pi_b \tag{5.68}$$

and α is the universal constant used in Theorem 9.

The values $p_{i,b}$ need only be relative probabilities. The exact probability that ball i lands in bin b is $p_{i,b}/\sum_a p_{i,a}$. Given that ball i does not land in some bin outside of some set $\mathcal{B}$, the probability that ball i lands in $b \in \mathcal{B}$ is $p_{i,b}/\sum_{a\in\mathcal{B}} p_{i,a}$

Proof. For notation, let $X_{i,b}$ be the indicator variable of whether ball i was thrown into bin b. We can express $N_b = \sum_{i=1}^n X_{i,b}$. Arrange the indices so that $\pi_1 \le \pi_2 \cdots \le \pi_q$.

First observe that

$$\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] = \prod_{b=1}^{q} \mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}]\,. \tag{5.69}$$

For $b = q$,

$$\mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}] = 1\,. \tag{5.70}$$

For $b < q$, we can compute for any i that

$$\min\left\{\frac{p_{i,b}}{\sum_{a=b}^q p_{i,a}}, 1 - \frac{p_{i,b}}{\sum_{a=b}^q p_{i,a}}\right\} = \min\left\{\frac{p_{i,b}}{\sum_{a=b}^q p_{i,a}}, \frac{\sum_{a>b}^q p_{i,a}}{\sum_{a=b}^q p_{i,a}}\right\} \tag{5.71}$$

$$= \min\left\{\frac{\pi_b \frac{p_{i,b}}{\pi_b}}{\sum_{a=b}^q \pi_a \frac{p_{i,a}}{\pi_a}}, \frac{\sum_{a>b}^q \pi_a \frac{p_{i,a}}{\pi_a}}{\sum_{a=b}^q \pi_a \frac{p_{i,a}}{\pi_a}}\right\} \tag{5.72}$$

$$\ge \frac{\min_a \frac{p_{i,a}}{\pi_a}}{\max_a \frac{p_{i,a}}{\pi_a}} \min\left\{\frac{\pi_b}{\sum_{a=b}^q \pi_a}, \frac{\sum_{a>b}^q \pi_a}{\sum_{a=b}^q \pi_a}\right\} \tag{5.73}$$

$$\ge \min_i \frac{c_-(i)}{c_+(i)} \frac{1}{\sum_{a=b}^q \pi_a} \min\left\{\pi_b, \sum_{a>b}^{q} \pi_a\right\} \tag{5.74}$$

$$= c_* \frac{\pi_b}{\sum_{a=b}^q \pi_a}\,. \tag{5.75}$$

We get the last equality because we have arranged π_b in increasing order. Hence by (5.62)

$$\mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}] \le \frac{\alpha}{\sqrt{\left(n - \sum_{a=1}^{b-1} n\pi_a\right) c_* \frac{\pi_b}{\sum_{a=b}^q \pi_a}}} \tag{5.76}$$

$$= \frac{\alpha}{\sqrt{\frac{n - \sum_{a=1}^{b-1} n\pi_a}{n} n c_* \frac{\pi_b}{\sum_{a=b}^q \pi_a}}} \tag{5.77}$$

$$= \frac{\alpha}{n^{1/2}\sqrt{c_* \pi_b}} \tag{5.78}$$

where we used that $n - \sum_{a=1}^{b-1} n\pi_a = n\sum_{a=b}^{q} \pi_a$ to get the last inequality. Taking a product of all terms in (5.69), gives

$$\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \leq \prod_{b=1}^{q-1} \frac{\alpha}{n^{1/2}\sqrt{c_* \pi_b}} \tag{5.79}$$

$$= \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{c_*^{q-1} \prod_{b=1}^{q-1} \pi_b}} \tag{5.80}$$

$$= \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{c_*^{q-1} \frac{\prod_b \pi_b}{\pi_q}}} \tag{5.81}$$

$$= \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{B}}\,. \tag{5.82}$$

□

5.3.4 Completing Proof of Theorem 7 and Determining Capacity.

Proof of Theorem 7. First we show the easier lower bound. Using Proposition 11, we need only to show that

$$\mathbb{E}\left[\log \frac{\mathbb{P}[A=1|Y^n]}{\mathbb{P}[A=1]}\middle| A=1\right] \geq 0\,. \tag{5.83}$$

We do this by,

$$\mathbb{E}\left[\log \frac{\mathbb{P}[A=1|Y^n]}{\mathbb{P}[A=1]}\middle| A=1\right] = \sum_{y^n} \mathbb{P}[Y^n = y^n|A=1] \log \frac{\mathbb{P}[A=1|Y^n=y^n]}{\mathbb{P}[A=1]} \tag{5.84}$$

$$= \sum_{y^n} \mathbb{P}[Y^n = y^n|A=1] \log \frac{\mathbb{P}[Y^n=y^n|A=1]\mathbb{P}[A=1]}{\mathbb{P}[Y^n=y^n]\mathbb{P}[A=1]} \tag{5.85}$$

$$= \sum_{y^n} \mathbb{P}[Y^n = y^n|A=1] \log \frac{\mathbb{P}[Y^n=y^n|A=1]}{\mathbb{P}[Y^n=y^n]} \tag{5.86}$$

$$= D(\mathbb{P}[Y^n|A=1] \| \mathbb{P}[Y^n]) \tag{5.87}$$

$$\geq 0 \tag{5.88}$$

since divergences are always positive.

To get the upper bound, we will express

$$\mathbb{E}\left[\log \frac{\mathbb{P}[A=1|Y^n]}{\mathbb{P}[A=1]}\middle| A=1\right] = \mathbb{E}\left[\log \mathbb{P}[A=1|Y^n]\middle| A=1\right] - \log \mathbb{P}[A=1] \tag{5.89}$$

and use Lemma 13 for the first term in the sum and Lemma 12 for the second term in sum.

To see why Lemma 13 applies to the first term, note that the first term is trying to calculate given some Y^n, what the probability that the type of X^n is equal to $T_n(P)$. This is under a distribution where $(X,Y)^n \sim (P_{Y|X} \times P)$.

We will express the type $T_n(P)$ with $P = (\pi_1, ..., \pi_q)$ where $P \in \mathcal{P}_n$. This implies that $\pi_b = \mathbb{P}[X = b]$. Let the balls described in Lemma 13 be each of the elements of Y^n. If $Y_i = y$, then let $p_{i,b} = \mathbb{P}[X = b, Y = y]$. This way $p_{i,b} = \mathbb{P}[X = b|Y = y]\mathbb{P}[Y = y]$ is appropriately the relative probability that $X_i = b$ given that $Y_i = y$. As in Lemma 13, N_b is the number of balls in bin b. Then the probability that $X^n \in T_n(P)$ is equivalent to $\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q]$. This is computed for a specific value of Y^n, but notice that the expression we derived for $\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q]$ in Lemma 13 does not depend on Y^n.

Before computing the rest of the expression, we need to pay particular attention to the case when there exists a b such that $\pi_b = 0$. If $\pi_b = 0$, then $\mathbb{P}[X = b] = 0$, which would also imply that $p_{i,b} = 0$ for all i. In this case, we can remove the symbol b (or bin b in the interpretation of Lemma 13) from consideration

and apply Lemma 13 to just the symbols with non-zero probability. We can always reorder the symbols, so that the first q' of the q symbols all have $\pi_b > 0$ and the remaining $b > q'$ are such that $\pi_b = 0$. Like in Lemma 13, we can define

$$B = c_*^{q'-1} \frac{\prod_{b=1}^{q'} \pi_b}{\pi_{max}} \tag{5.90}$$

$$c_* = \min_i \frac{c_-(i)}{c_+(i)} \tag{5.91}$$

$$c_-(i) = \min_{b:b\le q'} \frac{p_{i,b}}{\pi_b} \tag{5.92}$$

$$c_+(i) = \max_{b:b\le q'} \frac{p_{i,b}}{\pi_b}\,. \tag{5.93}$$

$$\mathbb{E}\left[\log \mathbb{P}[A=1|Y^n]\middle| A=1\right] = \log \mathbb{P}[N_1 = n\pi_1, \ldots, N_{q'} = n\pi_{q'}] \tag{5.94}$$

$$= \log \frac{\alpha^{q'-1}}{n^{(q'-1)/2}\sqrt{B}} \tag{5.95}$$

$$= \log \left(\frac{\alpha^{q'-1}}{n^{(q'-1)/2} c_*^{\frac{q'-1}{2}}} \left(\frac{\pi_{max}}{\prod_b \pi_b} \right)^{1/2} \right) \tag{5.96}$$

$$= \log \left(\frac{\alpha^{q'-1}}{c_*^{\frac{q'-1}{2}}} \left(\frac{n\pi_{max}}{\prod_{b=1}^{q'} n\pi_b} \right)^{1/2} \right) \tag{5.97}$$

$$= \frac{1}{2}\log n\pi_{max} - \sum_{b:\pi_b>0} \frac{1}{2}\log n\pi_b + (q'-1)\log\left(\frac{\alpha}{\sqrt{c_*}}\right) \tag{5.98}$$

$$\le \frac{1}{2}\log n - \sum_{b:\pi_b>0} \frac{1}{2}\log n\pi_b + c'\,. \tag{5.99}$$

where c' is constant that does not depend on n. Importantly, the value of c' also does not depend on π_b for any b. Even though c' depends on c_*, the latter quantity depends only on the quotient $p_{i,b}/\pi_b$.

$$\frac{p_{i,b}}{\pi_b} = \frac{\mathbb{P}[X=b, Y=y]}{\pi_b} = \frac{\mathbb{P}[Y=y|X=b]\mathbb{P}[X=b]}{\pi_b} = \mathbb{P}[Y=y|X=b] = P_{Y|X}(y|b)\,. \tag{5.100}$$

So c_* only depends on $P_{Y|X}$.

Combining these terms with those from Lemma 12 gives that the expression in (5.89) is a constant when $P_{Y|X}$ is strictly positive. This constant depends on q and $P_{Y|X}$ but not on n or π_b for any b.

□

Proof of Theorem 6. Using Theorem 7 with Proposition 9 completes the proof for strictly positive DMC. □

5.3.5 Comparing Theorem 7 to Stam

Here we give some details on how our result relates to that of Stam in [33]. First, Stam's setting has noiseless observations whereas our setting has noise. However, using data processing inequality, we can always use Stam's result as an upperbound for the noisy case. Second, Stam's result generalizes to m observations, where m can be less than n. Our result Theorem 7 only applies to exactly n observations. However, we can also use a version of Han's inequality for divergence [47, Proposition 5.5] (applies when the second probability argument is independent over the entries of the vector Y^n), to get the following corollary:

Corollary 3. *Let $Y^n \sim P_{Y|X}^n \circ U$, so that P_{Y^n} is the distribution as in Theorem 2. Then for every $m \leq n$ we have:*

$$D(P_{Y^m} \| Q_Y^m) \leq mD(P_Y \| Q_Y) + \frac{m}{n}c \tag{5.101}$$

where c is the same constant as in Theorem 7.

Here, Y^m are the first m entries of vector Y^n. In Stam's setting we have $Q_Y = P_Y$, so

$$D(P_{Y^m} \| P_Y^m) \leq \frac{m}{n}c\,. \tag{5.102}$$

Using Stam's result with data processing gives

$$D(P_{Y^m} \| P_Y^m) \leq D(P_{X^m} \| P_X^m) \leq \frac{(q-1)}{2} \frac{m(m-1)}{(n-1)(n-m+1)}\,. \tag{5.103}$$

The above equation, which is presented as Stam's final result, is actually not the tightest when m is close to n. For instance, when $m = n$, (5.103) gives $\frac{q-1}{2}n$ which is far from $\frac{q-1}{2}(\log n + c')$, the actual divergence when computed directly. Work in [48] is able to make for an improvement on Stam's bound when m is close to n. We show an easier improvement, using an intermediate result in the proof of (5.103). We can derive for larger m that

$$\begin{aligned}
D(P_{Y^m} \| P_Y^m) &\leq D(P_{X^m} \| P_X^m) && (5.104)\\
&\leq \frac{q-1}{n-1} \sum_{t=1}^{m-1} \frac{t}{n-t} && (5.105)\\
&= \frac{q-1}{n-1} \sum_{j=n-m+1}^{n-1} \frac{n-j}{j} && (5.106)\\
&= \frac{q-1}{n-1} \left(n \left(\sum_{j=n-m+1}^{n-1} \frac{1}{j} \right) - (m-1) \right) && (5.107)\\
&= \frac{q-1}{n-1} \left(n \left(\log(n-1) - \log(n-m) + c'' \right) - (m-1) \right) && (5.108)\\
&= \frac{q-1}{n-1} \left(n \log \frac{n-1}{n-m} + nc'' - (m-1) \right) && (5.109)\\
&= (q-1) \log \frac{n-1}{n-m} + O(q) && (5.110)
\end{aligned}$$

for $m < n$ and c'' is a constant leftover from approximating the harmonic sum by a log.

When $m << n$, (5.103) is the better bound. The bound in (5.102) will be a tighter bound than both (5.103) and (5.110) for the case when m is very close to n, such as when $n - m = o(n)$. When m is linear in n, then it becomes important to compare the constant factors. Let $\gamma = m/n$. To get an estimate on when our bound is tighter, we will first assume n is large and ignore the lower order constants which appear in the bounds. Using Remark 4, the bound in (5.102) can be tighter than (5.103) and (5.110) for large n if

$$\frac{1}{2} \log \frac{2\pi\alpha^2}{c_*} \leq \min \left\{ \frac{\gamma}{1-\gamma}, \frac{1}{\gamma} \log \frac{1}{1-\gamma} \right\}. \tag{5.111}$$

This can occur for certain values of γ depending on the size of c_*, which is a function of $P_{Y|X}$.

5.4 Block Diagonal Case

With a small modification to the proof of Theorem 6, we can show a converse bound for block diagonal matrices where each block is strictly positive. The modification uses that each block is independent from all

the other blocks, so we can apply the bound for strictly positive matrices separately to each block. We will need to show a separate achievability result to match this converse bound.

As a sanity check, the block diagonal case captures the situation where $P_{Z|X}$ is the identity matrix. In which case, it is possible to use all possible permutations of symbols as messages. No errors are allowed so decoding is straight-forward. Using an identity matrix of size $q \times q$ for the DMC, for each n,

$$R = \frac{\log M}{\log n} = \frac{\log \binom{n}{q-1}}{\log n} \approx \frac{\log(c^{q-1} n^{q-1}/(q-1)^{q-1})}{\log n} = q - 1 + \frac{c'}{\log n} \tag{5.112}$$

which goes to $q-1$ asymptotically. This matches our block diagonal converse result.

5.4.1 Converse

Proposition 12 (Block Diagonal Converse)**.** *Suppose $P_{Z|X}$ can be written as a block diagonal matrix with β blocks, so that each block is strictly positive. Then,*

$$C_{\mathsf{perm}}(P_{Z|X}) \leq \frac{\mathsf{rank}(P_{Z|X}) + \beta - 2}{2}\,. \tag{5.113}$$

Proof. We will use Proposition 9 but we need to show a version of the upper bound in Theorem 7 which applies to block diagonal matrices instead of strictly positive matrices.

Fix $\pi \in \mathcal{P}_n$. Arrange the matrix $P_{Z|X}$ in block diagonal form and let $\mathcal{X}_b$ be the set of symbols in $\mathcal{X}$ which are in the bth block. Let $(X,Y)^n$ be generated iid from $(\pi \times P_{Y|X})$. Let W_i be the number of X which equals i, i.e.

$$W_i = |\{t : X_t = i\}|\,. \tag{5.114}$$

Define

$$A_b = \left\{ \bigcap_{i \in \mathcal{X}_b} W_i = \pi_i n \right\}. \tag{5.115}$$

This is the event that all symbols i associated with block b occur with the count $\pi_i n$. Each block has its own separate set of output symbols in $\mathcal{Y}$. The probability of W_i is independent of what happens in other blocks. Let $Y^n(b)$ (and $y^n(b)$) be notation for the counts restricted to just the output symbols associated with the bth block.

Using the definition in Proposition 11, notice that $A = \mathbb{I}\left[\bigcap_{b=1}^{\beta} A_b\right]$. Then using (5.40) with Lemma 12,

$$D(P^n_{Y|X} \circ U \| Q^n_Y) = nD(P_Y\|Q_Y) + \mathbb{E}\left[\log \frac{\mathbb{P}[A=1|Y^n]}{\mathbb{P}[A=1]}\middle| A=1\right] \tag{5.116}$$

$$\leq nD(P_Y\|Q_Y) - \frac{1}{2}\log n + \sum_{i:\pi_i>0} \frac{1}{2}\log \pi_i n + c + \mathbb{E}\left[\log \mathbb{P}[A=1|Y^n]\middle| A=1\right]. \tag{5.117}$$

For any Y^n,

$$\mathbb{P}[A=1|Y^n] = \prod_{b=1}^{\beta} \mathbb{P}[A_b = 1 | Y^n(b) = y^n(b)]\,. \tag{5.118}$$

Each block is a strictly positive matrix. From Lemma 13 and following the same calculations that results in (5.99), we know that

$$\log \mathbb{P}[A_b = 1|Y^n(b) = y^n(b)] \leq \frac{1}{2}\log n - \sum_{i \in \mathcal{X}_b : \pi_i > 0} \frac{1}{2}\log n\pi_i + c' \tag{5.119}$$

and thus

$$\log \mathbb{P}[A=1|Y^n] \leq \sum_{b=1}^{\beta} \left(\frac{1}{2} \log n - \sum_{i \in \mathcal{X}_b : \pi_i > 0} \frac{1}{2} \log n\pi_i + c' \right) \tag{5.120}$$

$$= \frac{\beta}{2} \log n - \sum_{i:\pi_i>0} \frac{1}{2} \log n\pi_i + \beta c' \,. \tag{5.121}$$

This holds for all Y^n so it automatically gives the expected value. Putting all these terms together, for any π, we get

$$D(P_{Y|X}^n \circ U \| Q_Y^n) = nD(P_Y||Q_Y) + \frac{\beta-1}{2} \log n + c'' \tag{5.122}$$

where c'' combines all the constants. Using Proposition 9, gives

$$C_{\mathsf{perm}}(P_{Z|X}) \leq \frac{\mathsf{rank}(P_{Z|X})-1}{2} + \lim_{n\to\infty} \frac{\frac{\beta-1}{2}\log n + c''}{\log n} \tag{5.123}$$

$$= \frac{\mathsf{rank}(P_{Z|X})-1}{2} + \frac{\beta-1}{2} \tag{5.124}$$

$$= \frac{\mathsf{rank}(P_{Z|X})+\beta-2}{2} \,. \tag{5.125}$$

□

5.4.2 Achievability

Proposition 13 (Block Diagonal Achievability)**.** *Suppose $P_{Z|X}$ can be written as a block diagonal matrix with β blocks, so that each block is strictly positive. Then,*

$$C_{\mathsf{perm}}(P_{Z|X}) \geq \frac{\mathsf{rank}(P_{Z|X})+\beta-2}{2} \,. \tag{5.126}$$

Proof. The achievability proof encodes using two steps. The first step is a zero-error code based on which block in the block diagonal matrix the symbols are associated with. Let M_1 denote the total possible messages (or rather message stems) for the first step. The second step operates only on each block independently, and uses the achievability given by (5.3). Let M_2 denote the total messages (or message tails) possible here.

Label the β blocks in $P_{Z|X}$ as $B_1, ..., B_\beta$. Define the sets of input symbols $\mathcal{X}_1, .., \mathcal{X}_\beta$ and output symbols $\mathcal{Y}_1, ..., \mathcal{Y}_\beta$, so that $\mathcal{X}_b$ and $\mathcal{Y}_b$ are the input and output symbols respectively associated with block B_b. (In other words, if $p_{ij} > 0$ and p_{ij} falls into block B_b, then $i \in \mathcal{X}_b$ and $j \in \mathcal{Y}_b$.) These sets $\mathcal{X}_1, .., \mathcal{X}_\beta$ and $\mathcal{Y}_1, ..., \mathcal{Y}_\beta$ will both be disjoint.

Let $L = \mathsf{rank}(P_{Z|X})$ and let $L_b = \mathsf{rank}(B_b)$. Because of the block diagonal structure, $L = \sum_{b=1}^{\beta} L_b$.

For fixed n, set aside the first $n/2$ input symbol positions so that exactly $n/(2\beta)$ are from set $\mathcal{X}_b$ for each b. These will not be used for the first step of the two-step code and is used to make the analysis of the second step easier. The remaining $n/2$ positions can be encoded using symbols from any set and this is used to make the first step of the code. There are

$$\binom{n/2}{\beta-1} \geq \left(\frac{n/2}{\beta-1} \right)^{\beta-1} \tag{5.127}$$

possible combinations of symbols chosen from β blocks, disregarding order. The DMC will map the symbols in set $\mathcal{X}_b$ to symbols in set $\mathcal{Y}_b$ without any error. Hence, (5.127) is the number of messages M_1 the first step can encode without any error.

Once it is determined how many symbols of each set will be used, we can determine which symbol in the set will be used for the second step. Suppose there are n_b positions which are designated for symbols in set $\mathcal{X}_b$. This includes the $n/(2\beta)$ we set aside in the beginning and how ever many were chosen to make the first

step of the code. Using (5.3), we know there exists a encoder-decoder pair (f_{n_b}, g_{n_b}) so that the decoding error is vanishingly small as $n_b \to \infty$. Just by choosing which symbol in $\mathcal{X}_b$ to send, for some $\varepsilon_{n_b} > 0$ where $\varepsilon_{n_b} \to 0$, we can encode a set of messages with size M_b satisfying

$$\log M_b \geq \left(\frac{L_b - 1}{2} - \varepsilon_{n_b} \right) \log n_b \,. \tag{5.128}$$

The set of messages possible for all the β different sets is

$$\log M_2 = \log \prod_{b=1}^{\beta} M_b \tag{5.129}$$

$$\geq \sum_{b=1}^{\beta} \left(\frac{L_b - 1}{2} - \varepsilon_{n_b} \right) \log n_b \tag{5.130}$$

$$\geq \sum_{b=1}^{\beta} \left(\frac{L_b - 1}{2} - \varepsilon_{n_b} \right) \log \frac{n}{2\beta} \tag{5.131}$$

$$= \left(\frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b} \right) \log \frac{n}{2\beta} \,. \tag{5.132}$$

The total number of messages is the product of those available at the first and second steps.

$$\log M = \log M_1 + \log M_2 \tag{5.133}$$

$$\geq \log \left(\frac{n/2}{\beta - 1} \right)^{\beta - 1} + \left(\frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b} \right) \log \frac{n}{2\beta} \tag{5.134}$$

$$= (\beta - 1) \log \frac{n}{2\beta - 2} + \left(\frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b} \right) \log \frac{n}{2\beta} \tag{5.135}$$

$$\geq \left(\frac{2\beta - 2}{2} + \frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b} \right) \log \frac{n}{2\beta} \tag{5.136}$$

$$= \left(\frac{L + \beta - 2}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b} \right) \log \frac{n}{2\beta} \,. \tag{5.137}$$

Since each $n_b \geq \frac{n}{2\beta} \to \infty$ as $n \to \infty$, asymptotically the term $\sum_{b=1}^{\beta} \varepsilon_{n_b}$ disappears.
The achievable rate is given by

$$R = \frac{\log M}{\log n} \geq \left(\frac{L - \beta - 2}{2} + \sum_{b=1}^{\beta} \varepsilon_{n_b} \right) \frac{\log n - \log 2\beta}{\log n} \to \frac{L - \beta - 2}{2} \,. \tag{5.138}$$

□

Combining Proposition 12 and Proposition 13 gives the first result in Theorem 8.

5.5 Erasure and Z Channels

5.5.1 Preliminaries

The following lemma will be useful for computing the probability of event A (see Proposition 11) when the DMC matrix is not strictly positive. It is a straight-forward concentration bound which is a direct application of Bernstein's inequality. We choose to write the proof anyways for completeness.

Lemma 14. *Suppose that Y is a sum of n independent Bernoulli random variables. Let $\mathbb{E}[Y]$ be the expected value of Y.*

Fix γ. If $\mathbb{E}[Y] > 2\gamma \log n$, then with probability at least $1 - 2/n^{\gamma/4}$, we have that

$$\mathbb{E}[Y] > \mathbb{E}[Y] - \sqrt{\mathbb{E}[Y]\gamma \log n} \geq \frac{1}{5}\mathbb{E}[Y]\,. \tag{5.139}$$

Proof. Let $Y = \sum_{i=1}^{n} X_i$ where X_i is the ith Bernoulli random variable. Let $0 < p_i < 1$ be the probability of $X_i = 1$.

$$\sum_{i=1}^{n} \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = \sum_{i=1}^{n} p_i(1-p_i) \leq \sum_{i=1}^{n} p_i = \mathbb{E}[Y]\,. \tag{5.140}$$

We will use Bernstein's inequality for bounded variables [49, Theorem 2.8.4].

$$\mathbb{P}\left[Y - \mathbb{E}[Y] \leq -\sqrt{\mathbb{E}[Y]\gamma \log n}\right] \leq 2\exp\left(\frac{-\frac{1}{2}\mathbb{E}[Y]\gamma \log n}{\sum_{i=1}^{n}\mathbb{E}[(X_i - \mathbb{E}[X_i])^2] + \frac{1}{3}1\sqrt{\mathbb{E}[Y]\gamma \log n}}\right) \tag{5.141}$$

$$\leq 2\exp\left(\frac{-\frac{1}{2}\mathbb{E}[Y]\gamma \log n}{\mathbb{E}[Y] + \frac{1}{3}\sqrt{\mathbb{E}[Y]\gamma \log n}}\right) \tag{5.142}$$

$$\leq 2\exp\left(\frac{-\frac{1}{2}\gamma \log n}{1 + \frac{1}{3}\frac{\sqrt{\gamma \log n}}{\sqrt{\mathbb{E}Y}}}\right)\,. \tag{5.143}$$

Using that $\mathbb{E}[Y] > 2\gamma \log n$, we have $1 + \frac{1}{3}\frac{\sqrt{\gamma \log n}}{\sqrt{\mathbb{E}[Y]}} \leq 1 + \frac{1}{3}\frac{\sqrt{\gamma \log n}}{\sqrt{2\gamma \log n}} \leq 2$.

$$\mathbb{P}\left[Y - \mathbb{E}[Y] \leq -\sqrt{\mathbb{E}[Y]\gamma \log n}\right] \leq 2\exp\left(\frac{-1}{4}\gamma \log n\right) \leq \frac{2}{n^{\gamma/4}}\,. \tag{5.144}$$

Hence, with probability $1 - 2/n^{\gamma/4}$,

$$\mathbb{E}[Y] \geq \mathbb{E}[Y] - \sqrt{\mathbb{E}Y\gamma \log n} \geq \mathbb{E}[Y] - \sqrt{\mathbb{E}[Y]\frac{1}{2}\mathbb{E}[Y]} \geq \left(1 - \frac{1}{\sqrt{2}}\right)\mathbb{E}[Y] \geq \frac{1}{5}\mathbb{E}[Y]\,. \tag{5.145}$$

□

5.5.2 The q-ary Erasure Channel

We now can prove the converse bound for q-ary erasure channels, where q is the number of input symbols. Let $K = q + 1$ represent the erased symbol.

The matrix $P_{Z|X}$ for a q-ary erasure channel has the following structure:

$$P_{Z|X} = \begin{bmatrix} p_{11} & 0 & \cdots & 0 & p_{1K} \\ 0 & p_{22} & \cdots & 0 & p_{2K} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{qq} & p_{qK} \end{bmatrix}\,. \tag{5.146}$$

We will assume that each $p_{iK} > 0$.

Proof of Item 2 of Theorem 8. Fix $\pi = (\pi_1, ..., \pi_q)$ where $\pi \in \mathcal{P}_n$. (We will assume each $\pi_i > 0$, otherwise we can remove it.) Reorder the symbols in $\{1, ..., q\}$ so that $\pi_1 \leq \pi_2 \leq ... \leq \pi_q$. (We will use $P_{Y|X} = P_{Z|X}$.)

Following Proposition 11, let $(X,Y)^n$ be generated iid according to $(\pi \times P_{Y|X})$. To use Proposition 11 we need to determine $\mathbb{P}[\log \mathbb{P}[A=1|Y^n]|A=1]$.

Unlike the case of strictly positive $P_{Y|X}$, the value of $\mathbb{P}[A=1|Y^n]$ depends on Y^n. For instance, it is easy to see that when the erasure symbol K does not appear, then $\mathbb{P}[A=1|Y^n]=1$. While Y^n like this can occur under the event $A=1$, we want to show that these events are rare, this way the expected value of $\mathbb{P}[A=1|Y^n]$ given that $A=1$ will be much smaller than 1 and close to the value which will give our result. We will first show a concentration result on Y^n given that $A=1$.

Let U_b be the random variable which gives the count of the number of times the symbol b is erased, i.e

$$U_b = \sum_{i=1}^{n} \{(X_i, Y_i) = (b, K)\}. \tag{5.147}$$

Let $v_b(y^n) = \{U_b | A=1, Y^n = y^n\}$. Note that $v_b(y^n)$ is deterministic. If $A=1$ and Y^n is known, we can determine exactly what U_b is.

Define $S_b = \sum_{a \geq b} U_a$. Given Y^n, S_1 is deterministic. Given Y^n and $U_1, .., U_{b-1}$, S_b is deterministic.

Using Lemma 14, since $\mathbb{E}[S_b] = n \sum_{a \geq b} \pi_a p_{aK} \geq n \pi_q p_{qK} > 2\gamma \log n$ for some γ (chosen later) and all b for large enough n, we have

$$\mathbb{P}\left[S_b > \frac{1}{5} n \sum_{a \geq b} \pi_a p_{aK} \right] \geq 1 - 2/n^{\gamma/4}. \tag{5.148}$$

Using the union bound,

$$\mathbb{P}\left[\bigcap_{b=1}^{q} \left\{ S_b > \frac{1}{5} n \sum_{a \geq b} \pi_a p_{aK} \right\} \right] \geq 1 - 2q/n^{\gamma/4}. \tag{5.149}$$

Next, for any y^n which has positive probability given $A=1$,

$$\mathbb{P}[A=1|Y^n = y^n] = \mathbb{P}\left[\bigcap_{b=1}^{q} U_b = v_b(y^n) \middle| Y^n = y^n \right] \tag{5.150}$$

$$= \prod_{b=1}^{q-1} \mathbb{P}\left[U_b = v_b(y^n) \middle| \bigcap_{a=1}^{b-1} U_a = v_a(y^n), Y^n = y^n \right]. \tag{5.151}$$

We compute the following which is like the proof Lemma 13 of but with appropriate adjustments. Using (5.62),

$$\mathbb{P}\left[U_b = v_b(y^n) \middle| \bigcap_{a=1}^{b-1} U_a = v_a(y^n), Y^n = y^n \right] \leq \frac{\alpha}{\sqrt{\sum_{i=1}^{S_b} \min\left\{ \frac{\pi_b p_{bK}}{\sum_{a \geq b} \pi_a p_{aK}}, \frac{\sum_{a>b} \pi_a p_{aK}}{\sum_{a \geq b} \pi_a p_{aK}} \right\}}}. \tag{5.152}$$

Like in Lemma 13, define $c_- = \min_i p_{iK}$.

$$\min\left\{ \frac{\pi_b p_{bK}}{\sum_{a \geq b} \pi_a p_{aK}}, \frac{\sum_{a>b} \pi_a p_{aK}}{\sum_{a \geq b} \pi_a p_{aK}} \right\} = (\min_i p_{iK}) \min\left\{ \frac{\pi_b}{\sum_{a \geq b} \pi_a p_{aK}}, \frac{\sum_{a>b} \pi_a}{\sum_{a \geq b} \pi_a p_{aK}} \right\} \tag{5.153}$$

$$= \frac{c_- \pi_b}{\sum_{a \geq b} \pi_a p_{aK}}. \tag{5.154}$$

We get the last equality since π_i is in increasing order. Hence

$$\mathbb{P}\left[U_b = v_b(y^n) \middle| \bigcap_{a=1}^{b-1} U_a = v_a(y^n), Y^n = y^n \right] \leq \frac{\alpha}{\sqrt{S_b \frac{c_- \pi_b}{\sum_{a \geq b} \pi_a p_{aK}}}}. \tag{5.155}$$

We can now compute

$$\mathbb{E}[\log \mathbb{P}[A=1|Y^n]|A=1] = \sum_{y^n} \mathbb{P}[Y^n=y^n|A=1] \log \mathbb{P}[A=1|Y^n=y^n] \tag{5.156}$$

$$\leq \log \sum_{y^n} \mathbb{P}[Y^n=y^n|A=1]\mathbb{P}[A=1|Y^n=y^n] \tag{5.157}$$

$$\leq \log \sum_{y^n} \mathbb{P}[Y^n=y^n|A=1] \prod_{b=1}^{q-1} \frac{\alpha}{\sqrt{S_b \frac{c_- \pi_b}{\sum_{a\geq b} \pi_a p_{aK}}}} \tag{5.158}$$

$$\leq \log \left((2q/n^{\gamma/4}) + (1-2q/n^{\gamma/4}) \prod_{b=1}^{q-1} \frac{\alpha}{\sqrt{\left(\frac{1}{5} n \sum_{a\geq b} \pi_a p_{aK}\right) \frac{c_- \pi_b}{\sum_{a\geq b} \pi_a p_{aK}}}} \right) \tag{5.159}$$

$$\leq \log \left((2q/n^{\gamma/4}) + \frac{\alpha^q}{\left(\frac{c_-}{5}\right)^{\frac{q-1}{2}} n^{\frac{q-1}{2}} \sqrt{\frac{\prod_{b=1}^q \pi_b}{\pi_{max}}}} \right) \tag{5.160}$$

where in (5.159) we used (5.149). We can pick γ large enough[3] so that the first term in the logarithm is negligible compared to the second term for large n.

This gives

$$\mathbb{P}[\log \mathbb{P}[A=1|Y^n]|A=1] \leq \log \left(\frac{2\alpha^q}{\left(\frac{c_-}{5}\right)^{\frac{q-1}{2}} n^{\frac{q-1}{2}} \sqrt{\frac{\prod_{b=1}^q \pi_b}{\pi_{max}}}} \right) \tag{5.161}$$

$$\leq \frac{1}{2} \log n - \sum_{b=1}^{q} \frac{1}{2} \log \pi_b n + c' \,. \tag{5.162}$$

The value c' collects all the constants. Combining with Lemma 12, we get that for the q-ary erasure channel and sufficiently large n that

$$D(P_{Y|X} \circ U \| Q_Y^n) \leq n D(P_Y \| Q_Y) + c \tag{5.163}$$

where c does not depend on n or π. Using Proposition 9 completes the proof.

□

5.5.3 Z-Channel

The matrix for the Z-channel [1, p 225] is

$$\begin{bmatrix} p_{11} & p_{12} \\ 0 & 1 \end{bmatrix} \tag{5.164}$$

where we require that $p_{ij} > 0$. (Typically, $p_{11} = p_{12} = 1/2$, but we will consider a more general case.)

We can actually get the capacity of the noisy permutation channel with the Z-channel without altering the proof for the q-ary erasure channels. The transition matrix for the Z-channel can be written as

$$\begin{bmatrix} p_{11} & p_{12} & 0 \\ p_{21} & 0 & p_{23} \end{bmatrix} \tag{5.165}$$

setting $p_{23} = 0$. This does not change the rank of the matrix or the analysis in the proof.

[3] For instance, we can pick $\gamma = 40q$. For large enough n, we still get that $\mathbb{E}[S_b] = n \sum_{a\geq b} \pi_a p_{aK} > 2\gamma \log n$ is true for all b.

Corollary 4. *Let $P_{Z|X}$ be a stochastic matrix for the Z-channel, then*

$$C_{perm}(P_{Z|X}) = \frac{1}{2}\,. \tag{5.166}$$

This is item 3 of Theorem 8.

5.5.4 "Zigzag" Channel

In this section, we explore the limits of our approach. We have a particular DMC matrix which is similar to the q-ary erasure channel, but our method is not known to give a tight converse. We will use a matrix which could be considered a q-ary Z-channel. We will call it a "zigzag" channel since its edges in a transition diagram form a zigzag.

The matrix has the form:

$$\begin{bmatrix} p_{11} & p_{12} & 0 & \cdots & 0 & 0 \\ 0 & p_{22} & p_{23} & \cdots & 0 & 0 \\ 0 & 0 & p_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{q-1,q-1} & p_{q-1,q} \\ 0 & 0 & 0 & \cdots & 0 & p_{qq} \end{bmatrix} \tag{5.167}$$

where each $p_{ij} > 0$. This matrix has rank q.

Suppose that q is odd and that π is such that π_i is 0 for all even values of i. Following the notation and method in Proposition 11, $\mathbb{P}[A = 1|Y^n] = 1$, since any output symbol can be decoded to exactly one input symbol. For any π of this choice,

$$D(P_{Y|X}||Q_Y^n) = -nD(P_Y||Q_Y) - \frac{1}{2}\log n + \sum_{i:\pi_i>0}\frac{1}{2}\log \pi_i n + c + \mathbb{E}[\log \mathbb{P}[A=1|Y^n]|A=1] \tag{5.168}$$

$$= -nD(P_Y||Q_Y) - \frac{1}{2}\log n + \sum_{i:\pi_i>0}\frac{1}{2}\log \pi_i n + c \tag{5.169}$$

$$\leq -nD(P_Y||Q_Y) - \frac{1}{2}\log n + \frac{q+1}{2}\frac{1}{2}\log n + c \tag{5.170}$$

$$\leq -nD(P_Y||Q_Y) + \frac{q-1}{4}\log n + c\,. \tag{5.171}$$

If π of this form is the worst case π to use, meaning it gives the largest possible value of $D(P_{Y|X}\circ U\|Q_{Y^n})$ for any Q_Y, then we get that

$$C_{\mathsf{perm}}(P_{Z|X}) \leq \frac{q-1}{2} + \frac{q-1}{4} = \frac{3(q-1)}{4}\,. \tag{5.172}$$

If there is another π which is the worst, then our upper bound on the capacity will be larger than the value on the right side of (5.172). In either case, there is a gap between our upper bound for the capacity and the lower bound of $(q-1)/2$ given by (5.3).

5.6 Some Concluding Remarks

Our overall solution for finding converse bounds to the noisy permutation channel is to use divergence covering to upper bound mutual information followed by KL divergence results on the specific distributions which occur as outputs of the noisy permutation channel. The approach to find the converse bound in [3] is to

use Fano's inequality to bound mutual information and to compute the necessary quantities using entropies for binomial distributions. We note that both these approaches avoid the issue of needing to understand convergence rates for multidimensional central limit type problems.

The most straight-forward approach to finding a converse bound for the noisy permutation channel is to understand that each message, which can be sent through the noisy permutation channel, leads to a type at the output. This output type converges to its average value, and due to the intuition from multidimensional central limit theorem, we expect that the distribution of this output type has some variance when normalized appropriately. This variance is what limits the number of messages the transmitter can send. While dealing with 1-dimensional central limit theorem has corresponding results that show how fast the sums converge to the limiting cdf (particularly, Berry-Esseen), multidimensional results on how the sums converges are much harder. Part of this reason is that Fourier-type analysis (characteristic functions) can be used for showing Berry-Esseen type bounds for 1-dimensional variables, but this analysis does not hold as easily in the multidimensional case [50]. Without a strong enough result for convergence in the multidimensional case, we cannot take the straight-forward approach of using variance to find converse bounds.

The key to our divergence covering approach is that we have Theorem 7. The difficult distribution we cannot characterize the convergence rate of is the distribution of drawing balls out of an urn without replacement (hypergeometric). A much easier distribution to work with is the distribution of drawing balls out of an urn with replacement (multinomial), since each draw is iid. Theorem 7 shows that these distributions are close to one another as the length n increases. Thus, we can use the convergence rates of iid distributions to find the bounds we need, and side-step working with a difficult distribution.

Chapter 6

Average-Case Divergence Covering

In this chapter, we begin our journey of bounding the number of covering centers needed for average-case divergence covering. We dedicate two chapters for this exploration. In this chapter, we want to determine bounds on the average-case divergence covering when the prior W is a symmetric Dirichlet distribution. Our main technique for finding these bounds is to use rate distortion, treating each probability vector as a block and each symbol as a single letter. In the same spirit as Chapter 4, we also draw a connection between average-case divergence covering and the Bayes risk.

Chapter Organization We introduce the problem, with its motivation and connections to rate distortion in Section 6.1. We find the main lower bound in Theorem 11. We give the technique for the upper bound, which we name the *Interval Method*, in Section 6.4 and use this to find upper bounds in Section 6.5. We connect the rate distortion problem to average divergence covering in Section 6.6.

Notational Notes Since, the letter "D" appears frequently in this chapter as notation for various quantities, to clarify, we will use $D_{\mathrm{KL}}(P\|Q)$ to mean $D(P\|Q)$.

6.1 Introduction and Motivation

Suppose one wants to compress iid data over a large alphabet $[K]$ sampled from a distribution π. The distribution is considered to be known to the compressor (because it possesses a very large corpus of data) and unknown to the decompressor. A natural two-step compression scheme would be to first describe the distribution π and then use an optimal (Huffman, arithmetic, etc) compressor for it. Since π is to be represented with finitely many bits, only an approximation $\hat{\pi}$ can be conveyed to the decompressor. It is well known that this incurs penalty $D_{\mathrm{KL}}(\pi\|\hat{\pi})$ in compression length [1, Ch 5]. This motivates our definition for average-case divergence covering in Definition 3 in Chapter 1. The centers $\mathcal{Q}$ are the set of quantized distributions. We map each π to the element of $\mathcal{Q}$ closest[1] in KL divergence to π.

The quantity $M^*(W,\varepsilon)$ is intimately connected with Bayes risk or average redundancy (expressible as mutual information) in the universal compression setting.

Proposition 14. *Suppose that $\pi \sim W$ and X^n is generated iid from π. Then for any n,*

$$I(\pi; X^n) \le \inf_{\varepsilon} \left\{ n\varepsilon + \log M^*(W,\varepsilon) \right\} . \tag{6.1}$$

When $n = 1$, the above is equality.

The proof (shown in Section 6.7) is similar to techniques used in [6, 51]. Known results on $I(\pi; X^n)$ imply lower bounds on $M^*(W,\varepsilon)$ (discussed in Section 6.1.1 and more in Section 6.7).

[1] This is the ideal goal. Both in practice and in our analysis, we do not necessarily know which element of $\mathcal{Q}$ is closest. We usually just map to a center which is close enough.

Perhaps the most natural prior W to consider, in the absence of other knowledge, is the uniform prior which belongs to the family of symmetric Dirichlet distributions. Symmetric Dirichlet distributions only need to be specified with one scalar parameter α and an alphabet size K. We will use the notation $\mathrm{Dir}_K(\alpha)$ with scalar α to specify the symmetric Dirichlet on $[K]$ with all parameters set to α; the uniform distribution is $\mathrm{Dir}_K(1)$. Another important prior is Jeffreys' prior[21] which is $\mathrm{Dir}_K(1/2)$.

Dirichlet distributions are popular choices for modeling unknown discrete distributions in Bayesian statistics partly because they are the conjugate priors to multinomial distributions; they are used in many fields, including Dirichlet Processes [52] and other learning and estimation problems, such as in [53, 54, 55].

In [53], the authors use Dirichlet distributions for topic modeling under a hierarchical Bayesian model. Our problem fits nicely with this application. Each document has a probability over topics and each topic has a probability over words, resulting in many probability distributions which need to be saved. To save space on storage, all these probabilities can be quantized, reducing space requirements while keeping distortion (measured by KL divergence) low. The stored probabilities will then be well suited for tasks like compression or prediction with log-loss.

Our main result is the following:

Theorem 10. *For each α, there exists a constant $c_1(\alpha) > 0$ such that for all K and $\varepsilon > 0$,*

$$\log M^*(Dir_K(\alpha), \varepsilon) \geq \frac{K}{2} \log \frac{c_1(\alpha)}{\varepsilon + \frac{1}{2\alpha K}} \,. \tag{6.2}$$

Furthermore, there exists a $c_2(\alpha) > 0$ such that for all $\varepsilon > 0$ and all K such that $\alpha K > 1$

$$\log M^*(Dir_K(\alpha), \varepsilon) \leq \frac{K}{2} \log \frac{c_2(\alpha)}{\varepsilon(1 - (\alpha K)^{-1/3})^2} \,. \tag{6.3}$$

A practically important consequence of our analysis is that an almost optimal quantization of large-alphabet distributions is obtained by companding $\pi_j \mapsto f(s\pi_j)$ (for some constant s), followed by the uniform (scalar) quantizer and projection back onto the simplex. For $\alpha = 1$, our method recommends

$$f(x) = \begin{cases} c(x/\tau)^{2/3} & \text{for } x \leq \tau \\ (1-c)\big(1 - \exp(-(x-\tau)/3)\big) + c & \text{for } x > \tau \end{cases} \tag{6.4}$$

where $\tau = 2.954$, $c = 0.664$. (We will use the expression for f in Chapter 7. We name it the Exponential Density Intervals (EDI) method for quantization.)

Our approach is to reduce the single-sample covering question to the following rate-distortion function (corresponding to an iid multi-sample covering question):

Definition 23. *For distribution P_X over $\mathbb{R}_{\geq 0}$ and $D > 0$,*

$$R(P_X, D) \stackrel{\triangle}{=} \inf_{P_{Y|X}} \{I(X;Y) : \mathbb{E}[Y] = \mathbb{E}[X],\, \mathbb{E}[d(X,Y)] \leq D, Y > 0\} \,, \tag{6.5}$$

where

$$d(x,y) \stackrel{\triangle}{=} x \log(x/y) \tag{6.6}$$

(for $x \geq 0$ and $y > 0$).

We call $d(x,y)$ the *divergence distortion*. While divergence distortion is convex in both inputs and $d(x,y) = 0$ iff $x = y$, it has the unusual property of being negative when $x < y$ (with minimum at $x = y/e$). However, from $\mathbb{E}[Y] = \mathbb{E}[X]$ and convexity we still get $\mathbb{E}[d(X,Y)] \geq 0$. Note that without the constraint $\mathbb{E}[Y] = \mathbb{E}[X]$, $R(P_X, D)$ would be 0 trivially by increasing Y without bound.

It turns out (Section 6.6) that computing (6.5) with P_X being the Gamma distribution yields bounds on $M^*(W, \varepsilon)$ with $W = \mathrm{Dir}_K(\alpha)$ (and exponential P_X corresponds to uniform W).

Other work on quantizing probabilities includes [56], where they are quantized using the average L_p norm, and [57]. Particularly relevant is [5] which connects rate distortion to ε-nets, inspiring subsequent works on quantizing functions [58, 59].

In the remainder of this section, we will give a summary of the connection to universal compression, review the background of rate distortion and summarize the analysis in this paper.

6.1.1 Universal Compression

Much work has been done in the setting of universal compression on the asymptotic Bayes risk or average redundancy defined in [12, 23], mostly for the case where W is the least-favorable prior for maximin redundancy, the Jeffreys' prior (or $W = \mathrm{Dir}_K(1/2)$ for discrete alphabets) [21]. The dominant term of the Bayes risk is $\frac{K-1}{2} \log n$, where $K-1$ is the dimension of the parameter space and n is the number of samples [19, 60]; the other terms are constant or order $o(1)$ in n [61, 20, 21, 13, 22].

Based on the above results on Jeffreys' prior and Proposition 14, we conjecture that in general $\log M^*(W, \varepsilon)$ will have a dominant term of $\frac{K-1}{2} \log \frac{1}{\varepsilon}$ as $\varepsilon \to 0$. This is close to our upper bound from Theorem 10, especially for large alphabets (our bound behaves as $\frac{K}{2} \log \frac{1}{\varepsilon}$ as $\varepsilon \to 0$). An interesting direction for future work would be to close the gap between our upper and lower bound, using the conjectured asymptotic rate as a guide. The advantage of our results, however, is in yielding an explicit non-asymptotic upper bound on the Bayes risk. More details are discussed in Section 6.7.

6.1.2 Rate Distortion Background

For rate distortion problems which quantize X to Y, points in the support of Y are called *reconstruction points*. We typically use Q_Y to denote the pdf (probability density function) or pmf (probability mass function) of the reconstruction points. The probabilities with which each X maps to each Y are the transition probabilities. We usually use $Q_{Y|X}$ or $P_{Y|X}$ to specify this. The transition probabilities are optimal if they minimize $I(X;Y)$. We refer to P_X as the *source distribution*.

Reconstruction Points

Reconstruction points can be either discrete or continuous. The rate distortion problem for Gaussian sources and squared-error distortion gives continuous reconstruction points, but this is not the usual case for other problems. Even for squared-error distortion, the optimal reconstruction is discrete if the source distribution is not supported on the whole real line [62]. There is sometimes a transition point at a critical distortion, where for all D below the critical point, the reconstruction is continuous, and above the critical point it is discrete [63]. For other distortion measures the reconstruction can sometimes be continuous, such as for Gamma distributions with the Itakura-Saito measure. For a summary of results on several distortion measures on continuous sources, see [64].

Shannon Lower Bound (SLB)

Rate distortion functions are difficult to compute in general. For many rate distortion problems, the SLB is used for lower-bounding and even approximating $R(D)$. The concept behind the SLB is to minimize the mutual information in the rate distortion problem by maximizing the entropy $H(X|Y=y)$ (or differential entropy $h(X|Y=y)$) for a particular y. We get a lower bound by relaxing the constraint that $Q_{X|Y}$ needs to reproduce the original source distribution. This allows us to have a lower bound which depends on the source only through a $h(X)$ (differential entropy) term. See [65] for more on the SLB.

Several works show that asymptotically as the distortion D goes to zero, the SLB can be tight under various conditions [66, 67, 68].

When working with squared-error distortion, the SLB coincides with $R(D)$ if and only if the source X can be written as $X = Y + N$, where Y and N are independent and Y is the reproduction (or the "backward channel" condition). The reconstruction points are continuous when the SLB is met. Otherwise, the reconstruction points are discrete. [63].

Computing Upper Bounds

For large distortions, we can numerically compute the rate distortion function using Arimoto-Blahut [69] or [63]. The latter avoids the need for infinite discretization which Arimoto-Blahut would require. (The approach in [63] still needs discretization, but instead discretizes the unit interval.) Such algorithms are still complicated and intensive when dealing with continuous sources, especially for smaller and smaller distortion values. For small distortions, while not optimal, scalar quantization is useful for computing analytic upper bounds. In particular, [70] showed that quantizing to intervals is asymptotically optimal for the squared-error distortion measure. We will use a similar technique for our upper bounds.

Related Work

The following works solve the rate distortion problem for distortions measures which have similarities to the divergence distortion. In [71], the authors look at distortion measures which are a function of the quotient X/Y. Using this result and then taking $U = \log X$ and $V = \log Y$ gives results for other distortion measures. Their results show that for the Itakura-Saito measure the worst-case sources are Gamma distributions. In [72], the authors solve the rate distortion function using the absolute magnitude measure, $d(x, y) = |x - y|$. For their proof, their main technique is to use the SLB. In [64], the authors solve for the rate distortion function for Gamma sources using the distortion measure $d(x, y) = |\log(x) - \log(y)|$.

6.1.3 Summary

In Section 6.2 we show some preliminary results for problem (6.5). In Section 6.3 we show the following lower bound on (6.5) for continuous sources (i.e. described by density functions):

Theorem 11. *For any X with density $p_X(x)$, if $D \leq \mathbb{E}[X]$ we have*

$$R(p_X, D) \geq \frac{1}{2}\log\left(\frac{1}{D}\right) + h(X) - \frac{1}{2}\log \mathbb{E}[X] + c_L \tag{6.7}$$

where $h(\cdot)$ is the differential entropy and c_L is a constant.[a]

[a] The $-(1/2)\log \mathbb{E}[X]$ term might seem strange in light of the intuition that scaling X up will make it harder to approximate with low distortion (see Proposition 15). However, scaling X up increases $h(X)$ and the net effect is positive.

In Section 6.4 we describe a certain method, called the Interval Method, based on the technique of [70], for upper-bounding (6.5), as well as lemmas allowing easy manipulation of bounds using this method. In Section 6.5, we use the tools from Section 6.4 to give upper bounds for the following important sources:

Theorem 12. *For a uniform source $X \sim \text{Unif}_{[a,b]}$,*

$$R(p_X, D) \leq \frac{1}{2}\log\left(\frac{1}{D}\right) + \frac{1}{2}\log\left(\frac{(b-a)^2}{b}\frac{9}{32}\right) \tag{6.8}$$

Theorem 13. *For $X \sim \text{Exponential}(1)$,*

$$R(p_X, D) \leq \frac{1}{2}\log\left(\frac{1}{D}\right) + \frac{1}{2}\log(9) \leq \frac{1}{2}\log\left(\frac{1}{D}\right) + 1.1 \tag{6.9}$$

The exponential distribution is also a Gamma distribution. We can use the results from the Interval Method on uniform sources and exponential sources to get a result on general Gamma sources.

Theorem 14. *For source $X \sim \text{Gamma}(\alpha, \beta)$ for any α, β,*

$$R(p_X, D) \le \frac{1}{2} \log\left(\frac{1}{D}\right) + c(\alpha, \beta) \tag{6.10}$$

For all of these, our upper and lower bounds are tight up to an additive constant. This shows that $\frac{1}{2}\log\left(\frac{1}{D}\right)$ is the correct rate of growth for these sources under divergence distortion. Equating D with ε approximately gives the size of $M^*(\text{Dir}_K(\alpha), \varepsilon)$. We show this precisely in Section 6.6, where we use Theorem 11 and Theorem 14 to derive Theorem 10.

6.2 Preliminaries

We first give some preliminary results about the rate distortion function in (6.5).

Proposition 15 (Scaling)**.** *Suppose $X' = cX$ for $c > 0$, then $R(p_{X'}, cD) = R(p_X, D)$.*

Proof. Let $q_{Y|X}$ satisfy the constraints of the rate distortion problem on p_X with distortion bound D (i.e. $\mathbb{E}[X \log(X/Y)] \le D$ and $\mathbb{E}[Y] = \mathbb{E}[X]$ when $X \sim p_X$ and $Y \sim q_{Y|X}$). Then we use

$$q'_{Y'|X'}(y|x) = \frac{1}{c} q_{Y|X}(y/c|x/c) \tag{6.11}$$

This is equivalent to: given X', first extract $X = X'/c$, then produce Y given X as in $q_{Y|X}$, and then produce $Y' = cY$. It is also equivalent to generating X, then Y from X according to q, and finally extracting $X' = cX$ and $Y' = cY$. Thus,

$$\mathbb{E}_{X',Y' \sim p_{X'}, q'_{Y'|X'}}\left[X' \log\left(\frac{X'}{Y'}\right)\right] \tag{6.12}$$

$$= \mathbb{E}_{X,Y \sim p_X, q_{Y|X}}\left[cX \log\left(\frac{cX}{cY}\right)\right] \le cD \tag{6.13}$$

and so $q'_{Y'|X'}$ achieves distortion $\le cD$ with $p_{X'}$. Furthermore, it is trivial that $\mathbb{E}[Y] = \mathbb{E}[X] \implies \mathbb{E}[Y'] = \mathbb{E}[X']$, since $Y' = cY$ and $X' = cX$.

Then, since scaling the inputs of mutual information does not affect the result, we have

$$I(X'; Y') = I(cX; cY) = I(X; Y) \tag{6.14}$$

Thus, any value of the mutual information achievable on X with distortion $\le D$ is also achievable on $X' = cX$ with distortion $\le cD$, and since

$$R(p_X, D) = \inf_{q_{Y|X}} I(X; Y) \tag{6.15}$$

$$\text{s.t. } \mathbb{E}_{X,Y \sim p_X, q_{Y|X}}\left[X \log\left(\frac{X}{Y}\right)\right] \le D \text{ and } \mathbb{E}[Y] = \mathbb{E}[X] \tag{6.16}$$

we can conclude that $R(p_{X'}, cD) \le R(p_X, D)$. To get the converse, $R(p_X, D) \le R(p_{X'}, cD)$, we use the exact same argument, noting that if $c' = 1/c$ then $X = c'X'$ and $D = c'(cD)$, and conclude that $R(p_X, D) = R(p_{X'}, cD)$ as we wanted. □

The scaling property should fit with our intuition that quantizing variables X which are larger should be harder to achieve smaller distortion. For the rate to be the same, we must allow the larger variables a higher distortion.

Proposition 16 (Discrete Reconstruction)**.** *For any source probability density p_X where $\mathbb{E}[X] < \infty$, the optimal reconstruction for* (6.5) *is discrete.*

This proof follows the analysis of [63] except with some modifications to fit the divergence distortion and its constraints. We give in the proof in Appendix B.2.2.

Proposition 17. *For any source probability density p_X where $\mathbb{E}[X] < \infty$, the optimal reconstruction for (6.5) is discrete, and*

$$y_i = \frac{\int p_X(x) q_{Y|X}(y_i|x)\, x\, dx}{\int p_X(x) q_{Y|X}(y_i|x)\, dx} = \mathbb{E}[X|X \text{ maps to } y_i] \tag{6.17}$$

where $q_{Y|X}(\cdot|\cdot)$ are the optimal transition probabilities for (6.5).

There are multiple ways of proving this proposition. We present just one below.

Proof. Fix $\mathbb{E}[X] = 1$. We can scale otherwise.

Since we know the reconstruction must be discrete from Proposition 16, list the discrete values as $y_1, y_2, \ldots$. We construct a joint probability on X and J where J is the random index y_J which X maps to. Let $p_J(i)$ be the probability $J = i$.

Suppose $Y = w_i$ if $J = i$. Let $Z = z_i$ be the quantity defined in (6.17) if $J = i$. We have the constraint that $\mathbb{E}[Y] = \mathbb{E}[X]$ which implies

$$1 = \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|J]] = \sum_i p_J(i) w_i \,. \tag{6.18}$$

Similarly, $\sum_i p_J(i) z_i = 1$ Then

$$\begin{aligned}
&\mathbb{E}\left[X \log \frac{X}{Y}\right] - \mathbb{E}\left[X \log \frac{X}{Z}\right] && (6.19)\\
&= \mathbb{E}\left[X \log \frac{Z}{Y}\right] && (6.20)\\
&= \mathbb{E}\left[\mathbb{E}\left[X \log \frac{Z}{Y}\middle| J\right]\right] && (6.21)\\
&= \mathbb{E}\left[w_J \log \frac{w_J}{z_J}\middle| J\right] && (6.22)\\
&= \sum_i p_J(i) w_i \log \frac{p_J(i) w_i}{p_J(i) z_i} && (6.23)\\
&\geq 0 && (6.24)
\end{aligned}$$

where equality holds iff $w_i = z_i$ □

6.3 Lower Bound

Our method for finding a lower bound to (6.5) is to use the Shannon Lower Bound. First, we simplify by combining the distortion measure and expected value constraints into one inequality. We will relax (6.17) so that $\int_0^\infty q_{X|Y}(x|y_i) x\, dx \geq y_i$. For $q_{Y|X}(y|x)$ and $p_X(x)$, let $q_Y(y)$ be the resulting marginal on y. (The joint distribution can be expressed either as $q_Y(y) q_{X|Y}(x|y)$ or $p_X(x) q_{Y|X}(y|x)$.) We add the constraints together to get

$$\int_y \int_x q_Y(y) q_{X|Y}(x|y) \left(x \log \frac{x}{y} - x\right) dx\, dy \leq D - \mathbb{E}[Y] \tag{6.25}$$

Our new distortion for finding a lower bound is

$$d_{\text{SLB}}(x, y) = x \log(x/y) - x \,. \tag{6.26}$$

We will start with the following proposition, which is a slight variation on a result by Berger in [65, Ch 4]:

Proposition 18. *Let $p(x)$ be a probability density on X and $d(\cdot,\cdot)$ be a distortion measure. Let $\mathcal{A}_\lambda$ be the set of all nonnegative functions $\alpha_\lambda(x,y)$ satisfying*

$$c(y) = \int_0^\infty \alpha_\lambda(x,y)p(x)e^{-\lambda d(x,y)}dx \leq 1 \tag{6.27}$$

for all y. For all $D > 0$, suppose $\mathcal{B}$ is the set of conditional probabilities meeting $\mathbb{E}[d(X,Y)] \leq D$. Then

$$R(p,D) \geq \inf_{Q_{Y|X}\in\mathcal{B}} \sup_{\lambda\geq 0,\alpha_\lambda\in\mathcal{A}_\lambda} \left(-\lambda D + \int_y \int_x Q_{Y|X}(y|x)p(x)\log\alpha_\lambda(x,y)\,dx\,dy \right) \tag{6.28}$$

The proof is in Appendix B.2.3. The nice property of Proposition 18 is that it allows us to choose any nonnegative function $\alpha_\lambda(x,y)$ and relate it to a lower bound on rate distortion. To make computations easier, the Shannon Lower Bound chooses $\alpha_\lambda(x,y)$ to be the reciprocal of $p(x)$. We want to do this, but to make sure condition (6.27) is satisfied, we need to make sure that if we pick

$$\alpha_\lambda(x,y) = \frac{1}{p(x)\zeta} \tag{6.29}$$

that ζ (a value we choose to be independent of x) is set appropriately. To find a good value for ζ we will compute the integral in the following lemma.

Lemma 15. *For any $y > 0$ and any $\lambda > 0$,*

$$\int_0^\infty e^{-\lambda\left(x\log\frac{x}{y}-x\right)}dx \leq \frac{3}{2}e^{\lambda y}\sqrt{\frac{y}{\lambda}\frac{2\pi}{1-2c}} + \frac{1}{\lambda} \tag{6.30}$$

where $c = 2\log 2 - 3/2$.

We need a bound which holds even as $\lambda \to \infty$. We will show this using a modified version of Laplace's method. The key is to show that the integrand is Gaussian-like for $x < ey$ and is exponential-like for $x > ey$. When we integrate in (6.30), we get two terms, one related to the Gaussian part and one related to the exponential part.

Proof. Let

$$f_y(x) \stackrel{\triangle}{=} e^{-\lambda\left(x\log\frac{x}{y}-x\right)}. \tag{6.31}$$

We will compute the integral (6.30) by splitting up the domain into a few parts. The main focus will be to bound the portion where x is in $[0,2y)$. We will use that result to bound $[2y,3y)$. The interval $[3y,\infty)$ we will bound separately at the end.

Bound for $x \in [0,2y)$ First, observe that $\max_x f_y(x) = f_y(y) = e^{\lambda y}$. For the exponent part, the Taylor expansion around y is

$$x\log\frac{x}{y} - x = -y + \frac{(x-y)^2}{2y} + \sum_{n=1}^\infty \frac{(-1)^n(x-y)^{n+2}}{(n+1)(n+2)y^{n+1}} \tag{6.32}$$

$$= -y + \frac{(x-y)^2}{y}\sum_{n=0}^\infty \frac{(-1)^n}{(n+1)(n+2)}\left(\frac{x-y}{y}\right)^n \tag{6.33}$$

and this Taylor expansion expression has a radius of convergence on $(0,2y)$.

For any $-1 < r < 1$, we have that

$$g(r) \stackrel{\triangle}{=} \sum_{n=0}^\infty \frac{(-1)^n}{(n+1)(n+2)}r^n = \frac{-r+(r+1)\log(r+1)}{r^2}. \tag{6.34}$$

For $r \in [-1, 1]$,

$$g(r) \geq 2\log 2 - 1 \tag{6.35}$$

$$\geq \frac{1}{2} - c \tag{6.36}$$

where $c = -\frac{3}{2} + 2\log 2 \approx 0.11$. Notice also that (in the limit) $g(0) = 1/2$. Writing

$$x \log \frac{x}{y} - x = -y + \frac{(x-y)^2}{y} g\left(\frac{x-y}{y}\right) \tag{6.37}$$

$$x \log \frac{x}{y} - x \geq -y + \frac{(x-y)^2}{y} \frac{1}{2}(1-2c) \tag{6.38}$$

$$-\lambda\left(x \log \frac{x}{y} - x\right) \leq \lambda y - \lambda \frac{(x-y)^2}{y} \frac{1}{2}(1-2c) \tag{6.39}$$

we get

$$\int_0^{2y} e^{-\lambda\left(x \log \frac{x}{y} - x\right)} dx \leq \int_0^{2y} e^{\lambda y} e^{-\lambda\left(\frac{1}{2}\frac{(x-y)^2}{y}(1-2c)\right)} dx \tag{6.40}$$

$$< \int_{-\infty}^{\infty} e^{\lambda y} e^{-\lambda\left(\frac{1}{2}\frac{(x-y)^2}{y}(1-2c)\right)} dx \tag{6.41}$$

$$= e^{\lambda y} \sqrt{\frac{y}{\lambda} \frac{2\pi}{1-2c}}\,. \tag{6.42}$$

Bound for $x \in [2y, 3y)$ Since $f_y(x)$ decreases on $(y, 3y)$, we can upper bound the integral of $f_y(x)$ on $[2y, 3y]$ with the integral of $f_y(x)$ on $[y, 2y]$. Combining the two bounds together gives

$$\int_0^{3y} e^{-\lambda\left(x \log \frac{x}{y} - x\right)} dx \leq \frac{3}{2} e^{\lambda y} \sqrt{\frac{y}{\lambda} \frac{2\pi}{1-2c}}\,. \tag{6.43}$$

Bound for $x \in [3y, \infty)$ Note that $f_y(ey) = 1$.

For $x \in [3y, \infty) \subset [ey, \infty)$, the function $f_y(x) \leq e^{-\lambda(x-ey)}$. Once we have this, we can bound the integral

$$\int_{ey}^{\infty} e^{-\lambda\left(x \log \frac{x}{y} - x\right)} dx \leq \int_{ey}^{\infty} e^{-\lambda(x-ey)} dx \tag{6.44}$$

$$\leq \int_0^{\infty} e^{-\lambda x} dx \tag{6.45}$$

$$= \frac{1}{\lambda}\,. \tag{6.46}$$

To verify that $f_y(x) \leq e^{-\lambda(x-ey)}$ on $x \in [ey, \infty)$, we can use derivatives to check that

$$x \log \frac{x}{y} - x \geq x - ey \tag{6.47}$$

on this domain. Notice at $x = ey$,

$$x \log \frac{x}{y} - x = ey \log \frac{ey}{y} - ey = ey - ey = x - ey\,. \tag{6.48}$$

The derivative of $x - ey$ is 1. The derivative of $x \log \frac{x}{y} - x$ is $\log \frac{x}{y}$ which for $x \geq ey$ is always greater than or equal to 1. This shows (6.47).[2] □

[2]Our result in Lemma 15 is loose mostly from the crude upper bounding used in part b). Comparing to some selected values computed with numerical integration, our bound is off by less than a factor of 3/2.

To get our lower bound, we just need to combine the results and do the calculations.

Proof of Theorem 11. For this proof we assume $\mathbb{E}[X] = 1$ so that $D < 1$ based on the conditions and also $\log \mathbb{E}[X] = 0$. For $\mathbb{E}[X] \neq 1$ (but still finite) we can scale the source and the optimal reconstruction will scale with it to give the extra $\frac{1}{2} \log \mathbb{E}[X]$ term. We will discuss this scaling at the end. [3]

We will use Proposition 18 with (6.26) as the distortion measure. Let

$$\zeta(y, \lambda) \stackrel{\triangle}{=} Ce^{\lambda y} \sqrt{\frac{y}{\lambda}} + \frac{1}{\lambda} \geq \int_0^\infty e^{-\lambda d_{\mathrm{SLB}}(x,y)} dx \tag{6.49}$$

where we pick C appropriately using Lemma 15. Fix some y and some λ, we will choose

$$\alpha_\lambda(x, y) = \frac{1}{p_X(x)\zeta(y, \lambda)} \,. \tag{6.50}$$

It follows from Lemma 15 that

$$c(y) = \int_0^\infty \alpha_\lambda(x, y) p(x) e^{-\lambda d_{\mathrm{SLB}}(x,y)} dx \leq 1 \,. \tag{6.51}$$

Let $\mathcal{Q}(D)$ be the set of conditional probabilities meeting the constraints and let $q_{Y|X} \in \mathcal{Q}(D)$. Then $q_{Y|X}$ must satisfy $D - \mathbb{E}[Y] = \mathbb{E}[d_{\mathrm{SLB}}(x, y)]$, which is the distortion when applying Proposition 18. Define

$$f(D, \lambda, q_{Y|X}) \tag{6.52}$$

$$\stackrel{\triangle}{=} -\lambda D + \lambda \mathbb{E}[Y] + \int_x \int_y q_{Y|X}(y|x) p(x) \log \alpha_\lambda(x, y) \, dx \, dy \tag{6.53}$$

$$= -\lambda D + \lambda \mathbb{E}[Y] + h(X) - \int_y q_Y(y) \log \left(\zeta(y, \lambda) \right) \, dx \, dy \tag{6.54}$$

where $q_Y(y) = \int_x p_X(x) q_{Y|X}(y|x) \, dx$.

Next, we will upper bound

$$\zeta(y, \lambda) \leq 2 \max \left\{ Ce^{\lambda y} \sqrt{\frac{y}{\lambda}}, \frac{1}{\lambda} \right\} . \tag{6.55}$$

Define

$$A_{\mathrm{small}} \stackrel{\triangle}{=} \left\{ y : Ce^{\lambda y} \sqrt{\frac{y}{\lambda}} < \frac{1}{\lambda} \right\} \tag{6.56}$$

$$A_{\mathrm{large}} \stackrel{\triangle}{=} \left\{ y : Ce^{\lambda y} \sqrt{\frac{y}{\lambda}} \geq \frac{1}{\lambda} \right\} \tag{6.57}$$

$$p_{\mathrm{small}} \stackrel{\triangle}{=} \int_{y \in A_{\mathrm{small}}} q_Y(y) \, dy \tag{6.58}$$

$$p_{\mathrm{large}} \stackrel{\triangle}{=} \int_{y \in A_{\mathrm{large}}} q_Y(y) \, dy \tag{6.59}$$

[3] We still choose to write $\mathbb{E}[Y]$ in the proof even though it equals 1 so its clearer that it relates to quantities like conditional expectations on Y.

These quantities depend on λ and $q_{Y|X}$, though we do not explicitly write these dependencies. Then

$$f(D,\lambda,q_{Y|X}) \tag{6.60}$$

$$\geq -\lambda D + \lambda\mathbb{E}[Y] + h(X) - p_{\text{small}}\log\frac{2}{\lambda} - \int_{y\in A_{\text{large}}} q_Y(y)\log\left(2Ce^{\lambda y}\sqrt{\frac{y}{\lambda}}\right)dy \tag{6.61}$$

$$= -\lambda D + \lambda\mathbb{E}[Y] + h(X) - p_{\text{small}}\log\frac{2}{\lambda}$$
$$- p_{\text{large}}\log 2C - \int_{y\in A_{\text{large}}} q_Y(y)\lambda y\,dy - \int_{y\in A_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy - p_{\text{large}}\frac{1}{2}\log\frac{1}{\lambda} \tag{6.62}$$

$$\geq -\lambda D + h(X) + C' - p_{\text{small}}\log\frac{2}{\lambda} + p_{\text{small}}\mathbb{E}[\lambda Y|Y\in A_{\text{small}}] - \int_{y\in A_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy - p_{\text{large}}\frac{1}{2}\log\frac{1}{\lambda} \tag{6.63}$$

$$\geq -\lambda D + h(X) + C' - p_{\text{small}}\log\frac{2}{\lambda} - p_{\text{large}}\frac{1}{2}\log\frac{1}{\lambda}$$
$$- \int_{y\in A_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy\,. \tag{6.64}$$

We take out the term $p_{\text{small}}\mathbb{E}[\lambda Y|Y\in A_{\text{small}}]$ since it is positive. We will need to evaluate the last term in the sum. Using Jensen's inequality, we have

$$-\int_{y\in A_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy = \frac{1}{2}p_{\text{large}}\mathbb{E}[-\log Y|A_{\text{large}}] \tag{6.65}$$

$$\geq \frac{1}{2}p_{\text{large}}(-\log\mathbb{E}[Y|A_{\text{large}}])\,. \tag{6.66}$$

Then since $\mathbb{E}[Y] = p_{\text{small}}\mathbb{E}[Y|A_{\text{small}}] + p_{\text{large}}\mathbb{E}[Y|A_{\text{large}}]$ and $Y\geq 0$

$$\mathbb{E}[Y|A_{\text{large}}] = \frac{\mathbb{E}[Y] - p_{\text{small}}\mathbb{E}[Y|A_{\text{small}}]}{p_{\text{large}}} \tag{6.67}$$

$$\leq \frac{\mathbb{E}[Y]}{p_{\text{large}}}\,. \tag{6.68}$$

Recall that we assumed $\mathbb{E}[Y] = \mathbb{E}[X] = 1$.

$$-\int_{y\in A_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy \geq \frac{1}{2}p_{\text{large}}(-\log p_{\text{large}}) \tag{6.69}$$

$$\geq -\frac{1}{2e}\,. \tag{6.70}$$

Hence $-\int_{y\in A_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy$ can be bounded by a constant. Select $\lambda = 1/D$. Then using Proposition 18

$$\inf_{Q_{Y|X}\in\mathcal{Q}(D)} R(D,p_X) \geq \inf_{Q_{Y|X}\in\mathcal{Q}(D)} f(D,1/D,q_{Y|X}) \tag{6.71}$$

$$\geq \inf_{Q_{Y|X}\in\mathcal{Q}(D)} h(X) + c_L + p_{\text{small}}\log\frac{1}{D} + p_{\text{large}}\frac{1}{2}\log\frac{1}{D} \tag{6.72}$$

$$\geq h(X) + c_L + \frac{1}{2}\log\frac{1}{D}\,. \tag{6.73}$$

where

$$c_L = -1 - p_{\text{large}}\log\left(3\sqrt{\frac{2\pi}{1-2c}}\right) - 1/2p_{\text{large}}\log p_{\text{large}} - p_{\text{small}}\log 2 \tag{6.74}$$

$$\geq -3.84\,. \tag{6.75}$$

Now suppose we want to find a rate distortion lower bound for X' which has $\mathbb{E}[X'] = a$. Let random variable X be so that $X' = aX$ and thus $\mathbb{E}[X] = 1$. Then by Proposition 15,

$$R(p'_X, D) = R(p'_X, (D/a)a) \tag{6.76}$$
$$= R(p_X, D/a) \tag{6.77}$$
$$= h(X) + c_L + \frac{1}{2}\log\frac{1}{D/a} \tag{6.78}$$
$$= h(X) + c_L + \frac{1}{2}\log\frac{1}{D} + \frac{1}{2}\log a \tag{6.79}$$

We know that $p_{X'}(x) = \frac{1}{a}p_X(x/a)$.

$$h(X') = \int p'_X(x)\log\frac{1}{p_{X'}(x)}dx \tag{6.80}$$
$$= \int \frac{1}{a}p_X(x/a)\log\frac{1}{\frac{1}{a}p_X(x/a)}dx \tag{6.81}$$
$$= \int \frac{1}{a}p_X(u)\log\frac{1}{\frac{1}{a}p_X(u)}a\,du \tag{6.82}$$
$$= h(X) + \log a \tag{6.83}$$

where we substituted $u = x/a$. This gives

$$R(p'_X, D) = h(X') - \log a + c_L + \frac{1}{2}\log\frac{1}{D} + \frac{1}{2}\log a \tag{6.84}$$
$$= h(X') + c_L + \frac{1}{2}\log\frac{1}{D} - \frac{1}{2}\log\mathbb{E}[X]\,. \tag{6.85}$$

□

6.4 Interval Method

We develop the *Interval Method* for upper-bounding (6.5), which is an instance of scalar quantization, similar to the technique of Gish and Pierce [70]. The Interval Method partitions the support of X into n intervals I_j. We set Y by interval, i.e. $(Y|X \in I_j) = y_j \triangleq \mathbb{E}[X|X \in I_j]$ (we refer to y_j as the *center* that $X \in I_j$ maps to). Note that this assigns Y from X in a deterministic way, i.e. $P_{Y|X}(\cdot|\cdot) \in \{0, 1\}$, which is not necessarily optimal but simplifies the analysis. The distortion of any such quantization must, of course, upper bound the minimum possible distortion. We present some definitions and lemmas which help us derive upper bounds using the Interval Method.

Definition 24. *For an interval $I \subseteq \mathbb{R}_{\geq 0}$, let $\mathcal{F}_I$ be the set of functions $p : I \to \mathbb{R}_{\geq 0}$ such that $\int_I p(x)\,dx < \infty$ and $\int_I p(x)x\,dx < \infty$.*

Since we will upper bound some probability density functions, it is important that this definition include functions that do not integrate to 1.

Definition 25. *We define for $p \in \mathcal{F}_I$ the values*

$$y^{(p,I)} \triangleq \frac{\int_I p(x)x\,dx}{\int_I p(x)\,dx} \tag{6.86}$$

and

$$D^{(p,I)} \triangleq \int_I p(x)x\log\left(\frac{x}{y^{(p,I)}}\right)dx\,. \tag{6.87}$$

We will write $y^{(p,I)}$ and $D^{(p,I)}$ even when p is defined outside of I as well, to mean $y^{(p|_I,I)}$ and $D^{(p|_I,I)}$ where $p|_I$ is the restriction of p to I.

Note that when p is a probability distribution (i.e. $\int p(x)dx = 1$), this definition is equivalent to $y^{(p,I)} = \mathbb{E}_{X\sim p}[X|X \in I]$. Therefore the Interval Method automatically satisfies $y_i = \mathbb{E}[X|X \text{ maps to } y_i]$ and $\mathbb{E}[Y] = \mathbb{E}[X]$ so we will ignore this constraint moving forward.

Definition 26. *The minimum distortion on intervals of density $p \in \mathcal{F}_I$ onto n centers is*

$$D(p,n) \triangleq \inf_{\{I_1,\dots,I_n\}} \sum_{j=1}^{n} D^{(p,I_j)} \tag{6.88}$$

where the infimum is taken over $I_1, \dots, I_n$ partitioning I.

The goal is to determine the intervals $I_1, ..., I_n$ in order to minimze $D(p,n)$. We will generally find intervals which are easy to perform calculations and get an upper bound on $D(p,n)$. The intervals will couple X to a discrete Y. Limiting Y to n values limits the mutual information $I(X;Y)$ (since it is at most the entropy of Y, which is at most $\log(n)$ on n values), giving us:

Proposition 19. *Let X be a random variable with probability density p_X; f be a decreasing nonnegative function; and $b, c > 0$ be constants. Suppose for all $\varepsilon > 0$, the following holds for all sufficiently large n: (i) $D(p_X, n) \le c\, f(n)$ and (ii) $n^{-b-\varepsilon} \le f(n) \le n^{-b+\varepsilon}$. Then the rate distortion function satisfies*

$$R(p_X, D) \le \log(f^{-1}(D)) + \frac{1}{b}\log(c) + o(1) \tag{6.89}$$

(where f^{-1} is the inverse of f). Furthermore, if $f(n) = n^{-b}$, then

$$R(p_X, D) \le \frac{1}{b}\log\left(\frac{1}{D}\right) + \frac{1}{b}\log(c) \tag{6.90}$$

The proof for this is in Appendix B.2.4. Note that even if the constant c is not known, this proposition gives $R(p_X, D) \le \log(f^{-1}(D)) + O(1)$.

In principle, it may be possible to show better results in particular cases where $I(X;Y)$ is significantly smaller than $\log(n)$. However, the process of minimizing the distortion seems to produce Y which are close enough to uniform that $H(Y) = \log(n) - O(1)$; additionally, for simplicity we work with joint distributions where X determines Y, so $H(Y|X) = 0$. Thus, we get $I(X;Y) = \log(n) - O(1)$ and so Proposition 19 is the best one can do in these cases with the upper bound $D(p_X, n) = O(\alpha(n))$.

Next we discuss useful lemmas for determining $D(p,n)$.

Lemma 16. *For $p \in \mathcal{F}_I$ and constant $c > 0$ (where cI and $I + c$ scale and shift I by c, respectively), let $p_{\times c}$ and p_{+c} be p on cI and $I + c$ with the input scaled or shifted accordingly:*

- *$p_{\times c} : cI \to [0, \infty)$ such that $p_{\times c}(x) = p(x/c)$*
- *$p_{+c} : I + c \to [0, \infty)$ such that $p_{+c}(x) = p(x - c)$.*

Then if $p_{\times c} \in \mathcal{F}_{cI}$ and $p_{+c} \in \mathcal{F}_{I+c}$:

i. $D^{(cp,I)} = cD^{(p,I)}$;

ii. $D^{(p_{\times c},cI)} = c^2 D^{(p,I)}$;

iii. $D^{(p_{+c},I+c)} \le \frac{y^{(p,I)}}{y^{(p,I)}+c} D^{(p,I)}$ *if p is uniform on I*

Parts i) and ii) of this lemma can be combined to give a similar result to Proposition 15. For part iii), this lemma gives a bound on the distortion for translating an interval to the right, but not to the left. The

proof of part iii) will also show that for $c > 0$ and any p (and interval I with nonzero distortion),

$$D^{(p_{+c}, I_{+c})} < D^{(p,I)}\,. \tag{6.91}$$

So translation to the right always decreases the distortion of a given interval. For the proof, the first two parts of the lemma is straightforward. The third part is more involved. We give the proofs in Appendix B.2.5

The next lemma makes it possible to upper-bound the density function p_1 of a source by another function p_2 (which will not actually be a probability distribution since it will sum to more than 1), then upper-bound the "distortion" for the new function using the Interval Method, and immediately get an upper bound for the distortion of p_1 using the same intervals.

Lemma 17. *Let $p_1, p_2 \in \mathcal{F}_I$ for an interval $I \subseteq \mathbb{R}_{\geq 0}$. Then*

$$p_1 \leq p_2 \implies D^{(p_1,I)} \leq D^{(p_2,I)}\,. \tag{6.92}$$

The principal difficulty is that p_1 and p_2 can have different averages $y^{(p_1,I)}, y^{(p_2,I)}$. We need to show that changing these averages does not decrease the distortion. The following proof sketch removes some computation steps. The same proof but with all computation steps is in Appendix B.2.6.

Proof Sketch. We prove this by showing that for any $p \in \mathcal{F}_I$, adding a tiny bit of mass (however distributed as long as its nonnegative everywhere) cannot decrease the distortion. Then the result follows because p_2 can be generated from p_1 by smoothly adding mass (e.g. adding $\lambda(p_2 - p_1)$ and taking λ from 0 to 1 continuously) and at no point can the distortion be decreasing.

We define for $p \in \mathcal{F}_I$, any $z \in \mathcal{F}_I$ (WLOG $\int_I p(x)\,dx = \int_I z(x)\,dx = 1$ because we can scale) and $\varepsilon > 0$ the distribution $p_\varepsilon \triangleq p + \varepsilon z$; our objective will be to show that for all such p, z, $\left[\frac{d}{d\varepsilon} D^{(p_\varepsilon, I)}\right]_{\varepsilon=0} \geq 0$. We accomplish this by looking at how adding εz moves the average, defining $y^* \triangleq y^{(z)} - y^{(p,I)}$; then (thanks to normalization making things easier),

$$y^{(p_\varepsilon)} = y^{(p,I)}\Big(1 + \frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p,I)}}\Big) \tag{6.93}$$

Then we can take the derivative

$$\frac{d}{d\varepsilon} D^{(p_\varepsilon)} = \frac{d}{d\varepsilon}\int_I p(x) x \log\Big(\frac{x}{y^{(p_\varepsilon)}}\Big)\,dx \tag{6.94}$$

$$+ \frac{d}{d\varepsilon}\varepsilon\int_I z(x) x \log\Big(\frac{x}{y^{(p_\varepsilon)}}\Big)\,dx \tag{6.95}$$

and since we only need the value at $\varepsilon = 0$, it simplifies nicely. First, we note that the first term depends only on ε through $y^{(p_\varepsilon)}$, whose dependence on ε was noted above, so we get:

$$\left[\frac{d}{d\varepsilon}\int_I p(x) x \log\Big(\frac{x}{y^{(p_\varepsilon)}}\Big)\,dx\right]_{\varepsilon=0} = -y^* \tag{6.96}$$

Next, we note by the derivative product rule, the derivative of the second part consists of $\int_I z(x) x \log\left(\frac{x}{y^{(p_\varepsilon)}}\right) dx$ plus a term multiplied by ε, which disappears at $\varepsilon = 0$ (and $y^{(p_\varepsilon)}$ reduces just to $y^{(p)}$). Thus,

$$\Big[\frac{d}{d\varepsilon} D^{(p_\varepsilon)}\Big]_{\varepsilon=0} = \int_I z(x) x \log\Big(\frac{x}{y^{(p)}}\Big)\,dx - y^* \tag{6.97}$$

But (since $z(x)$ is a probability density over I),

$$\int_I z(x) x \log\Big(\frac{x}{y^{(p)}}\Big)\,dx = \mathbb{E}_{X\sim z}\Big[X \log\Big(\frac{X}{y^{(p)}}\Big)\Big] \tag{6.98}$$

i.e. the expectation over a convex function. Thus, we may apply Jensen's inequality to it, which reduces nicely to give

$$\mathbb{E}_{X\sim z}\Big[X\log\Big(\frac{X}{y^{(p)}}\Big)\Big] \geq \mathbb{E}_{X\sim z}[X]\log\Big(\frac{\mathbb{E}_{X\sim z}[X]}{y^{(p)}}\Big)\Big] \geq y^* \tag{6.99}$$

and hence $\left[\frac{d}{d\varepsilon}D^{(p_\varepsilon)}\right]_{\varepsilon=0} \geq y^* - y^* = 0$. □

The following lemma is especially useful because we can bound a density function p_1 by a larger density function p_2 and have a distortion bound over n intervals (rather than just treating their domain as one interval).

Lemma 18. *If $p_1 \leq p_2 \in \mathcal{F}_{[0,\infty)}$, then $D(p_1, n) \leq D(p_2, n)$.*

Proof. We note that $p_1, p_2 \in P_I$ mean that their restrictions to any $I_j \subseteq I$ are in P_{I_j}, and that $p_1 \geq p_2$ on I_j as well. Thus, for any $\{I_1, \ldots, I_n\}$ partitioning I, we have

$$\sum_{j=1}^{n} D^{(p_1,I_j)} \leq \sum_{j=1}^{n} D^{(p_2,I_j)} \tag{6.100}$$

But we note that this means we can take the infimum of both sides and get

$$D(p_1, n) = \inf_{\{I_1,\ldots,I_n\}} \sum_{j=1}^{n} D^{(p_1,I_j)} \tag{6.101}$$

$$\leq \inf_{\{I_1,\ldots,I_n\}} \sum_{j=1}^{n} D^{(p_2,I_j)} = D(p_2, n) \tag{6.102}$$

and we are done. □

6.5 Upper Bounds

In this section we derive upper bounds for (6.5) when p_X is a) uniform and b) exponential, and use them to show upper bounds for when p_X is Gamma. Let the support of X be $[\ell, L]$ (typically $[0, 1]$ or $[0, \infty)$); we denote the intervals as $I_j = [a_{j-1}, a_j]$ where

$$\ell = a_0 \leq a_1 \leq \cdots \leq a_{n-1} \leq a_n = L\,. \tag{6.103}$$

Let $y_j \triangleq y^{(p_X,I_j)} = \mathbb{E}[X|X \in I_j]$, and let $r_j \triangleq a_j - a_{j-1}$ (width of interval I_j).

6.5.1 Upper Bound for Uniform X

Since X is uniform, the computation of the centers is greatly simplified: $y_j = \frac{a_j+a_{j-1}}{2} = a_j - \frac{1}{2}r_j$.

Lemma 19. *The distortion on interval I_j is at most $\frac{1}{12}\frac{r_j^3}{y_j}$.*

Proof. Since $p_X(x) = 1$ in I_j, by the fact that $x > 0$ and log is concave, the distortion is

$$\int_{y_j-\frac{r_j}{2}}^{y_j+\frac{r_j}{2}} x\log\Big(\frac{x}{y_j}\Big)dx = \int_{-\frac{r_j}{2}}^{\frac{r_j}{2}} (x+y_j)\log\Big(1+\frac{x}{y_j}\Big)dx \tag{6.104}$$

$$\leq \int_{-\frac{r_j}{2}}^{\frac{r_j}{2}} (x+y_j)\frac{x}{y_j}dx = \frac{1}{12}\frac{r_j^3}{y_j}\,. \tag{6.105}$$

□

Theorem 15. *For $X \sim \mathrm{Unif}_{[0,1]}$,*

$$D(p_X, n) \leq \frac{9}{32}\frac{1}{n^2}\,. \tag{6.106}$$

Proof. Using the Interval Method, we set the interval boundaries as $a_j = j^{3/2}/n^{3/2}$. Then the width of each interval is

$$r_j = \frac{j^{3/2} - (j-1)^{3/2}}{n^{3/2}} \leq \frac{3}{2}\frac{(j-\frac{1}{2})^{1/2}}{n^{3/2}}, \tag{6.107}$$

and the midpoint is

$$y_j = \frac{j^{3/2} + (j-1)^{3/2}}{2n^{3/2}} \geq \frac{(j-\frac{1}{2})^{3/2}}{n^{3/2}} \tag{6.108}$$

The second follows from Jensen's inequality because $j^{3/2}$ is convex, while the first follows because if $f'(t)$ is concave (for simplicity we'll use Newton notation here), then

$$\begin{aligned} f(t) - f(t-1) &= \int_{t-1}^{t} f'(x)\,dx && (6.109)\\ &\leq \int_{-1/2}^{1/2} \big(f'(t-1/2) + f''(t-1/2)x\big)\,dx && (6.110)\\ &= f'(t-1/2) && (6.111) \end{aligned}$$

(we upper bound the integral over $f'(x)$ with the integral of the tangent line at $x = t - 1/2$, which holds because f' is concave, and the linear term in the second integral cancels itself out since no matter what f'' is, it evaluates equally at $x = 1/2$ and $x = -1/2$). Therefore, letting $f(t) = t^{3/2}$, we can conclude that

$$t^{3/2} - (t-1)^{3/2} \leq \frac{3}{2}(t-1/2)^{1/2} \tag{6.112}$$

Therefore, using Lemma 19 we can upper bound the total distortion on I_j by

$$\frac{1}{12}\frac{r_j^3}{y_j} \leq \frac{1}{12}\frac{(3/2)^3(j-\frac{1}{2})^{3/2}}{(j-\frac{1}{2})^{3/2}}\frac{1/n^{9/2}}{1/n^{3/2}} = \frac{9}{32}\frac{1}{n^3}\,. \tag{6.113}$$

Since there are n intervals, the total distortion is bounded above by $(9/32)/n^2$, and we are done. □

An interesting thing to note is that an upper bound of $O(1/n^2)$ is achievable by intervals with boundaries defined by $a_{j+1} = (j+1)^b/n^b$ for any $b > 1$. (When $b = 1$ the intervals are equally spaced, and decay rate becomes $O(\log(n)/n^2)$.) Here is an intuitive argument. Roughly speaking,

$$r_j \approx \frac{d}{dj}\frac{j^b}{n^b} = b\frac{j^{b-1}}{n^b} \tag{6.114}$$

and $y_j \approx j^b/n^b$, giving a bound of

$$D(p_X, n) \leq \frac{1}{12}\frac{r_j^3}{y_j} \approx \frac{1}{12}b^3 j^{2b-3} n^{-2b} \tag{6.115}$$

Then, approximating the sum of j^{2b-3} for $j = 1, \ldots, n$ as

$$\int_0^n t^{2b-3}dt = \frac{1}{2b-2}n^{2b-2}\,, \tag{6.116}$$

we combine for a bound of

$$\sum_{j=1}^{n} \frac{1}{12}\frac{r_j^3}{y_j} \approx \frac{1}{12}\frac{b^3}{2b-2}\frac{1}{n^2}\,. \tag{6.117}$$

Setting $b = 3/2$ not only allows for a very clean proof but also minimizes the constant $b^3/(2b-2)$. This reasoning also makes clear why $b = 1$ introduces the log factor, as the sum of $1/j$ for $j = 1, \ldots, n$ grows like $\log(n)$.

We then use Lemma 16 with Theorem 15 to get Theorem 12. This is shown in Appendix B.2.7.

6.5.2 Upper Bound for Exponential X

Proof of Theorem 13. We split the distribution into the regions $x \le \tau$ and $x > \tau$ for some $\tau \ge 1$ and use the results in Section 6.4 to bound the distortion from each region separately. We will give n_1 intervals to the $[0, \tau]$ region and n_2 to the $[\tau, \infty)$ region; thus, if $n = n_1 + n_2$,

$$D(p_X, n) \le D(p_X|_{[0,\tau]}, n_1) + D(p_X|_{[\tau,\infty)}, n_2). \tag{6.118}$$

We will discuss how to set n_1, n_2 and τ later.

For the $[0, \tau]$ region, we use the upper bound $p_X(x) \le 1$. Therefore, by Lemmas 16 and 18 and Theorem 15,

$$D(p_X|_{[0,\tau]}, n_1) \le \tau^2 \frac{9}{32}\frac{1}{n_1^2}\,. \tag{6.119}$$

For the $[\tau, \infty)$ region, we use $a_j = 3\log\big(n_2/(n_2 - j)\big) + \tau$ as our interval boundaries (noting that $a_{n_2} = \infty$, as it should).

First, we consider the infinite interval $I_{n_2} = [a_{n_2-1}, a_{n_2}) = [3\log(n_2) + \tau, \infty)$.

$$\int_{a_{n_2-1}}^{\infty} e^{-x} x \log\left(\frac{x}{y_{n_2}}\right) dx \tag{6.120}$$

$$= \int_{a_{n_2-1}}^{\infty} e^{-x} x \log x\, dx - \int_{a_{n_2-1}}^{\infty} e^{-x} x \log y_{n_2}\, dx\,. \tag{6.121}$$

For simplicity, denote $a = a_{n_2-1}$. Using integration by parts on the first integral, we get that

$$\int_a^{\infty} e^{-x} x \log x\, dx \tag{6.122}$$

$$= -(\log x)(x+1)e^{-x}\Big|_a^{\infty} + \int_a^{\infty} e^{-x}\frac{x+1}{x} dx \tag{6.123}$$

$$= e^{-a}(a+1)\log a + e^{-a} + \int_a^{\infty} \frac{e^{-x}}{x}\, dx \tag{6.124}$$

$$\le e^{-a}(a+1)\log a + \Big(1 + \frac{1}{a}\Big)e^{-a} \tag{6.125}$$

$$= e^{-a}(a+1)\Big(\log a + \frac{1}{a}\Big) \tag{6.126}$$

For computing the second integral, from memorylessness of the exponential distribution, $y_{n_2} = a + 1$.

$$-\int_a^{\infty} e^{-x} x \log y_{n_2}\, dx = -\int_a^{\infty} e^{-x} x \log(a+1)\, dx \tag{6.127}$$

$$= e^{-a}(a+1)\log\frac{1}{a+1} \tag{6.128}$$

Combining the integrals gives

$$e^{-a}(a+1)\log\frac{1}{a+1}+e^{-a}(a+1)\Big(\log a+\frac{1}{a}\Big) \tag{6.129}$$

$$=e^{-a}(a+1)\Big(\log\frac{a}{a+1}+\frac{1}{a}\Big) \tag{6.130}$$

$$\le e^{-a}(a+1)\Big(-\frac{1}{a+1}+\frac{1}{a}\Big) \tag{6.131}$$

$$=\frac{e^{-a}}{a} \tag{6.132}$$

$$\le \frac{e^{-\tau}}{\tau}\frac{1}{n_2^3} \tag{6.133}$$

The last inequality happens because $a=3\log(n_2)+\tau$ so $a>\tau$ (for the denominator) and $e^{-a}=e^{-\tau}\frac{1}{n_2^3}$. We then take the rather unusual step of using

$$\frac{e^{-\tau}}{\tau}\frac{1}{n_2^3}\le 18\frac{e^{-\tau}}{\tau}\frac{1}{n_2^3} \tag{6.134}$$

This is meant to match it to the bound we get for other intervals $[a_{j-1},a_j]$, so that we can sum them all nicely.

For the other intervals I_j, since $p_X(x)=e^{-x}$ is decreasing,

$$p_X(x)\le e^{-a_{j-1}}=e^{-\tau}\left(\frac{n_2-j+1}{n_2}\right)^3 \tag{6.135}$$

over I_j. Also,

$$r_j=3\log\left(\frac{n_2}{n_2-j}\right)+\tau-3\log\left(\frac{n_2}{n_2-j+1}\right)-\tau \tag{6.136}$$

$$=3\left(\log(n_2-j+1)-\log(n_2-j)\right) \tag{6.137}$$

$$\le 3\frac{1}{n_2-j}\,. \tag{6.138}$$

Therefore, by Lemmas 16, 17 and 19 we get, for all j,

$$D^{(p_X,I_j)}\le D^{(e^{-\tau}(\frac{n_2-j+1}{n_2})^3,I_j)} \tag{6.139}$$

$$\le \frac{1}{12}e^{-\tau}\left(\frac{n_2-j+1}{n_2}\right)^3\frac{r_j^3}{y_j} \tag{6.140}$$

$$\le \frac{1}{12}e^{-\tau}\left(\frac{n_2-j+1}{n_2}\right)^3\frac{27}{(n_2-j)^3y_j} \tag{6.141}$$

$$\le 18\frac{e^{-\tau}}{\tau}\frac{1}{n_2^3} \tag{6.142}$$

since $y_j\ge\tau$ and $\left(\frac{n_2-j+1}{n_2-j}\right)^3\le 8$ (a very loose bound for most j). Summing gives $D(p_X|_{[\tau,\infty)},n_2)\le 18\frac{e^{-\tau}}{\tau}\frac{1}{n_2^2}$.

Therefore, we have our bound over $[0,\infty)$ for $n=n_1+n_2$:

$$D(p_X,n)\le\tau^2\frac{9}{32}\frac{1}{n_1^2}+18\frac{e^{-\tau}}{\tau}\frac{1}{n_2^2}. \tag{6.143}$$

Let $n_1=cn$ and $n_2=(1-c)n$; we know $c=j/n$ for some $j\in[n-1]$, but for now let's require only $c\in(0,1)$. Thus,

$$D(p_X,n)\le\Big(\tau^2\frac{9}{32}\frac{1}{c^2}+18\frac{e^{-\tau}}{\tau}\frac{1}{(1-c)^2}\Big)\frac{1}{n^2}. \tag{6.144}$$

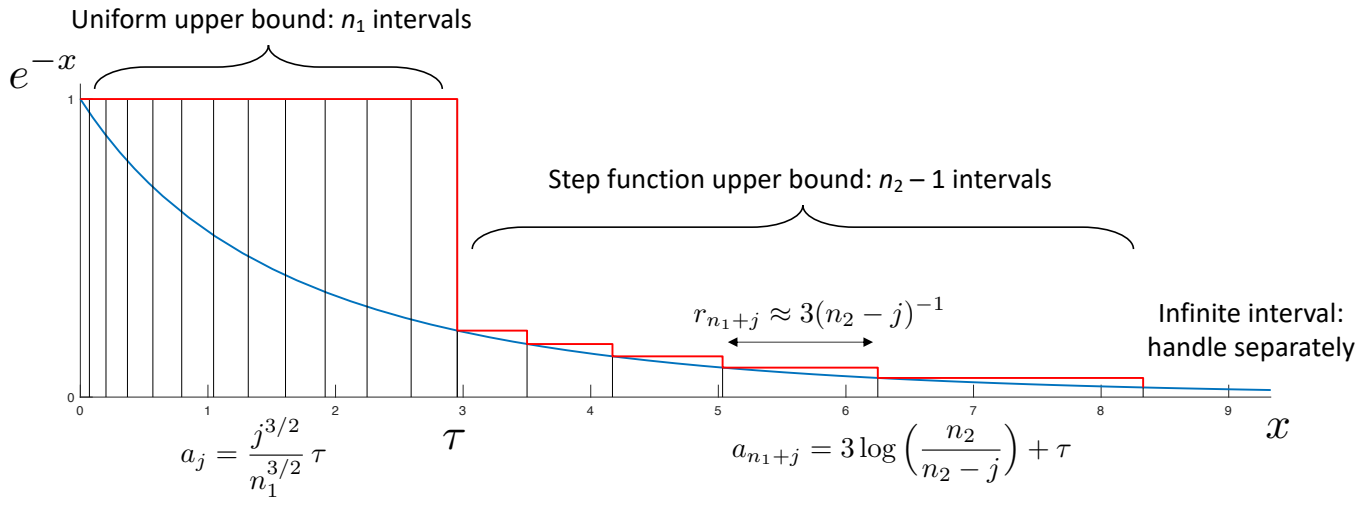


Figure 6.1: Diagram of interval sizes and function $p'(x)$ used for computing the upper bound for the exponential function. (NOTE: Exponential function not drawn to scale for illustrative purposes.)

Numerical optimization gives a minimum at $\tau \approx 2.954$ and $c \approx 0.664$, giving a constant of 8.4.

Now we bring back the $c = j/n$ condition. We note that any $c \in [0.614, 0.714]$ (and $\tau = 2.954$) gives $D(p_X, n) \leq 9/n^2$; so this holds for all $n \geq 10$ since $c = j/n \in [0.614, 0.714]$ always exists for $n \geq 10$. For $n = 5, 6, 7, 8, 9$ we can set $c = 3/5, 4/6, 5/7, 5/8, 6/9$ respectively (with the same τ), which gives a constant of ≤ 9 as well; and for $n \leq 4$ direct numerical optimization of the intervals gives the desired bound. Therefore, $D(p_X, n) \leq 9/n^2$ as we wanted.

□

Some of the bounds we used were quite loose, so more detailed analysis (omitted) can improve the constant, though not the decay rate.

Corollary 5. *For $X \sim$ Exponential(λ), $D(p_X, n) \leq \frac{9}{\lambda n^2}$*

Proof. Let $p_E(x) = e^{-x}$. The random variable X has pdf $p_X(x) = \lambda e^{-\lambda x}$. Using i) and ii) of Lemma 16,

$$D(p_X, n) = D(\lambda p_E(\lambda x), n) \tag{6.145}$$

$$= \frac{1}{\lambda} D(p_E, n)\,. \tag{6.146}$$

□

The corresponding rate distortion function for $X \sim$ Exponential(λ) has an additional $(1/2)\log(1/\lambda)$ term.

6.5.3 Upper Bound for Gamma X

We will combine the bounds we found for the uniform and exponential case to get bounds for Gamma distributions. A random variable $X \sim$ Gamma(α, β) where $\alpha, \beta > 0$ has the probability density

$$P_X(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{6.147}$$

where α is the shape parameter and β is the rate parameter.

When $\alpha < 1$, "head" part (portion near zero) has probability density which goes to infinity as $x \to 0$. We need the following lemma to find intervals for this "head" part before deriving a bound for all Gamma distributions.

Lemma 20. *For any $-1 < s < 0$, if $p_X(x) = (1+s)x^s$ on $[0,1]$ (and 0 elsewhere),*

$$D(p_X, n) \leq \frac{19/3}{n^2}\,. \tag{6.148}$$

While give the formal proof in Appendix B.2.8, the following illustrates the main ideas of the proof. Suppose we set the interval boundaries at $a_j = j^\gamma/n^\gamma$ (where γ is to be determined later). Ignoring the first interval, we have interval lengths

$$r_j = a_j - a_{j-1} = \frac{j^\gamma - (j-1)^\gamma}{n^\gamma} \approx \frac{\gamma(j-1/2)^{\gamma-1}}{n^\gamma} \tag{6.149}$$

(The above is a less-than if x^s is convex.) We also approximate $y_j \approx \frac{(j-1/2)^\gamma}{n^\gamma}$ and the density over I_j as $(1+s)\big(\frac{(j-1/2)^\gamma}{n^\gamma}\big)^s$. Therefore, we can get the distortion over I_j to be approximately

$$D(p_X, I_j) \leq \frac{1}{12}(1+s)\Big(\frac{(j-1/2)^\gamma}{n^\gamma}\Big)^s \frac{r_j^3}{y_j} \tag{6.150}$$

$$= \frac{1}{12}\gamma^3(1+s)\frac{(j-1/2)^{(2+s)\gamma-3}}{n^{(2+s)\gamma}} \tag{6.151}$$

The overall distortion is a sum over all this, i.e.

$$D(p_X, n) \leq \frac{1}{12}\gamma^3(1+s)n^{-(2+s)\gamma}\sum_{j=1}^{n}(j-1/2)^{(2+s)\gamma-3} \tag{6.152}$$

$$\approx \frac{1}{12}\gamma^3(1+s)n^{-(2+s)\gamma}\int_0^n t^{(2+s)\gamma-3}dt \tag{6.153}$$

$$= \frac{1}{12}\gamma^3(1+s)n^{-(2+s)\gamma+1}\int_0^1 (nu)^{(2+s)\gamma-3}du \tag{6.154}$$

$$= \frac{1}{12}\gamma^3(1+s)n^{-2}\int_0^1 u^{(2+s)\gamma-3}du \tag{6.155}$$

$$= \frac{1}{12}\gamma^3(1+s)n^{-2}\frac{1}{(2+s)\gamma-2} \tag{6.156}$$

(using change of variables $u = t/n$). This holds for all $\gamma > \frac{2}{2+s}$. Now we ask, what γ minimizes this? All the terms except $\frac{\gamma^3}{(2+s)\gamma-2}$ are irrelevant to that question. Taking the derivative, we get

$$\frac{d}{d\gamma}\frac{\gamma^3}{(2+s)\gamma-2} = \frac{3\gamma^2((2+s)\gamma-2)-(2+s)\gamma^3}{((2+s)\gamma-2)^2} \tag{6.157}$$

$$= \frac{2(2+s)\gamma^3 - 6\gamma^2}{((2+s)\gamma-2)^2} \tag{6.158}$$

Setting the above to 0 (when $\gamma > \frac{2}{2+s}$ yields $\gamma = \frac{3}{2+s}$, and plugging back into the bound we get approximately gives

$$D(p_X, n) \lesssim \frac{9}{4}\frac{1+s}{(2+s)^3}\frac{1}{n^2} \tag{6.159}$$

We can ask: what s maximizes this (clearly $s \to -1$ and $s \to \infty$ both push the constant down towards 0)? It turns out that $s = -1/2$ is the worst constant, giving

$$D(p_X, n) \lesssim \frac{1}{3}\frac{1}{n^2} \tag{6.160}$$

and this corresponds to the Jeffreys prior (Dirichlet with $\alpha = 1/2$).

These 'head' (as opposed to 'tail') bounds can be combined with Lemma 17 to obtain bounds for other common ways that densities can increase to ∞ as $x \to 0$:

Proposition 20. *For $X \sim \text{Gamma}(\alpha, \beta)$,*

$$D(p_X, n) = \frac{C_G(\alpha)}{\beta n^2}\,. \tag{6.161}$$

where

$$C_G(\alpha) = \begin{cases} \frac{144}{\Gamma(\alpha)} + \frac{(\alpha-1)^4}{\sqrt{2\pi(\alpha-1)}}\frac{81}{8}, & \alpha > 1 \\ 9, & \alpha = 1 \\ \left(36 + \frac{4}{\alpha}\right)\frac{1}{\Gamma(\alpha)}, & \alpha < 1 \end{cases} \tag{6.162}$$

Proof. The key to this proof is that, using the Interval Method, we will bound the tail portion of the Gamma distribution by upper bounding with the exponential function. For the front or head portion, depending on the parameter α of the Gamma distribution (6.147), we will either bound with uniform or the function from Lemma 20.

Let p_X be the pdf of $X \sim \text{Gamma}(\alpha, 1)$. (Since the rate parameter β just scales the function, we will account for general values of β at the end after solving for the case where $\beta = 1$.) We will use $I_L = [0, b)$ and $I_R = [b, \infty)$ for some b. The intended purpose is to indicate that the values in I_R will be those upper-bounded by the exponential, and I_L to be the values bounded by either the uniform or the function in Lemma 20. The value of b will need to specified. We will use $p_{X,R}$ to be the function which is equal to p_X on I_R and zero otherwise. Similarly, $p_{X,L}$ is the function equal to p_X on I_L and is zero otherwise. Let $h_E(x) = e^{-x}$.

We will start with the case where $\alpha > 1$. First we find the point b where we switch from using the uniform upper bound to the exponential upper bound.

Let b be such that

$$\frac{\log x}{x} \leq \frac{1/2}{\alpha - 1} \tag{6.163}$$

for all $x \geq b$. This must exist since $\frac{\log x}{x} \to 0$ as $x \to \infty$. For a crude value of b, we can choose $b = 3(\alpha-1)^2$.

We chose this particular condition on b because

$$\frac{\log x}{x} \leq \frac{1/2}{\alpha - 1} \tag{6.164}$$

$$(\alpha - 1)\log x \leq \frac{x}{2} \tag{6.165}$$

$$x^{\alpha-1} \leq e^{x/2} \tag{6.166}$$

$$\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x} \leq \frac{1}{\Gamma(\alpha)}e^{-x/2}\,. \tag{6.167}$$

Define $h_R(x) = \frac{1}{\Gamma(\alpha)}e^{-x/2}$ on I_R and zero elsewhere. Using parts i) and ii) of Lemma 16 with the optimal intervals $I_1, ..., I_n$ for the exponential distribution,

$$D(p_{X,R}, n) \leq D(h_R, n) \tag{6.168}$$

$$\leq D\left(\frac{1}{\Gamma(\alpha)}h_E(x/2), n\right) \tag{6.169}$$

$$\leq \sum_{j=1}^{n} \frac{2^2}{\Gamma(\alpha)} D^{(h_E, I_j)} \tag{6.170}$$

$$= \frac{36}{\Gamma(\alpha)n^2} \tag{6.171}$$

This upper-bounds the tail portion when $\alpha > 1$. Now for the head portion. Let $m = \max_{x \in I_L} p_X(x)$ and define $h_L(x) = m$ on I_L and zero otherwise. The mode of the distribution Gamma$(\alpha, 1)$ occurs at $\alpha - 1$

when $\alpha > 1$. This means that

$$m = \frac{1}{\Gamma(\alpha)}(\alpha-1)^{\alpha-1}e^{-(\alpha-1)} \tag{6.172}$$

$$= \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha-1}{e}\right)^{\alpha-1} \leq \frac{1}{\sqrt{2\pi(\alpha-1)}}\,. \tag{6.173}$$

where to simplify, we used the Stirling approximation lower bound for when $\alpha > 1$.

Let $u(x) = 1$ on the interval $[0,1]$ and zero elsewhere. Using parts i) and ii) of Lemma 16 with the optimal intervals $I_1, ..., I_n$ for the uniform distribution on $[0,1]$

$$D(p_{X,L}, n) \leq D(h_L, n) \tag{6.174}$$

$$\leq D(m \cdot u(x/b), n) \tag{6.175}$$

$$\leq \sum_{j=1}^{n} mb^2 D^{(u,I_j)} \tag{6.176}$$

$$= \frac{9mb^2}{32n^2} \tag{6.177}$$

$$\leq \frac{9(\alpha-1)^4}{\sqrt{2\pi(\alpha-1)}}\frac{9}{32}\frac{1}{n^2}\,. \tag{6.178}$$

Then

$$D(p_X, n) \leq D(p_{X,L}, n/2) + D(p_{X,R}, n/2) \tag{6.179}$$

$$\leq \frac{36}{\Gamma(\alpha)(n/2)^2} + \frac{(\alpha-1)^4}{\sqrt{2\pi(\alpha-1)}}\frac{81}{32}\frac{1}{(n/2)^2} \tag{6.180}$$

$$\leq \left(\frac{144}{\Gamma(\alpha)} + \frac{(\alpha-1)^4}{\sqrt{2\pi(\alpha-1)}}\frac{81}{8}\right)\frac{1}{n^2}\,. \tag{6.181}$$

This completes the proof for when $\alpha > 1$. We do not need to consider when $\alpha = 1$ because that is the exponential distribution showed in Theorem 13. We now show the result for $\alpha < 1$. The process is very similar to the case of $\alpha > 1$. We will choose $b = 1$.

Let $h_R(x) = \frac{1}{\Gamma(\alpha)}e^{-x}$ on I_R and zero elsewhere. Because we are restricting to $x \geq 1$, $p_{X,R}(x) \leq h_R(x)$. Like above,

$$D(p_{X,R}, n) \leq D(h_R, n) \tag{6.182}$$

$$\leq D\left(\frac{1}{\Gamma(\alpha)}h_E(x), n\right) \tag{6.183}$$

$$\leq \sum_{j=1}^{n}\frac{1}{\Gamma(\alpha)}D^{(h_E,I_j)} \tag{6.184}$$

$$= \frac{9}{\Gamma(\alpha)n^2}\,. \tag{6.185}$$

Let $h_L(x) = \frac{1}{\Gamma(\alpha)}x^{\alpha-1}$ on I_L and zero elsewhere. Let $h_S(x) = \alpha x^{\alpha-1}$. We have shown in Lemma 20 that $D(h_S, n)$ is bounded above by $19/(3n^2)$). Using parts i) and ii) of Lemma 16 with the optimal intervals

$I_1, ..., I_n$ for h_S on $[0,1]$,

$$D(p_{X,L}, n) \leq D(h_L, n) \tag{6.186}$$
$$\leq D\left(\frac{1}{\alpha\Gamma(\alpha)} h_S(x), n\right) \tag{6.187}$$
$$\leq \sum_{j=1}^{n} \frac{1}{\alpha\Gamma(\alpha)} D^{(h_S, I_j)} \tag{6.188}$$
$$= \frac{1}{\alpha\Gamma(\alpha) n^2}\,. \tag{6.189}$$

Like above,

$$D(p_X, n) \leq D(p_{X,L}, n/2) + D(p_{X,R}, n/2) \tag{6.190}$$
$$\leq \frac{9}{\Gamma(\alpha)(n/2)^2} + \frac{1}{\alpha\Gamma(\alpha)(n/2)^2} \tag{6.191}$$
$$\leq \left(36 + \frac{4}{\alpha}\right) \frac{1}{\Gamma(\alpha)} \frac{1}{n^2} \tag{6.192}$$

This completes the proof for when $\alpha < 1$.

If we have the distribution Gamma(α, β), we can use parts i) and ii) of Lemma 16 to scale the distortion. We need a scaling that transforms

$$\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \to \frac{\beta}{\Gamma(\alpha)} (\beta x)^{\alpha-1} e^{-\beta x} \tag{6.193}$$
$$p_X(x) \to \beta p_X(\beta x)\,. \tag{6.194}$$

The overall distortion is scaled by $1/\beta$. □

Proof of Theorem 14. From Proposition 20 and Proposition 19, we then get the rate distortion result for Gamma distributions. □

6.6 Expected Divergence Results

Finally, we connect divergence rate distortion on Gamma sources to quantizing $\mathcal{P}([K])$ with symmetric Dirichlet prior.

Fact 10. *Let $X_i \overset{iid}{\sim}$ Gamma(α, β) for $i \in [K]$ (giving random vector X^K) and let $S \triangleq \sum_{i=1}^{K} X_i$. Then*

$$P \triangleq X^K / S \sim Dir_K(\alpha) \tag{6.195}$$

and S and P are independent. (The division above treats P and X^K as K-length vectors.)

Lemma 21. *For $S \sim$ Gamma(α, α), $\mathbb{E}[S \log S] \leq 1/(2\alpha)$.*

This can be proved using properties of the Gamma function. We show this in Appendix B.2.9.

Proposition 21. *Let p_Z be the probability density of $Z \sim$ Gamma(α, α). Then, for $K > 0$,*

$$K \cdot R\left(p_Z, D + 1/(2\alpha K)\right) \leq \log M^*(Dir_K(\alpha), D). \tag{6.196}$$

Proof. Let random vector $X^K = (X_1, ..., X_K)$ be such that each $X_i \sim \text{Gamma}(\alpha, \alpha K)$. Let

$$S = \sum_{i=1}^{K} X_i \,. \tag{6.197}$$

Then, $S \sim \text{Gamma}(\alpha K, \alpha K)$ and $\mathbb{E}[S] = 1$.

From Fact 10, we know that $X^K/S \sim \text{Dir}_K(\alpha)$. Also S is independent of X^K/S and therefore it is also independent of $\min_j D_{\text{KL}}((X^K/S)||Q(j))$. Fixing J as the argmin, we denote the components of $Q(J)$ as $(Y_1, \ldots, Y_K)$. Using Lemma 21,

$$\sum_{i=1}^{K} \mathbb{E}\left[X_i \log \frac{X_i}{Y_i}\right] \tag{6.198}$$

$$= \mathbb{E}\left[\sum_{i=1}^{K} S\frac{X_i}{S} \log \frac{X_i}{SY_i} + X_i \log S\right] \tag{6.199}$$

$$= \mathbb{E}[S]\,\mathbb{E}[D_{\text{KL}}((X^K/S)||Q(J))] + \mathbb{E}[S \log S] \tag{6.200}$$

$$\leq D + \frac{1}{2\alpha K} \tag{6.201}$$

Since we showed that a K-ary quantizer with the given average distortion exists, the standard single-letter lower bound from rate distortion[43, Ch 25] implies

$$\log M^*(\text{Dir}_K(\alpha), D) \geq KR\left(p_X, \frac{D + 1/(2\alpha K)}{K}\right) . \tag{6.202}$$

If p_X is the density of each X_i, then by scaling (Proposition 15)

$$R\left(p_X, \frac{D + 1/(2\alpha K)}{K}\right) = R(p_Z, D + 1/(2\alpha K)) \tag{6.203}$$

where $Z \sim \text{Gamma}(\alpha, \alpha)$. □

Proposition 22. *For any α, K such that $\alpha K > 1$, there exist*

$$m = \left(\frac{C_G(\alpha)/\alpha}{D(1 - (\alpha K)^{-1/3})^2}\right)^{K/2} \tag{6.204}$$

centers $Q(1), \ldots, Q(m)$ such that

$$\mathbb{E}_{P \sim Dir_K(\alpha)} \min_j D_{\text{KL}}(P||Q(j)) \leq D \,. \tag{6.205}$$

The values for $C_G(\alpha)$ are given in Proposition 20.

Proof. Let $S \sim \text{Gamma}(\alpha K, \alpha K)$ and $P = (p_1, ..., p_K) \sim \text{Dir}_K(\alpha)$. Let $X_i = Sp_i$ so $X_i \sim \text{Gamma}(\alpha, \alpha K)$.

From Proposition 20, we know there exists a scheme using n quantization points for X_i, so that

$$\mathbb{E}[d(X_i, Y_i)] \leq \frac{C_G(\alpha)}{\alpha K n^2} \text{ and } \mathbb{E}[X_i] = \mathbb{E}[Y_i] \,. \tag{6.206}$$

Set $D' = C_G(\alpha)/(\alpha n^2)$. Then for

$$m = \left(\frac{C_G(\alpha)}{\alpha D'}\right)^{K/2} = \left(\frac{C_G(\alpha)}{\alpha K(D'/K)}\right)^{K/2} \tag{6.207}$$

centers, we have that

$$\sum_{i=1}^{K} \mathbb{E}[d(X_i, Y_i)] \leq \sum_{i=1}^{K} \frac{D'}{i} = D' \,. \tag{6.208}$$

Next we will relate this to $D_{\mathrm{KL}}(P\|Q)$.

$$D' \geq \sum_{i=1}^{K} \mathbb{E}[d(X_i, Y_i)] = \mathbb{E}\left[\sum_{i=1}^{K} X_i \log \frac{X_i}{Y_i}\right] \tag{6.209}$$

$$= \mathbb{E}\left[S \sum_{i=1}^{K} \frac{X_i}{S} \log \frac{X_i/S}{Y_i/\sum_{i=1}^{K} Y_i} + S \log \frac{S}{\sum_{i=1}^{K} Y_i}\right] \tag{6.210}$$

$$= \mathbb{E}\left[S D_{\mathrm{KL}}(P||Q) + S \log \frac{S}{\sum_{i=1}^{K} Y_i}\right] \tag{6.211}$$

$$\geq \mathbb{E}\left[S D_{\mathrm{KL}}(P||Q)\right] \,. \tag{6.212}$$

To get the last inequality, we can apply the log sum inequality [1, Ch. 2]. Let $T = \sum_{i=1}^{K} Y_i$ and $p(s,t) = \mathbb{P}\left[S = s, T = t\right]$. to show that

$$\mathbb{E}\left[S \log \frac{S}{T}\right] = \int p(s,t) s \log \frac{p(s,t)s}{p(s,t)t} \, ds \, dt \tag{6.213}$$

$$\geq \int p(s,t) s \, ds \log \frac{\int p(s,t)s \, ds}{\int p(s,t)t \, dt} \tag{6.214}$$

$$= \mathbb{E}[S] \log \frac{\mathbb{E}[S]}{\mathbb{E}\left[T\right]} \tag{6.215}$$

$$= \mathbb{E}[S] \log \frac{\mathbb{E}[S]}{\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]} \tag{6.216}$$

$$= 0 \,. \tag{6.217}$$

Note that Q depends on S and P. We use the following fact: if $Z_1, Z_2 \geq 0$ are random variables then for any $\sigma > 0$,

$$\mathbb{E}[Z_2 \,|\, Z_1 \geq \sigma] \leq \frac{\mathbb{E}[Z_1 Z_2]}{\sigma \mathbb{P}[Z_1 \geq \sigma]} \,. \tag{6.218}$$

Therefore, we have for any $\sigma > 0$,

$$\mathbb{E}[D_{\mathrm{KL}}(P||Q) \,|\, S \geq \sigma] \leq \frac{\mathbb{E}[S D_{\mathrm{KL}}(P||Q)]}{\sigma \mathbb{P}[S \geq \sigma]} \leq \frac{D'}{\sigma \mathbb{P}[S \geq \sigma]} \,. \tag{6.219}$$

We then fix $s \geq \sigma$ to minimize $\mathbb{E}[D_{\mathrm{KL}}(P||Q) \,|\, S = s]$ to get

$$\min_{s \geq \sigma} \mathbb{E}[D_{\mathrm{KL}}(P||Q) \,|\, S = s] \leq \frac{D'}{\sigma \mathbb{P}[S \geq \sigma]} \,. \tag{6.220}$$

We then want to set σ to maximize $\sigma \mathbb{P}[S \geq \sigma]$. We have $\mathbb{E}[S] = \alpha K/(\alpha K) = 1$ and $\mathrm{Var}[S] = \alpha K/(\alpha K)^2 = 1/(\alpha K)$. Thus, defining $t = 1 - \sigma$ (assuming $\sigma \leq 1$), applying Chebyshev's inequality, and setting $t = (\alpha K)^{-1/3}$ gives

$$\mathbb{P}[S \geq 1 - t] \geq 1 - 1/(\alpha K t^2) \tag{6.221}$$

which implies (for $t = (\alpha K)^{-1/3}$),

$$\mathbb{E} D_{\mathrm{KL}}(P||Q) \leq \frac{D'}{(1-t)\left(1 - \frac{1}{(\alpha K t^2)}\right)} = \frac{D'}{(1 - (\alpha K)^{-1/3})^2} \tag{6.222}$$

(here the expectation is only over P, and use fixed $S = s$).

We can then define our coding of $P \in \mathcal{P}([K])$ to a center $Q \in \{Q(j)\}_{j=1}^m$ to be the result of the following procedure:

$$P \to X^K \to Y^K \to Q \tag{6.223}$$

where $X^K = sP$, Y^K is the encoding of X^K using the Interval Method and $Q = Y^K / \sum_{i=1}^K Y_i$.

Now that we have a bound on the distortion produced by this mapping of each P to a center Q, we need to consider how many centers it maps to. Recall that each X^K maps to one of $m = (C_G(\alpha)/(\alpha D'))^{K/2}$ centers. Thus, letting

$$D = \frac{D'}{(1 - (\alpha K)^{-1/3})^2}, \tag{6.224}$$

we get a coding scheme on m centers where $\mathbb{E}[D_{\mathrm{KL}}(P||Q)] \leq D$, so long as

$$m = \Big(\frac{C_G(\alpha)}{\alpha D'}\Big)^{K/2} = \Big(\frac{C_G(\alpha)/\alpha}{D(1 - (\alpha K)^{-1/3})^2}\Big)^{K/2}. \tag{6.225}$$

□

While we proved the existence of such a set of centers, note that we do not explicitly compute s and so finding an efficient covering remains open. We suspect that $s = 1$ will work in practice.

Finally, we can put these together to prove Theorem 10:

Proof of Theorem 10. The first part follows from Proposition 21 and Theorem 11, and the second from Proposition 22. □

6.7 Connection to Universal Compression

Since we are working with KL divergence, our work ties in directly to universal compression. This connection to universal compression also connects this work to Minimum Description Length (MDL) [73, 60, 13, 74] and especially to two-step compression algorithms.

Our method suggests a two-step compression algorithm to losslessly compress X^n by (i) selecting a distribution Q_i (over the symbol set) from a finite set $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$, and (ii) compressing X^n according to Q_i (using for example Huffman coding or arithmetic coding). This incurs additional loss (above the loss needed to compress X^n given the true distribution) from two sources: (a) not having exactly the right distribution, and (b) from the need to specify i from $[m]$. The term at (a) causes at most $\mathbb{E}_{\pi \sim W} \min_i D_{\mathrm{KL}}(\pi||Q_i)$ extra loss per symbol; (b) causes $\log(m)$ loss.

Proof of Proposition 14. Fix W, and for simplicity let $m := M^*(W, \varepsilon)$. Let $P_{X^n|\pi}(X^n|\pi) = \prod_{t=1}^n \pi(X_t)$.

For ε, let $\mathcal{Q}$ be the set $\{Q_1, ..., Q_m\}$ so that $\mathbb{E}_{\pi \sim W} \min_i D_{\mathrm{KL}}(\pi \| Q_i) \leq \varepsilon$. We will use the notation $Q_i^{\otimes n}$ to be a product distribution on X^n (i.e. $Q_i^{\otimes n}(X^n) = \prod_{t=1}^n Q_i(X_t)$).

Let

$$\bar{Q}(X^n) = \frac{1}{m} \sum_{i=1}^m Q_i^{\otimes n}(X^n). \tag{6.226}$$

$$
\begin{aligned}
&I(\pi; X^n) \\
&\quad = \min_Q D(P_{X^n|\pi} \| Q | P_\pi) && (6.227)\\
&\quad \le D(P_{X^n|\pi} \| \bar{Q} | P_\pi) \\
&\quad = \mathbb{E}_{\pi\sim W} \sum_{x^n} P_{X^n|\pi}(x^n|\pi) \log \frac{P_{X^n|\pi}(x^n|\pi)}{\frac{1}{m}\sum_{i=1}^m Q_i^{\otimes n}(x^n)} \\
&\quad \le \mathbb{E}_{\pi\sim W} \min_i \sum_{x^n} P_{X^n|\pi}(x^n|\pi) \log \frac{P_{X^n|\pi}(x^n|\pi)}{\frac{1}{m} Q_i^{\otimes n}(x^n)} \\
&\quad = \mathbb{E}_{\pi\sim W} \log m + \min_i \sum_{x^n} \prod_{t=1}^n \pi(x_t) \log \frac{\prod_{t=1}^n \pi(x_t)}{\prod_{t=1}^n Q_i(x_t)} \\
&\quad = \log m + \mathbb{E}_{\pi\sim W} \min_i n D_{\mathrm{KL}}(\pi \| Q_i) \\
&\quad \le \log M^*(W, \varepsilon) + n\varepsilon
\end{aligned}
$$

In (6.227) we used [43, Corollary 3.1] where Q is any distribution on X^n. Since the equation holds for all $\varepsilon \ge 0$, we can use the ε which gives the infimum.

For the case when $n = 1$, suppose we set $\varepsilon = I(\pi; X)$. Then for $Q_1(x) = \int \pi(x) W(d\pi)$,

$$
\begin{aligned}
\varepsilon &= I(\pi; X) && (6.228)\\
&= \int W(d\pi) \sum_x \pi(x) \log \frac{\pi(x)}{Q_1(x)} && (6.229)\\
&= \mathbb{E}_{\pi\sim W} D_{\mathrm{KL}}(\pi || Q_1) && (6.230)
\end{aligned}
$$

Only one center, Q_1, is needed to achieve $\mathbb{E}_{\pi\sim W} D_{\mathrm{KL}}(\pi||Q_1) \le \varepsilon$. So $\log M^*(W, \varepsilon) = 0$, and thus there exists a ε for which

$$
I(\pi; X) = \log M^*(W, \varepsilon) + \varepsilon \,. \tag{6.231}
$$

□

In the universal compression setting, the quantity $I(\pi; X^n)$ equals the Bayes risk (or average redundancy) $R_n(W)$ with respect to prior W, which is defined as

$$
R_n(W) = \int D(P_{X^n|\pi} \| Q^{(n)}) W(d\pi) \tag{6.232}
$$

where the best strategy $Q^{(n)}$ is to chose the posterior distribution

$$
Q^{(n)}(X^n) = \int P_{X^n|\pi}(X^n|\pi) W(d\pi) \,. \tag{6.233}
$$

A long line of work has been dedicated to computing (6.232), though most focus on when W is the worst case prior or least-favorable prior [19, 12]. (The worst case occurs at $W = \mathrm{Dir}(1/2)$ which also known as the Jeffreys' prior [21].)

In [20], the authors computed that given suitable conditions on W, for any π_0 in the interior of the simplex

$$
D(P_{X^n|\pi_0} \| Q^{(n)}) = \frac{K-1}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log \det I(\pi_0) + \log \frac{1}{W(\pi_0)} + o(1) \,. \tag{6.234}
$$

This bound computes all the constant terms in the asymptotic expression. (Previously, an asymptotic lower bound of $((K-1)/2) \log n$ for almost all θ was shown in [60]. The same term without specifying constants was also shown to be the order of growth for [73].)

It is believed that the the term $((K-1)/2)\log n$ plus some constant is the correct asymptotic limit. (Issues with uniform convergence prevents us from simply taking an integral of (6.234) over the prior W. Unfortunately, the Dirichlet prior can blow up near the boundary of the simplex.)

Using (14) and setting $\varepsilon = 1/n$, we get the (supposed) asymptotic lower bound

$$M^*(W,\varepsilon) \geq \frac{K-1}{2}\log\frac{1/\varepsilon}{2\pi e} + c(W) - 1 + o(1)\,. \tag{6.235}$$

Conditions when the Bayes risk equals the integration of (6.234) are explored in [61].

We can use our upper bounds on $M^*(W,\varepsilon)$ to upper bound the Bayes risk. Setting $\varepsilon = 1/n$, we can get a non-asymptotic bound for any $W = \mathrm{Dir}(\alpha)$ on a alphabet of size K, that

$$R_n(W) \leq \frac{K}{2}\log\frac{nC_G(\alpha)/\alpha}{(1-(\alpha K)^{-1/3})^2}\,. \tag{6.236}$$

When W is least favorable (i.e. $W = \mathrm{Dir}(1/2)$), it is known that [22]

$$R_n(W) = \frac{K-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma(1/2)^K}{\Gamma(K/2)} + o(1) \tag{6.237}$$

and there is a corresponding result in the MDL literature [13]. Our bounds do not quite have this behavior, illustrating that neither of our upper or lower bounds are asymptotically tight. However, the advantage of our bounds is that they hold even in the non-asymptotic case.

Chapter 7

Companders for Quantization

In this chapter, we continue to study quantizing large alphabet discrete probability distributions using Kullback-Leibler (KL) divergence as the primary loss between the true distribution and its reconstruction. In order to quantize quickly and efficiently, we focus on using *companders*. This method applies predetermined function to each value of the probability distribution independently, so that quantization is computationally easy. (Also recall we showed in the previous chapter that for symmetric Dirichlet priors, methods using scalar quantization are close to optimal.) We develop and analyze a *minimax compander* which gives performance guarantees even when the probability distribution we are trying to quantize is chosen to be the worst case. Empirically, this compander is able to produce small KL divergence loss on real-world data.

Chapter Organization We go over what a compander is in Section 7.2 and show our main results in Section 7.3. A history of companders and other previous works is given in Section 7.4. The proof for the asymptotic KL loss of companders is in Section 7.5. The development of the minimax compander is in Section 7.6. Worst-case divergence covering results for the minimax compander are given in Section 7.7.

7.1 Introduction and Motivation

As more and more data is collected, more and more data needs to be stored. While many current technological innovations are aimed at increasing the amount of data we can store per amount of physical hardware, another direction which can help meet increasing data storage demands to is to reduce the physical resources needed to store data by considering *lossy* storage. This method quantizes data which loses some information, but in a way which preserves the information content that we consider most important. Typically, we quantify what is the important content by using a loss or distortion function. We will choose the distortion function to the be the KL divergence between the original and quantized version.

The precise scenario we will study is quantizing discrete probability distributions on large alphabets. For each P over K symbols, we want to quantize P to another probability distribution Q, where Q is selected from a predetermined set $\mathcal{Q} = \{Q_{(1)}, ..., Q_{(M)}\}$ of size M. This allows us to store each P with $\log_2 M$ bits. Choosing M so that $\log_2 M$ is small reduces the amount of storage data needed to store P.

While reducing storage is a our objective, we also prioritize making the quantization algorithm simple to implement, so that it can be added to existing systems which store probabilities without much overhead.

Our main result is that we developed a technique called the *(approximate) minimax compander*, which we advocate for regular use. When we are quantizing to 8-bits per symbol and when the number of symbols is large ($K = 10^5$), compared to uniform quantizing (likely default method), using the minimax compander reduces the KL divergence loss from 10^{-1} to 10^{-4}, a significant amount of saving in terms of the fidelity of the quantization.

Notational Notes Instead of P or Q, we use vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to represent probabilities.

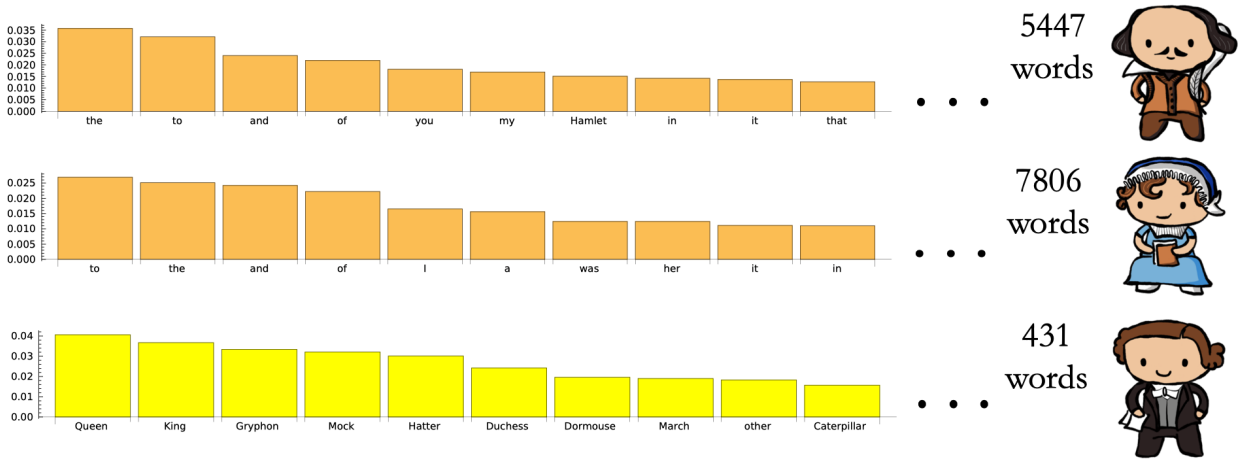


Figure 7.1: An application of quantizing probabilities is to store word frequencies in books. We give a few examples of frequencies of words (in sorted order by largest probability) in different books. The third frequency shown is from a bigram model, where we show the frequency of words appearing after the word "the".

7.2 Compander Basics and Definitions

Our goal is to find a *quantization scheme* on $\triangle_{K-1}$ (probability simplex of alphabet size K) minimizing the KL (Kullback-Leibler) divergence between probability vectors and their representations. In this chapter, we consider only scalar quantization methods, since we showed in Chapter 6 that for Dirichlet priors on the simplex, methods using scalar quantization perform nearly as well as optimal vector quantization; scalar quantization is typically simpler and faster to use, and can be parallelized easily. Our scalar quantization function is based on *companders* (portmanteau of 'compressor' and 'expander').

Encoding

Companders require two things: a monotonically increasing function $f : [0,1] \to [0,1]$ (we denote the set of such functions as $\mathcal{F}$) and an integer N representing the number of quantization levels, or *granularity*. To quantize $x \in [0,1]$, the compander computes $f(x)$ and applies a uniform quantizer with N levels, i.e. encoding x to $\text{Enc}(x) = n \in [N]$ if $f(x) \in (\frac{n-1}{N}, \frac{n}{N}]$; this is equivalent to $\text{Enc}(x) = \lceil f(x)N \rceil$.

This encoding system induces *bins* over $[0,1]$:

$$x \in f^{-1}\Big(\Big(\frac{n-1}{N}, \frac{n}{N}\Big]\Big) \iff \text{Enc}(x) = n \tag{7.1}$$

where f^{-1} denotes the preimage in f. We denote these bins as $I^{(n)}$, and the bin containing x at granularity N as $I^{(n_N(x))}$.

To simplify the problem and algorithm, we use the same f for each element of the probability vector $\boldsymbol{x} \in \triangle_{K-1}$.

For a graphical example of compander encoding, see Figure 7.2. We give a graphical example of the bins given by compander function f in Figure 7.3.

Decoding

To decode $n \in [N]$, we pick some $\widehat{y}^{(n)} \in I^{(n)}$ to represent all $x \in I^{(n)}$; for a given x (at granularity N), its representation is denoted $\widehat{y}(x) = \widehat{y}^{(n_N(x))}$. This is usually the *midpoint* of the bin or, if x is drawn

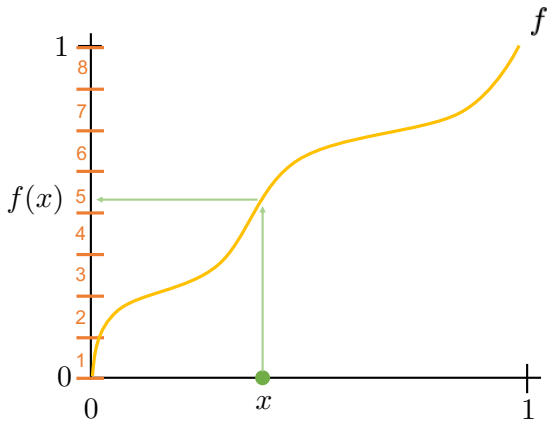


Figure 7.2: For compander function f, to find what input x encodes to, we will use a uniform quantizer on the range $f(x)$. Since in the illustration, $f(x)$ lands in the fifth bin, $\text{Enc}(x) = 5$. (Note that here $N = 9$ since in our experiments we make zero its own interval.)

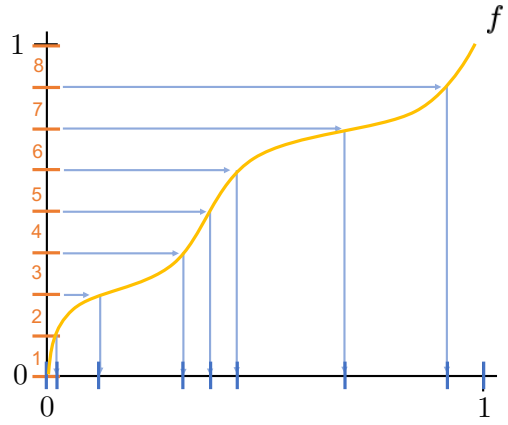


Figure 7.3: Illustration of the bins given by compander function f.

randomly from a prior,[1] the *centroid* (the mean within bin $I^{(n_N(x))}$). The midpoint of $I^{(n)}$ can be computed as $(f^{-1}(\frac{n-1}{N}) + f^{-1}(\frac{n}{N}))/2$.

The use of scalar quantization makes it difficult to ensure that the decoded values will sum to one, so we normalize. Let $\boldsymbol{x} = (x_1, .., x_K)$ be the input probability vector; then, letting

$$y(x_i) = \frac{\widehat{y}(x_i)}{\sum_{j=1}^{K} \widehat{y}(x_j)} \tag{7.2}$$

the vector $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) = (y(x_1), \ldots, y(x_K)) \in \triangle_{K-1}$ is the output of the compander. We refer to $\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{y}}(\boldsymbol{x}) = (\widehat{y}(x_1), \ldots, \widehat{y}(x_K))$ as the *raw* reconstruction of $\boldsymbol{x}$, and $\boldsymbol{y}$ as the *normalized* reconstruction. (We generally use the $\widehat{\cdot}$ accent to mark values dependent on the raw reconstruction.)

Thus, any $\boldsymbol{x} \in \triangle_{K-1}$ requires $K \log_2 N$ bits to store; to encode and decode, only f and N need to be stored (as well as the prior if using centroid decoding). Another major advantage of companders is that a single f can work well over all choices of N, making the design more flexible.

KL divergence loss

The loss incurred by representing $\boldsymbol{x}$ as $\boldsymbol{y}(\boldsymbol{x})$ is the KL divergence

$$D_{\mathrm{KL}}(\boldsymbol{x} \| \boldsymbol{y}(\boldsymbol{x})) = \sum_{i=1}^{K} x_i \log \frac{x_i}{y(x_i)} \,. \tag{7.3}$$

Distributions from a prior

Much of our work concerns the case where $\boldsymbol{x} \in \triangle_{K-1}$ is drawn from some prior $P_{\boldsymbol{x}}$ (to be commonly denoted as simply P), whose marginals are absolutely continuous. Using a single f for each entry also means we can WLOG assume that P is symmetric over the alphabet, since permuting the symbol indices doesn't affect the KL divergence. We denote the set of such priors as $\mathcal{P}_K^{\triangle}$.

We let $\mathcal{P}$ denote the class of absolutely continuous probability distributions on $[0, 1]$, represented by their probability density functions (PDFs), denoted p. Let $\mathcal{P}_{1/K} \subset \mathcal{P}$ be the set of p where $\mathbb{E}_{X \sim p}[X] = 1/K$. Note that $P \in \mathcal{P}_K^{\triangle}$ implies its marginals are in $\mathcal{P}_{1/K}$.

Expected loss and preliminary results

For $P \in \mathcal{P}_K^{\triangle}$, $f \in \mathcal{F}$ and granularity N, we define the *expected loss*:

$$\mathcal{L}_K(P, f, N) = \mathbb{E}_{\boldsymbol{X} \sim P}[D_{\mathrm{KL}}(\boldsymbol{X} \| \boldsymbol{y}(\boldsymbol{X}))] \,. \tag{7.4}$$

This is the value we want to minimize.

Note that $\mathcal{L}_K(P, f, N)$ can almost be decomposed into a sum of K separate expected values (one per entry), except the normalization step (7.2) depends on the vector as a whole. Hence, we define the *raw loss*:

$$\widehat{\mathcal{L}}_K(P, f, N) = \mathbb{E}_{\boldsymbol{X} \sim P}\Big[\sum_{i=1}^{K} X_i \log(X_i / \widehat{y}(X_i))\Big] \tag{7.5}$$

We also define for $p \in \mathcal{P}$, the *single-symbol loss* as

$$\widehat{L}(p, f, N) = \mathbb{E}_{X \sim p}\big[X \log(X / \widehat{y}(X))\big] \tag{7.6}$$

The raw loss is useful because it bounds the (normalized) expected loss and is decomposable into single-symbol losses:

[1] Priors on $\triangle_{K-1}$ induce priors over $[0, 1]$ for each entry.

Proposition 23. *For $P \in \mathcal{P}_K^{\triangle}$ with marginals p, under centroid decoding*

$$\mathcal{L}_K(P, f, N) \le \widehat{\mathcal{L}}_K(P, f, N) = K\,\widehat{L}(p, f, N) \tag{7.7}$$

To derive our results about worst-case priors (for instance, Theorem 18), we will also be interested in $\widehat{L}(p, f, N)$ even when p is not known to be a marginal of some $P \in \mathcal{P}_K^{\triangle}$.

Proof. Since

$$\begin{aligned}\mathcal{L}(P, (f_k), N) &= \mathbb{E}_{\boldsymbol{X}\sim P} D_{\mathrm{KL}}(\boldsymbol{X}||\boldsymbol{Y}) \\ &= \widehat{\mathcal{L}}_K(P, (f_k), N) + \mathbb{E}_{\boldsymbol{X}\sim P}\left[\log\left(\sum_{k=1}^{K}\widetilde{Y}_k\right)\right]\end{aligned} \tag{7.8}$$

it remains to show that

$$\mathbb{E}_{\boldsymbol{X}\sim P}\left[\log\left(\sum_{k=1}^{K}\widetilde{Y}_k\right)\right] \le 0 \tag{7.9}$$

Since $\mathbb{E}[\widetilde{Y}_k] = \mathbb{E}[X_k]$ for all k, $\sum_{k=1}^{K}\mathbb{E}[\widetilde{Y}_k] = \sum_{k=1}^{K}\mathbb{E}[X_k] = 1$. Because log is concave, by Jensen's Inequality

$$\mathbb{E}\Big[\log\Big(\sum_{k=1}^{K}\widehat{Y}_k\Big)\Big] \le \log\Big(\mathbb{E}\Big[\sum_{k=1}^{K}\widehat{Y}_k\Big]\Big) = \log(1) = 0\,. \tag{7.10}$$

□

Remark 6. *Raw and single-symbol loss lose much of their meaning without centroid decoding, because $\widehat{L}(p, f, N)$ will be dominated by the difference between $\mathbb{E}[X]$ and $\mathbb{E}[\widehat{y}(X)]$, potentially even making it negative (this 'first-order' effect gets canceled out by the normalization step). Centroid decoding solves this problem by ensuring $\mathbb{E}[X] = \mathbb{E}[\widehat{y}(X)]$.*

As we will show, when N is large these values are roughly proportional to N^{-2} (for well-chosen f) and hence we define the *asymptotic average single-symbol loss*:

$$\widehat{L}(p, f) = \lim_{N\to\infty} N^2\widehat{L}(p, f, N) \tag{7.11}$$

We similarly define $\widehat{\mathcal{L}}_K(P, f)$ and $\mathcal{L}_K(P, f)$. While the limit in (7.11) does not exist for *every* p, f, one can ensure it exists by choosing an appropriate f (which works against any $p \in \mathcal{P}$), and cannot gain much by not doing so, as we will show.

Finally, we define the following function on p and f:

$$L^{\dagger}(p, f) = \frac{1}{24}\int_0^1 p(x)f'(x)^{-2}x^{-1}\,dx \tag{7.12}$$

For a wide class of functions f, we will show that $\widehat{L}(p, f) = L^{\dagger}(p, f)$. Furthermore, $L^{\dagger}(p, f)$ is a lower bound on the asymptotic performance of f against p for *any* $f \in \mathcal{F}$.

7.3 Main Results

We demonstrate, theoretically and experimentally, the efficacy of companding for quantizing probability distributions with KL divergence loss, and show their optimal use. Though our theoretical results are asymptotic as $N \to \infty$ and focus on raw loss, the experimental (normalized) loss of the various companders closely tracks the (raw) loss predicted theoretically, even for quantization levels as low as $N = 256$ (i.e. where each value can be stored on 8 bits).

Theoretical Results

Definition 27. *Define $\mathcal{F}^\dagger \subseteq \mathcal{F}$ to be the set of f such that there exists $c > 0$ and $\alpha \in (0,1)$ for which $f(x) - cx^\alpha$ is still monotonically increasing.*

This is equivalent to $f'(x) \geq c\,\alpha\, x^{\alpha-1}$ for all x where f' is defined (which is almost everywhere since f is monotonic).

Theorem 16. *For any $p \in \mathcal{P}$ and $f \in \mathcal{F}^\dagger$,*

$$\widehat{L}(p,f) = L^\dagger(p,f)\,. \tag{7.13}$$

Additionally, for all $f \in \mathcal{F}$,

$$\liminf_{N\to\infty} N^2 \widehat{L}(p,f,N) \geq L^\dagger(p,f) \tag{7.14}$$

This shows how to compute the asymptotic single-symbol loss for any $f \in \mathcal{F}^\dagger$ (against any $p \in \mathcal{P}$). The lower bound shows that $f \notin \mathcal{F}^\dagger$ cannot outperform the formula $L^\dagger(p,f)$.

Theorem 17. *The best loss against source $p \in \mathcal{P}$ is*

$$\inf_{f\in\mathcal{F}} \widehat{L}(p,f) = \min_{f\in\mathcal{F}} L^\dagger(p,f) = \frac{1}{24}\Big(\int_0^1 (p(x)x^{-1})^{1/3} dx\Big)^3 \tag{7.15}$$

where the optimal compander against *p is*

$$f_p(x) = \arg\min\nolimits_{f\in\mathcal{F}} L^\dagger(p,f) = \frac{\int_0^x (p(t)t^{-1})^{1/3}\,dt}{\int_0^1 (p(t)t^{-1})^{1/3}\,dt} \tag{7.16}$$

(satisfying $f_p'(x) \propto (p(x)x^{-1})^{1/3}$).

If $f_p \in \mathcal{F}^\dagger$, it achieves the value from (7.15); otherwise, there is a sequence $f_{p,\delta} \in \mathcal{F}^\dagger$ such that

$$\lim_{\delta\to 0} \widehat{L}(p, f_{p,\delta}) = \lim_{\delta\to 0} L^\dagger(p, f_{p,\delta}) = L^\dagger(p, f_p) \tag{7.17}$$

Thus, even if $f_p \notin \mathcal{F}^\dagger$, there is a way to approximate $L^\dagger(p,f_p)$ by functions in $\mathcal{F}^\dagger$ and hence achieve asymptotic performance arbitrarily close to optimal, as per Theorem 16. This shows there is no real advantage to using $f \notin \mathcal{F}^\dagger$; thus we can now restrict our analysis to $f \in \mathcal{F}^\dagger$, for which (7.12) holds.

While in practice we do not always know the prior P on the simplex, we can show the existence of a particular compander which performs well against *any* prior.

Proposition 24. *For alphabet size $K > 4$, there is a unique $\frac{1}{4} \leq c_K \leq \frac{3}{4}$ such that if $a_K = (4/(c_K K \log K + 1))^{1/3}$ and $b_K = 4/a_K^2 - a_K$, then the following density is in $\mathcal{P}_{1/K}$:*

$$p_K^*(x) = (a_K x^{1/3} + b_K x^{4/3})^{-3/2} \tag{7.18}$$

Furthermore, $\lim_{K\to\infty} c_K = 1/2$.

We call p_K^* the *maximin single-symbol density.*

The optimal compander against p_K^* is the *minimax compander*:

$$f_K^*(x) = \frac{\mathrm{ArcSinh}(\sqrt{c_K K \log K}\, x)}{\mathrm{ArcSinh}(\sqrt{c_K K \log K})} \tag{7.19}$$

Note that $f_K^* \in \mathcal{F}^\dagger$ (see Remark 9). The source p_K^* and compander f_K^* then form an 'equilibrium':

Theorem 18. *The minimax compander f_K^* and maximin single-symbol density p_K^* satisfy*

$$\sup_{p\in\mathcal{P}_{1/K}} \widehat{L}(p, f_K^*) = \inf_{f\in\mathcal{F}^\dagger} \sup_{p\in\mathcal{P}_{1/K}} \widehat{L}(p, f) = \sup_{p\in\mathcal{P}_{1/K}} \inf_{f\in\mathcal{F}^\dagger} \widehat{L}(p, f) = \inf_{f\in\mathcal{F}^\dagger} \widehat{L}(p_K^*, f) \tag{7.20}$$

which is equal to $\widehat{L}(p_K^*, f_K^*)$ and satisfies

$$\widehat{L}(p_K^*, f_K^*) = \Theta(K^{-1}\log^2 K) \tag{7.21}$$

This theorem importantly implies the following:

Corollary 6. *For any prior $P \in \mathcal{P}_K^\triangle$,*

$$\mathcal{L}_K(P, f_K^*) \le \widehat{\mathcal{L}}_K(P, f_K^*) = \Theta(\log^2 K) \tag{7.22}$$

Furthermore, there exists a prior $P^* \in \mathcal{P}_K^\triangle$ such that

$$\inf_{f\in\mathcal{F}} \widehat{\mathcal{L}}_K(P^*, f) \ge \frac{K-1}{K}\frac{1}{2} \sup_{P\in\mathcal{P}_K^\triangle} \widehat{\mathcal{L}}_K(P, f_K^*) = \Theta(\log^2 K) \tag{7.23}$$

The constant-factor gap exists because $P^* \in \mathcal{P}_K^\triangle$ is a stronger constraint than $p_K^* \in \mathcal{P}_{1/K}$.

Using the minimax compander f_K^* on $P \in \mathcal{P}_K^\triangle$ with granularity N, we have a bound on the average KL divergence:

$$\mathbb{E}_{\boldsymbol{X}\sim P}[D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{Y})] = O\left(N^{-2}\log^2 K\right) + o\left(N^{-2}\right) \tag{7.24}$$

Remark 7. *For a given K, c_K can be computed numerically. While p_K^* must use the exact value (otherwise it's not in $\mathcal{P}_{1/K}$), the compander can approximate f_K^* (using an approximation of c_K, such as $1/2$ (see Appendix C.6) for large K) and still perform well. This suggests the* approximate minimax compander

$$f(x) = \frac{\mathrm{ArcSinh}(\sqrt{(1/2)K\log K\, x})}{\mathrm{ArcSinh}(\sqrt{(1/2)K\log K})} \tag{7.25}$$

as a simple compander whose performance is close to optimal.

Remark 8. *When b is the number of bits used to quantize each value in the probability vector, we get a loss on the order of $2^{-2b}\log^2 K$. If we use our divergence covering numbers (for worst-case loss instead of average-case), then the loss will an order between $2^{-2b\frac{K}{K-1}}$ and $2^{-2b\frac{K}{K-1}}\log K$. Thus, our result using companders is within a factor $2^{2b/(K-1)}\log^2 K$ of the optimal loss. (The bound $2^{-2b\frac{K}{K-1}}\log K$ from Theorem 2 is not associated with an explicit quantization scheme. One is only shown to exist.)*

Remark 9. *While the minimax compander might appear complicated,* $\mathrm{ArcSinh}(\sqrt{z}) = \log(\sqrt{z}+\sqrt{z+1})$ *is fairly simple. Taking the Taylor expansion also confirms that $f_K^* \in \mathcal{F}^\dagger$.*

Note that (7.16) *(Theorem 17) suggests that the natural form of an optimal compander against p is a normalized incomplete integral, which is hard to use; thus, the closed-form expression of the minimax compander is a welcome surprise.*

The above are all 'average case' results, where $\boldsymbol{X}$ is drawn from a prior P (which is fixed as $N\to\infty$). In the worst-case problem, $\boldsymbol{x}$ is chosen to maximize loss and can depend on N:

Theorem 19. *There exists constant c such that if $N > c\log K$, the minimax compander with midpoint decoding achieves*

$$\max_{\boldsymbol{x}\in\triangle_{K-1}} D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{y}) = \Theta\left(N^{-2}\log^2 K\right). \tag{7.26}$$

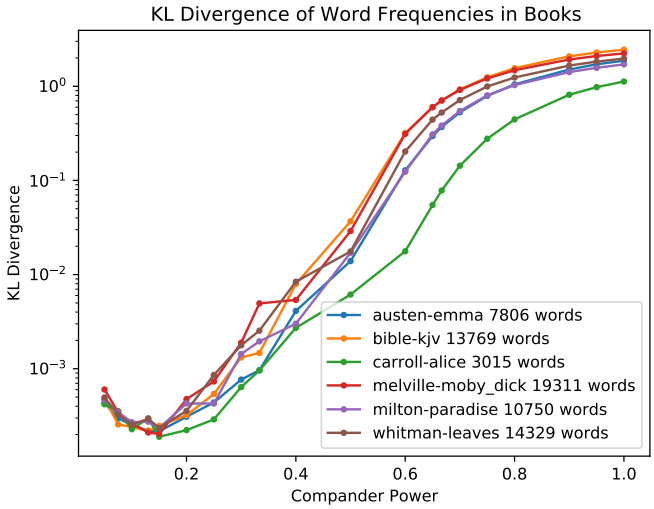


Figure 7.4: Power compander $f(x) = x^s$ performance with different powers s used to quantize frequency of words in books. Number of distinct words in each book is shown in the legend.

Experimental results

Our experiments verify that the approximate minimax compander (7.25) behaves as the theory predicts. We test it on three types of datasets: (i) random synthetic distributions drawn from the uniform prior over the simplex; (ii) frequency of words in books; and (iii) frequency of k-mers in DNA (for more details about these datasets, see Appendix C.1). We compare the minimax compander against three alternatives:

- **Truncation**: Values are quantized uniformly (equivalent to $f(x) = x$ or the "identity" compander), which truncates the least significant bits. This is the 'default' quantization option for values in $[0, 1]$.
- **Exponential Density Interval (EDI)**: This is the quantization method we used in an achievability proof for Theorem 13 given in (6.4). Recall that the EDI is designed for the uniform prior over the simplex.
- **Power Compander**: The compander where $f(x) = x^s$. We optimize s, both in theory and experimentally, and find that $s = \frac{1}{\log K}$ minimizes KL divergence. To see the effects of different powers s on the performance of the power compander, see Figure 7.4. For more on the power compander, see Appendix C.2.

Our main experimental results are given in Figure 7.5, showing the KL divergence between the original distribution $\boldsymbol{x}$ and its quantized version $\boldsymbol{y}$ versus alphabet size K. In the plots, the approximate minimax compander (7.25) performs well against all sources. For truncation, as the alphabet size K increases, the KL divergence increases as expected (larger alphabets are harder to quantize). The EDI quantizer works well for the synthetic uniform prior (as it should), but for real-world datasets like word frequency in books, it performs worse even than truncation. The power compander performs similarly to the minimax compander and is worse only by a constant[2].

The experiments demonstrate that the minimax compander achieves low loss on the entire ensemble of data (even for relatively small granularity, such as $N = 256$), validating our theoretical analysis. Currently, most systems which store probability values likely use the truncation method by default. However, our theoretical and experimental results show that using the minimax compander decreases the KL divergence loss by several orders of magnitude across a wide variety of datasets; additionally, its closed-form expression (and entrywise application) makes it simple to implement and computationally inexpensive. Thus it can be easily added to existing storage systems, yielding significant gains in fidelity for little cost.

[2]Theorem 19 also holds for the power compander with different constants.

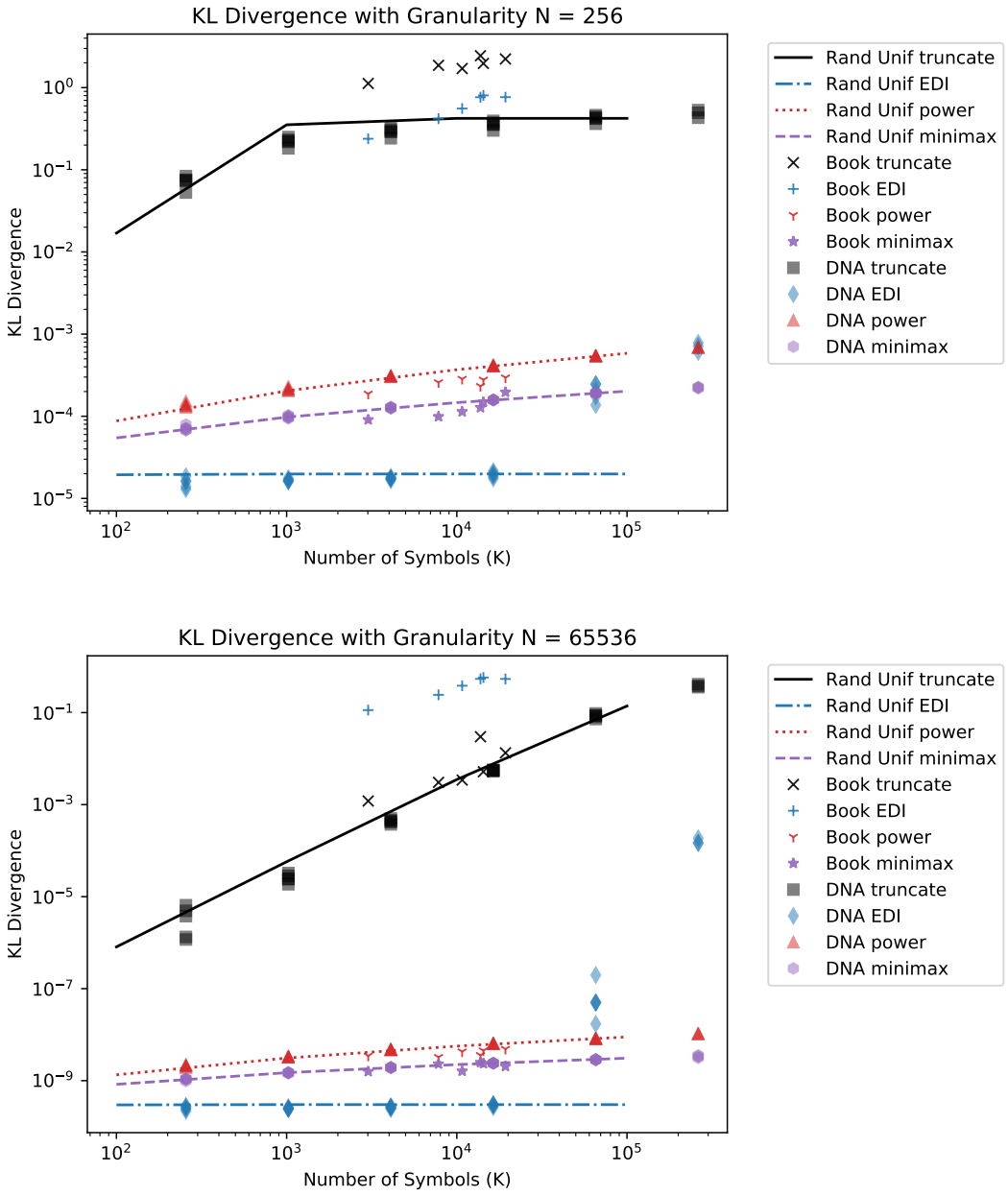


Figure 7.5: Plot comparing the performance of the truncation compander, the EDI compander, the power compander, and the approximate minimax compander (7.25) on probability distributions of various sizes.

7.4 Related Previous Works

7.4.1 Compander History

Companders (spelled "compandors" in some references) appeared in 1948 in a paper by Bennett [75], where he considers quantizing signals for newly developed high speed electronic devices. These devices deal with speech signals, where it is advantageous to give finer quantization levels to weaker signals and coarser levels to larger signals. This method was implemented by a transducer before and after the uniform quantizer. The transducer before the quantizer functions as a "compressor" and the transducer after functions an "expander". Bennett gives a first order approximation that the mean-square error in this system is given by

$$\frac{1}{12N^2}\int_a^b \frac{p(x)}{(E'(x))^2}dx \tag{7.27}$$

where N is the number quantization levels, a and b are the minimum and maximum values of the input signal, p is the probability density of the input signal, and E' is the slope of the compressor function E. This becomes the companding model for studying non-uniform quantization. This expression for mean-squared error is very similar to our result for average KL divergence loss in Theorem 17. The only two differences are that the x^{-1} term is missing and the constant out front is different.

Other have expanded on this line of work by Bennett. In Panter-Dite [76], the authors studied the same problem and determined the optimal compressor. They assume that the intervals they are quantizing to are small enough so that $p(x)$ is constant in the interval. If $\lambda(x)$ is the density of the quantization points, then the mean-squared error, denoted as $L_2(p,\lambda)$, is

$$L_2(p,\lambda) = \frac{1}{12N^2}\int_a^b p(x)\frac{1}{\lambda(x)^2}dx\,. \tag{7.28}$$

The optimal density for $\lambda(x)$ is to choose

$$\lambda^*(x) = \frac{p(x)^{1/3}}{\int_a^b p(x)^{1/3}dx} \tag{7.29}$$

which gives

$$L_2(p,\lambda^*) = \frac{1}{12N^2}\left(\int_a^b p(x)^{1/3}dx\right)^3. \tag{7.30}$$

Generalizations of Bennett's formula are also studied when instead of mean-square error, the loss is the expected rth moment loss $\mathbb{E}\|\cdot\|^r$. This is computed for vectors of length K in [77] and [78]. The loss in this scenario is given by

$$L_r(p,\lambda^*) = \beta N^{-r/K}\left(\int (p(\boldsymbol{x}))^{K/(K+r)}d\boldsymbol{x}\right)^{(K+r)/K} \tag{7.31}$$

for some value β which depends on K and r but not on p.

Practical Use of Companders

The typical examples of companders used in engineering are μ-law and A-law companders [79]. Both these companders have international standards[3] and both are primarily used for compressing human speech. It is believed that the human auditory system processes the logarithm of sound amplitudes. The idea is that a uniform quantizer would not give enough resolution to the sounds humans can distinguish, and thus the μ-law and A-law use companding by applying a logarithm to the signal amplitude. Panter-Dite and [80] also consider what happens when the compander function is (essentially) the μ-Law compander. They argue

[3]The μ-law compander is used in the US and Japan and the A-law is used in Europe.

for large enough constant μ, in the case of mean-squared error, the distortion becomes independent of the signal.

Though it is not relevant to any of our results, we analyze the performance of the μ-law and A-law companders in Appendix C.5. These companders do not preform too badly for quantizing large alphabet distributions if we replace the constants used in both companders with values which grow with K. The results for these have a slightly worse order of growth. They behave as $O\left(\frac{\log N}{N^2}\right)$ instead of $O\left(\frac{1}{N^2}\right)$ (which is the growth of the optimal compander for each prior.)

7.4.2 Information k-means

Companders have not been studied with KL divergence (as far as we know). However, a similar problem which does use KL divergence is the problem information k-means. The goal of information k-means is to cluster points $a \in \mathcal{A}$ which represents a set of features in high-dimensional space. Each a is associated with a distribution on some set $\mathcal{B}$, which we can write as $p_{A|B}(b|a)$ for each $b \in \mathcal{B}$. The goal of information k-means is to cluster each point a to a center $\tilde{a}$ to minimize the distortion

$$d(a, \tilde{a}) = D\big(p_{B|A}(\cdot|a) \| p_{B|A}(\cdot|\tilde{a})\big)\,. \tag{7.32}$$

The authors in [81] and [82] show that this distortion is the natural consequence for a problem where we want the cluster center to be such that the mutual information $I(A; \tilde{A})$ is small while $I(B; \tilde{A})$ is large, where $A, \tilde{A}, B$ represent the random variables for $a, \tilde{a}, b$ respectively.

The authors in [83] use (7.32) has the distortion for clustering nouns, where $p_{B|A}(b|a)$ gives the frequency over verbs b, that noun a is the direct object of verb b. This approach is able to produce semantically significant clusters. Other applications include using information k-means to cluster a distribution of ratings [84] and to cluster images of landmarks [85]. There are also other works that study clustering with a slightly different but related metric, such as Jensen-Shannon distortion [86, 87] or symmetric KL distortion [88].

Clustering and quantization are inherently similar since the goal in both is to find centers which are representative of the data in order to reduce complexity. However our framework fundamentally has a different goal than clustering. While our method will cluster points, the objective is that the reconstruction of each point can be easily mapped backed to a probability distribution for immediate use. This is also why we choose to use scalar quantization, so that this mapping can be done quickly and possibly in parallel. Using general clustering for this task will require that we store information about each cluster center, increasing the overhead for storage. Also, information k-means is generally used as a tool for analyzing existing data rather than for reducing storage bits.

7.5 Asymptotic Analysis

Following the ideas of Panter-Dite [76], the analysis in this section is intended for the case where the quantization is very fine. While this is not the case for our applications, we see that empirically, the same consequences apply when the quantization levels N is not that large. Similar to the work in Panter-Dite, we want to determine asymptotically what the correct compander is given a prior distribution. To do that, we want to show

$$\widehat{L}(p, f) = L^{\dagger}(p, f) = \frac{1}{24} \int_0^1 p(x) f'(x)^{-2} x^{-1}\, dx\,. \tag{7.33}$$

This is the key idea we are trying to capture with Theorem 16.

Remark 10. *Note that the expression $L^{\dagger}(p, f)$ resembles* (7.28) *except there is an additional divide by x and a difference in the constant.*

7.5.1 Sketch of Intuition Behind Theorem 16

Unfortunately, showing (7.33) is for every p and f is very difficult and possibly not true for pathological choices. (We note that [76] proved (7.28) by making assumptions that densities are uniform over small

intervals. We do not make this assumption, causing much of the difficulty.) This is why Theorem 16 requires that we define the classes $\mathcal{P}$ and $\mathcal{F}^\dagger$, to limit p to having densities and f to very nice functions. Still, the proof of Theorem 16 is very complicated. We are interested in showing (7.33) for functions p and f which might have discontinuities, areas where values are zero, and places where $f'(x) = 0$, which all complicate any proof.

However, serving the purpose of giving an explanation of where the formula for $L^\dagger(p, f)$ comes from, we give the following proof sketch for when:

1. Density p is continuous, bounded and greater than 0 everywhere on $[0, 1]$
2. Derivative f' is continuous and greater than 0 everywhere on $[0, 1]$ (this implies that f is strictly monotone).

We note that these conditions are not sufficient enough for problems we are interested in, but the proof of this is helpful for understanding, and thus we show it.

Proof sketch for very nice p and f'. We want to show that

$$\lim_{N\to\infty} N^2 \widehat{L}(p, f, N) = \frac{1}{24} \int_0^1 p(x)(f'(x))^{-2} x^{-1} dx \,. \tag{7.34}$$

For each $I^{(n)}$ where $n \in [N]$, we use y_n to denote the conditional expectation. We use $\tilde{I}^{(n)}$ to denote image of $I^{(n)}$ under f.

First, using Taylor expansion, we have that

$$\int_{I^{(n)}} p(x) x \log \frac{x}{y_n} dx = \int_{I^{(n)}} p(x) \left((x - y_n) + \frac{(x - y_n)^2}{2y_n} - O((x - y)^3) \right) dx \tag{7.35}$$

$$= \frac{1}{2} \int_{I^{(n)}} p(x) \frac{(x - y_n)^2}{y_n} dx + O\left(\frac{1}{N^4}\right) \tag{7.36}$$

and thus

$$N^2 \widehat{L}(p, f, N) = N^2 \left(\frac{1}{2} \sum_{n\in[N]} \int_{I^{(n)}} p(x) \frac{(x - y_n)^2}{y_n} dx \right) + O\left(\frac{1}{N}\right) \tag{7.37}$$

We can ignore the $O(1/N)$ since it will not affect our asymptotic result.

To find the term in the integral, we will do a substitution where $z = f(x)$. To make notation easier to read, let $g = f^{-1}$. The function g is well-defined since we can assume that f is strictly monotone . Then

$$dx = g'(z) dz \,. \tag{7.38}$$

We will also use $\tilde{z}_n = f(y_n)$. This gives

$$N^2 \widehat{L}(p, f, N) = N^2 \frac{1}{2} \sum_{n\in[N]} \int_{\tilde{I}^{(n)}} p(g(z)) \frac{(g(z) - g(\tilde{z}_n))^2}{g(\tilde{z}_n)} g'(z) dz \tag{7.39}$$

$$= N^2 \frac{1}{2} \sum_{n\in[N]} \int_{\tilde{I}^{(n)}} \frac{p(g(z))}{p(g(\tilde{z}_n))} \frac{p(g(\tilde{z}_n))}{g(\tilde{z}_n)} (z - \tilde{z}_n)^2 \frac{(g(z) - g(\tilde{z}_n))^2}{(z - \tilde{z}_n)^2} \frac{g'(z)}{g'(\tilde{z}_n)} g'(\tilde{z}_n) dz \tag{7.40}$$

$$= N^2 \frac{1}{2} \sum_{n\in[N]} \frac{p(g(\tilde{z}_n))}{g(\tilde{z}_n)} (g'(\tilde{z}_n))^3 \int_{\tilde{I}^{(n)}} (z - \tilde{z}_n)^2 \left(\frac{p(g(z))}{p(g(\tilde{z}_n))} \right) \frac{\frac{(g(z) - g(\tilde{z}_n))^2}{(z - \tilde{z}_n)^2}}{(g'(\tilde{z}_n))^2} \left(\frac{g'(z)}{g'(\tilde{z}_n)} \right) dz \,. \tag{7.41}$$

Next, we bound many of the terms in the integral. Since p is continuous, because p is a function on a compact set, we can assume p must also be uniformly continuous. Thus, we can find a sequence of $\varepsilon_N \to 0$ as $N \to \infty$ so that for each $n \in [N]$,

$$\left(\frac{p(g(z))}{p(g(\tilde{z}_n))} \right) = 1 + \varepsilon_N \tag{7.42}$$

Similarly, if f' is continuous on an interval (thus uniformly continuous), for some sequence $\nu_N \to 0$ as $N \to \infty$ so that for each $n \in [N]$ except where a discontinuity occurs,

$$\left(\frac{g'(z)}{g'(\tilde{z}_n)}\right) = 1 + \nu_N \tag{7.43}$$

Finally, using the definition of a derivative, for some sequence $\eta_N \to 0$ as $N \to \infty$ so that for each $n \in [N]$

$$\frac{\frac{(g(z)-g(\tilde{z}_n))^2}{(z-\tilde{z}_n)^2}}{(g'(\tilde{z}_n))^2} = 1 + \eta_N \tag{7.44}$$

Next, each interval $\tilde{I}^{(n)}$ is has length $1/N$. The value of the integral $\int_{\tilde{I}^{(n)}}(z-\tilde{z}_n)^2 dz$ depends on the location of $\tilde{z}_n$ within the interval. Let $z_{n,mid}$ be the midpoint of $\tilde{I}^{(n)}$. Let $\delta_{n,N} = z_{n,mid} - \tilde{z}_n$. Then

$$\int_{\tilde{I}^{(n)}} (z-\tilde{z}_n)^2 dz = \int_{\tilde{z}_n+\delta_{n,N}-\frac{1}{2N}}^{\tilde{z}_n+\delta_{n,N}+\frac{1}{2N}} (z-\tilde{z}_n)^2 dz \tag{7.45}$$

$$= \int_{\delta_{n,N}-\frac{1}{2N}}^{\delta_{n,N}+\frac{1}{2N}} z^2 dz \tag{7.46}$$

$$= \frac{1}{3}\left(\frac{1}{2N}+\delta_{n,N}\right)^3 - \frac{1}{3}\left(\frac{-1}{2N}+\delta_{n,N}\right)^3 \tag{7.47}$$

$$= \frac{1}{12N^3} + \frac{\delta_{n,N}^2}{N} \tag{7.48}$$

The value in (7.48) is at minimum $1/(12N^3)$ and at maximum $1/(3N^3)$. As the size of $\tilde{I}^{(n)}$ decreases, we expect the value of p over the integral to be flat. Thus, the value of $\delta_{n,N}$ should go to zero. We will write (7.48) as $\frac{\kappa_{n,N}}{12N^3}$ where $\kappa_{n,N} \to 1$.

This gives that

$$N^2\widehat{L}(p,f,N) = \frac{1}{2}N^2 \sum_n \frac{p(g(\tilde{z}_n))}{g(\tilde{z}_n)}(g'(\tilde{z}_n))^3 \left(\frac{\kappa_{n,N}}{12N^3}\right)(1+\varepsilon_N)(1+\nu_N)(1+\eta_N) \tag{7.49}$$

$$= (1+\varepsilon_N)(1+\nu_N)(1+\eta_N)\left(\max_n \kappa_{n,N}\right)\frac{1}{24}\sum_n \frac{1}{N}\frac{p(g(\tilde{z}_n))}{g(\tilde{z}_n)}(g'(\tilde{z}_n))^3 \tag{7.50}$$

$$= \frac{\alpha_N}{24}\sum_n \frac{1}{N}\frac{p(g(\tilde{z}_n))}{g(\tilde{z}_n)}(g'(\tilde{z}_n))^3 \tag{7.51}$$

$$\to \frac{1}{24}\int_0^1 \frac{p(g(z))}{g(z)} g'(z)^3 dz \tag{7.52}$$

where α_N collects the constants approaching 1. The other terms in (7.51) are exactly the terms used to approach a Riemann integral over all $\tilde{z}_n$. Integrating with respect to x instead of z gives

$$N^2\widehat{L}(p,f,N) \to \frac{1}{24}\int_0^1 \frac{p(x)}{x}\frac{1}{(f'(x))^2}dx \tag{7.53}$$

□

Showing the proof of Theorem 16 is much more complicated. We only give a sketch and provide more detail on one aspect of the proof.

7.5.2 Proof Sketch of Theorem 16

In this section we sketch the proof of Theorem 16. For an interval $I \subseteq [0,1]$, we denote its size as r_I, its midpoint as $\bar{y}_I$, and its centroid (under fixed p) as $\widetilde{y}_I := \mathbb{E}_{X\sim p}[X \,|\, X \in I]$. If $\mathbb{P}_{X\sim p}[X \in I] = 0$ we set $\widetilde{y}_I = \bar{y}_I$ by default.

Given single-source density p and compander f, to prove Theorem 16 we construct the following functions: the *local loss function at granularity* N, defined as

$$g_N(x) = N^2 \mathbb{E}[X \log(X/\widehat{y}(X)) \,|\, X \in I^{(n_N(x))}] \tag{7.54}$$

and the *asymptotic local loss function*, defined as

$$g(x) = \frac{1}{24} f'(x)^{-2} x^{-1} \,. \tag{7.55}$$

The function g_N basically takes each x and returns the expected loss for $X \sim p$ which fall in the same bin as x, thus averaging the losses in each bin (and keeping the overall average the same). The goal is thus to show that

$$\lim_{N\to\infty} \int g_N \, dp = \int g \, dp \,. \tag{7.56}$$

We get this by showing $\lim_{N\to\infty} g_N(X) \to g(X)$ almost surely for $X \sim p$.

Let $r_N(x) = r_{I^{(n_N(x))}}$ be the size of the bin $I^{(n_N(x))}$.

Lemma 22. *If f' is well-defined at x, then*

$$\lim_{N\to\infty} N \, r_N(x) = f'(x)^{-1} \tag{7.57}$$

(including the limit going to infinity when $f'(x) = 0$).

(The key idea of this lemma is that the size of the intervals can be approximated by the derivative of f.)

For any interval I, we want to compare p over I with a uniform distribution; to do this, we define a probability distribution p_I over $[-1, 1]$ (but with support of width 1) which is a shifted and scaled version of $p|_I$ (p restricted to I) such that $\mathbb{E}_{Z\sim p_I}[Z] = 0$. Specifically, if we have $X \sim p \,|\, X \in I$ then

$$Z = (X - \widetilde{y}_I)/r_I \tag{7.58}$$

is distributed according to p_I. It turns out that if $p(x)$ is well-defined (as a derivative of the cumulative distribution function) and positive, then any sufficiently small interval I around x should produce p_I close to the uniform distribution on $[-1/2, 1/2]$. Letting

$$d(p, q) = \|F_p - F_q\|_\infty \tag{7.59}$$

denote the maximum difference between the CDFs of p and q, and letting $\mathbf{1}_{[-1/2,1/2]}$ be the indicator function of $[-1/2, 1/2]$:

Proposition 25. *For $p \in \mathcal{P}$, if $X \sim p$ then almost surely*

$$\lim_{r\to 0} d(p_{I_r}, \mathbf{1}_{[-1/2,1/2]}) = 0 \tag{7.60}$$

for any set of intervals containing X where I_r has width r.

Thus, if $X \sim p$ we know that p is almost surely close to uniform on sufficiently small I around p; and Lemma 22 shows that the bins about X shrink as $N^{-1} f'(X)^{-1} + o(N^{-1})$.

Then we can estimate the expected loss within the bin as:

Proposition 26. *Let I be such that $d(p_I, \mathbf{1}_{[-1/2,1/2]}) \le \varepsilon < 1/2$ and $r_I \le \bar{y}_I$. Then,*

$$\left| \mathbb{E}[X \log(X/\widetilde{y}_I) \,|\, X \in I] - \frac{1}{24} \frac{r_I^2}{\widetilde{y}_I} \right| \le \frac{1}{2} \frac{r_I^2}{\widetilde{y}_I} \varepsilon + O\left(\frac{r_I^3}{\widetilde{y}_I^2} \right). \tag{7.61}$$

(To prove the above proposition, we do need a step where we bound KL divergence by the chi-squared divergence.) When N is large,

$$r_N(X) = N^{-1} f'(X)^{-1} + o(N^{-1}) \tag{7.62}$$

(and hence $\widetilde{y}(X) = X + O(N^{-1})$ since $\widetilde{y}(X)$ must fall in the same interval). Since $I^{(n_N(X))}$ shrinks to size 0 about X as $N \to \infty$, the above results give:

$$d(p_{I^{(n_N(X))}}, \mathbf{1}_{[-1/2,1/2]}) = o(1) \tag{7.63}$$

$$\implies \left|g_N(X) - g(X)\right| = \left|g_N(X) - \frac{1}{24}\frac{f'(X)^{-2}}{X}\right| \tag{7.64}$$

$$= \left|g_N(X) - \frac{N^2}{24}\frac{r_N(X)^{-2}}{\widetilde{y}(X)}\right| + o(1) \tag{7.65}$$

$$= o(1) \tag{7.66}$$

almost surely if $X \sim p$. Fatou's Lemma then gives the lower bound (7.14) from Theorem 16.

To get the full result (7.12) for $f \in \mathcal{F}^\dagger$, we build a function h^* for which $h^*(x) \geq g_N(x)$ for all x and $\int h^* \, dp < \infty$ (assuming $\int g \, dp < \infty$), allowing us to apply the Dominated Convergence Theorem. Since $f \in \mathcal{F}^\dagger$, there are $c > 0$ and $\alpha \in (0, 1)$ such that $f(x) - cx^\alpha$ is monotonic. Hence all the bins induced by f are smaller than those induced by cx^α (treating it as a compander). Using the bound

$$\mathbb{E}[X \log(X/\widetilde{y}_I) \,|\, X \in I] \leq \frac{1}{2}\frac{r_I^2}{\overline{y}_I} \tag{7.67}$$

we show that

$$h^*(x) = 2^{2/\alpha+1} 24 g(x) + 2^{1/\alpha+2} c^{-1/\alpha} \tag{7.68}$$

dominates all local loss functions induced by cx^α, thus also dominating all local loss functions induced by f. Since g is integrable over p, so is h^*, completing the proof of Theorem 16.

In the next section, we discuss how to get (7.68).

7.5.3 Proof For Power Companders

Then, we define for $x \in [0, 1]$ and $\varepsilon \in (0, 1]$ the following:

$$J_\varepsilon(x) = f^{-1}\big([f(x) - \varepsilon/2, f(x) + \varepsilon/2] \cap [0, 1]\big) \tag{7.69}$$

$$r_\varepsilon(x) = |J_\varepsilon(x)| \tag{7.70}$$

$$h_\varepsilon(x) = 4\varepsilon^{-2} r_\varepsilon(x)^2 / \max(x, f^{-1}(\varepsilon/2)) \tag{7.71}$$

We now consider what happens for a compander f such that $f'(x) \geq c\,\alpha\, x^{\alpha-1}$ for some $\alpha \in (0, 1), c > 0$ for all x; such f satisfy $f(x) \geq cx^\alpha$, but the converse is not necessarily true (since $f(x) \geq cx^\alpha$ might have flat, or nearly flat, sections).

For such a function, the intervals $r_\varepsilon(x)$ are always smaller than they would be for cx^α as a compander, and hence an h^* which dominated all h_ε for the compander $c\,x^\alpha$ will work for f as well. Therefore, we assume that $f(x) = c\,x^\alpha$ and the result will apply for all f such that $f'(x) \geq c\,\alpha\, x^{\alpha-1}$. Since $f(x) = cx^\alpha$, we have $f^{-1}(z) = (z/c)^{1/\alpha} = c^{-1/\alpha} z^{1/\alpha}$.

We split things into two cases (each expressed in three different equivalent ways):

(i) $f(x) \geq \varepsilon/2 \iff x \geq c^{-1/\alpha}(\varepsilon/2)^{1/\alpha} \iff \varepsilon \leq 2cx^\alpha$

(ii) $f(x) \leq \varepsilon/2 \iff x \leq c^{-1/\alpha}(\varepsilon/2)^{1/\alpha} \iff \varepsilon \geq 2cx^\alpha$

Case (i) is when $J_\varepsilon(x)$ does not reach 0 on the lower end, while Case (ii) is when it does. Note that in Case (i), since $f(x) \geq \varepsilon/2$, we have $\max(x, f^{-1}(\varepsilon/2)) = x$, while in Case (ii) we have

$$\max(x, f^{-1}(\varepsilon/2)) = f^{-1}(\varepsilon/2) = c^{-1/\alpha}(\varepsilon/2)^{1/\alpha} \tag{7.72}$$

In both cases we can derive that:

$$J_\varepsilon(x) \subseteq \left[c^{-1/\alpha}(f(x)-\varepsilon/2)^{1/\alpha}, c^{-1/\alpha}(f(x)+\varepsilon/2)^{1/\alpha}\right] \tag{7.73}$$

$$r_\varepsilon(x) \le c^{-1/\alpha}\big((f(x)+\varepsilon/2)^{1/\alpha} - (f(x)-\varepsilon/2)^{1/\alpha}\big) \tag{7.74}$$

Case (i) Since $\alpha \in (0,1)$, we know that $z^{1/\alpha}$ is convex, and hence we can bound with the right-hand derivative:

$$\big((f(x)+\varepsilon/2)^{1/\alpha} - (f(x)-\varepsilon/2)^{1/\alpha}\big) = \int_{f(x)-\varepsilon/2}^{f(x)+\varepsilon/2} \alpha^{-1} z^{1/\alpha-1}\,dz \tag{7.75}$$

$$\le \int_{f(x)-\varepsilon/2}^{f(x)+\varepsilon/2} \alpha^{-1}(f(x)+\varepsilon/2)^{1/\alpha-1}\,dz \tag{7.76}$$

$$= \varepsilon\alpha^{-1}(f(x)+\varepsilon/2)^{1/\alpha-1} \tag{7.77}$$

Thus we can get:

$$r_\varepsilon(x) \le \varepsilon c^{-1/\alpha}\alpha^{-1}\big(f(x)+\varepsilon/2\big)^{1/\alpha-1} \tag{7.78}$$

$$\le \varepsilon c^{-1/\alpha}\alpha^{-1}(2f(x))^{1/\alpha-1} \tag{7.79}$$

$$= \varepsilon c^{-1/\alpha}\alpha^{-1}2^{1/\alpha-1}(cx^\alpha)^{1/\alpha-1} \tag{7.80}$$

$$= 2^{1/\alpha-1}\varepsilon\big(c^{-1}\alpha^{-1}x^{1-\alpha}\big) \tag{7.81}$$

$$= 2^{1/\alpha-1}\varepsilon f'(x)^{-1} \tag{7.82}$$

Noting that Case (i) is exactly when $h_\varepsilon(x) = 4\varepsilon^{-2}r_\varepsilon(x)/x$, we get

$$h_\varepsilon(x) = 4\varepsilon^{-2}r_\varepsilon(x)^2x^{-1} \tag{7.83}$$

$$\le 4\varepsilon^{-2}\big(2^{1/\alpha-1}\varepsilon f'(x)^{-1}\big)^2x^{-1} \tag{7.84}$$

$$= 2^{2/\alpha+1}f'(x)^{-2}x^{-1} \tag{7.85}$$

which holds for all $\varepsilon \le 2cx^\alpha$.

Case (ii) In this case, the lower bound of $f(J_\varepsilon(x))$ is 0 and the upper bound is no more than ε (since $f(x)+\varepsilon/2 \le \varepsilon/2+\varepsilon/2 = \varepsilon$). Noting that $f^{-1}(0) = 0$, we get that

$$r_\varepsilon(x) \le f^{-1}(\varepsilon) = c^{-1/\alpha}\varepsilon^{1/\alpha} \tag{7.86}$$

Thus, we can plug this in straight away (noting that in this case $h_\varepsilon(x)$ divides by $f^{-1}(\varepsilon/2)$):

$$h_\varepsilon(x) = 4\varepsilon^{-2}r_\varepsilon(x)/f^{-1}(\varepsilon/2) \tag{7.87}$$

$$\le 4c^{-2/\alpha}\varepsilon^{2/\alpha-2}\big(c^{-1/\alpha}(\varepsilon/2)^{1/\alpha}\big)^{-1} \tag{7.88}$$

$$= 2^{1/\alpha+2}c^{-1/\alpha}\varepsilon^{1/\alpha-2} \tag{7.89}$$

When $\alpha \le 1/2$, the $\varepsilon^{1/\alpha-2}$ term is maximized by making ε as large as possible, which is 1 (because we only define $h_\varepsilon(x)$ for ε up to 1 and hence $h_\varepsilon(x) \le 2^{1/\alpha+2}c^{-1/\alpha}$, which is a constant, for all $\varepsilon \ge 2cx^\alpha$. When $\alpha > 1/2$, the value is maximized by making ε as small as possible, which in Case (ii) means $\varepsilon = 2cx^\alpha$. This means

$$h_\varepsilon(x) \le 2^{1/\alpha+2}c^{-1/\alpha}(2cx^\alpha)^{1/\alpha-2} \tag{7.90}$$

$$= 2^{2/\alpha}c^{-2}x^{1-2\alpha} \tag{7.91}$$

$$= 2^{2/\alpha}\alpha^2(c^{-1}\alpha^{-1}x^{1-\alpha})^2x^{-1} \tag{7.92}$$

$$= 2^{2/\alpha}\alpha^2 f'(x)^{-2}x^{-1} \tag{7.93}$$

This means we have one bound for $\varepsilon \leq 2cx^\alpha$ and another for $\varepsilon \geq 2cx^\alpha$; the *overall* bound can then be found by taking the maximum of the two, and if necessary by adding them. When $\alpha \in [1/2, 1)$, we see that $2^{2/\alpha+1} = 2 \cdot 2^{2/\alpha} > 2^{2/\alpha}\alpha^2$, and when $\alpha \in (0, 1/2)$ we have $h_\varepsilon(x) \leq 2^{1/\alpha+2}c^{-1/\alpha}$ and hence our overall bound is

$$h_\varepsilon(x) \leq 2^{2/\alpha+1} f'(x)^{-2} x^{-1} + 2^{1/\alpha+2} c^{-1/\alpha} = h^*(x) \tag{7.94}$$

which holds for every x, ε pair for every $\alpha \in (0, 1)$ and $c > 0$. For any fixed α, c, we have that h^* is a simply a constant multiple of g plus a constant, so $\int g\, dp < \infty$ implies

$$\int_{[0,1]} h^*\, dp = 24\cdot \leq 2^{2/\alpha+1} \int_{[0,1]} g\, dp + 2^{1/\alpha+2} c^{-1/\alpha} < \infty \tag{7.95}$$

Since h^* dominates all h_ε, it also dominates all g_N (as $h_{2/N}$ dominates g_N; and therefore we can apply the Dominated Convergence Theorem to achieve our result.

7.6 Minimax Compander

The purpose of this section is to optimize over the expression for $L^\dagger(p, f)$. This is possible due to the following:

Proposition 27. *The function $L^\dagger(p, f)$ is convex in f for any p and linear (and therefore concave) in p for any f.*

Proof. First, we show that $L^\dagger(p, f)$ is convex in f for any p. For this, we will be primarily using the fact that $1/x^2$ for $x \geq 0$, is a convex function. For any $f_1, f_2 \in \mathcal{F}$ and a fixed p,

$$L^\dagger(p, \lambda f_1 + (1-\lambda) f_2) = \frac{1}{24} \int_0^1 p(x) x^{-1} \frac{1}{(\lambda f_1'(x) + (1-\lambda) f_2'(x))^2} dx \tag{7.96}$$

$$\leq \frac{1}{24} \int_0^1 p(x) x^{-1} \left(\lambda \frac{1}{(f_1'(x))^2} + (1-\lambda) \frac{1}{(f_2'(x))^2} \right) dx \tag{7.97}$$

$$= \lambda L^\dagger(p, f_1) + (1-\lambda) L^\dagger(p, f_2) \tag{7.98}$$

For any $p_1, p_2 \in \mathcal{P}_K$ and a fixed f,

$$L^\dagger(\lambda p_1 + (1-\lambda) p_2) = \frac{1}{24} \int_0^1 (\lambda p_1(x) + (1-\lambda) p_2(x)) x^{-1} \frac{1}{f'(x)} dx \tag{7.99}$$

$$= \lambda L^\dagger(p_1, f) + (1-\lambda) L^\dagger(p_2, f) \tag{7.100}$$

□

In this section we use $g(x)$ and $h(x)$ as auxiliary functions; note that these have nothing to do with the asymptotic local loss function or dominating function from the previous section.

7.6.1 Optimizing for Best Compander

Theorem 17 follows from Theorem 16 by finding $f \in \mathcal{F}$ which minimizes $L^\dagger(p, f)$; however, the form of $L^\dagger(p, f)$ makes it more natural to optimize over f'. Since $f : [0, 1] \to [0, 1]$ is monotonic, we use constraints $f'(x) \geq 0$ and $\int_0^1 f'(x)\, dx = 1$ (equal since using the whole range is optimal).

We solve the following:

$$\text{minimize } L^\dagger(p, f) = \frac{1}{24} \int_0^1 p(x) f'(x)^{-2} x^{-1}\, dx$$

$$\text{subject to } \int_0^1 f'(x)\, dx = 1 \text{ and } f'(x) \geq 0 \text{ for all } x \in [0, 1]$$

The function $L^\dagger(p,f)$ is convex in f', and thus first order conditions show optimality. Let $g(x)$ be a function such that $\int_0^1 g(x)dx = 0$. We derive:

$$\frac{d}{dt}\frac{1}{24}\int_0^1 p(x)\big(f'(x)+t\,g(x)\big)^{-2}x^{-1}\,dx = \frac{1}{24}\int_0^1 p(x)x^{-1}\frac{d}{dt}\big(f'(x)+t\,g(x)\big)^{-2}\,dx \tag{7.101}$$

$$= -\frac{1}{12}\int_0^1 p(x)x^{-1}\big(f'(x)+t\,g(x)\big)^{-3}g(x)\,dx \tag{7.102}$$

$$= -\frac{1}{12}\int_0^1 p(x)x^{-1}f'(x)^{-3}g(x)\,dx \quad (\text{at } t=0) \tag{7.103}$$

If f is such that $f'(x) \propto (p(x)x^{-1})^{1/3}$, then for all $g(x)$ integrating to 0, the function (7.101) has value 0, making it optimal. This gives

$$f_p'(x) \propto (p(x)x^{-1})^{1/3} \tag{7.104}$$

and $f(0) = 0$ and $f(1) = 1$, from which (7.15) and (7.16) follow. If $f_p \in \mathcal{F}^\dagger$, then $f_p = \arg\min_f \widehat{L}(p,f)$ because for any other $f \in \mathcal{F}$,

$$\widehat{L}(p,f_p) = L^\dagger(p,f_p) \le L^\dagger(p,f) \le \liminf_{N\to\infty} N^2\widehat{L}(p,f,N) \tag{7.105}$$

If $f_p \notin \mathcal{F}^\dagger$, for any $\delta > 0$ define $f_{p,\delta} \in \mathcal{F}^\dagger$ where

$$f_{p,\delta}(x) = (1-\delta)f_p(x) + \delta x^{1/2} \tag{7.106}$$

$$\implies \widehat{L}(p,f_{p,\delta}) = L^\dagger(p,f_{p,\delta}) \le L^\dagger(p,f_p)(1-\delta)^{-2}\,. \tag{7.107}$$

Taking $\delta \to 0$ (and noting that $\mathcal{F}^\dagger \subseteq \mathcal{F}$) thus shows that

$$L^\dagger(p,f_p) = \inf_{f\in\mathcal{F}^\dagger} \widehat{L}(p,f)\,. \tag{7.108}$$

With (7.104), we can derive the best compander for symmetric Dirichlet priors and compare this to the EDI method. We show this in Appendix C.4

7.6.2 Most Difficult Density to Quantize

What density p maximizes (7.15) (i.e. is hardest to quantize with a compander)? Using calculus of variations to maximize

$$\int_0^1 (p(x)x^{-1})^{1/3}\,dx \tag{7.109}$$

(which of course maximizes (7.15)) subject to $p(x) \ge 0$ and $\int_0^1 p(x)\,dx = 1$, we find that maximizer is $p(x) = \frac{1}{2}x^{-1/2}$. However, while interesting, this is only for a single symbol; and because $\mathbb{E}[X] = 1/3$ under this distribution, it is clearly impossible to construct a prior over the simplex (whose output vector *must* sum to 1) with this marginal (unless $K = 3$). (See Appendix C.3 for commentary on $p(x) = \frac{1}{2}x^{-1/2}$.)

Hence, we add an expected value constraint while maximizing (7.109). This gives the problem:

$$\text{maximize} \int_0^1 \big(p(x)x^{-1}\big)^{1/3}\,dx \tag{7.110}$$

$$\text{subject to} \int_0^1 p(x)\,dx = 1; \tag{7.111}$$

$$\int_0^1 p(x)x\,dx = \frac{1}{K}; \tag{7.112}$$

$$\text{and } p(x) \ge 0 \text{ for all } x \tag{7.113}$$

We can solve this again using variational methods (we are maximizing a concave function so satisfying first order conditions are enough to ensure optimality). A function $p(x) > 0$ is optimal if, for any $g(x)$ where

$$\int_0^1 g(x)\,dx = 0 \text{ and } \int_0^1 g(x)x\,dx = 0 \tag{7.114}$$

the following holds:

$$\frac{d}{dt}\int_0^1 x^{-1/3}\big(p(x) + t\,g(x)\big)^{1/3}\,dx = 0\,. \tag{7.115}$$

We have by the same logic as before:

$$\frac{d}{dt}\int_0^1 x^{-1/3}\big(p(x) + t\,g(x)\big)^{1/3}\,dx = \frac{1}{3}\int_0^1 x^{-1/3}\big(p(x) + t\,g(x)\big)^{-2/3} g(x)\,dx \tag{7.116}$$

$$= \frac{1}{3}\int_0^1 x^{-1/3} p(x)^{-2/3} g(x)\,dx \quad (\text{at } t = 0) \tag{7.117}$$

Thus, if we can arrange things so that there are constants a_K, b_K such that

$$x^{-1/3}p(x)^{-2/3} = a_K + b_K x \tag{7.118}$$

this ensures (7.117) holds. In that case,

$$x^{-1/3}p(x)^{-2/3} = a_K + b_K x \tag{7.119}$$

$$\iff \quad p(x)^{-2/3} = a_K x^{1/3} + b_K x^{4/3} \tag{7.120}$$

$$\iff \quad p(x) = \big(a_K x^{1/3} + b_K x^{4/3}\big)^{-3/2} \tag{7.121}$$

This yields the *maximin density* p_K^* (7.18) from Theorem 18, where a_K, b_K are set to meet the constraints (7.111) and (7.112). We can determine a_K and b_K numerically given K, but getting a formula for a_K or b_K is difficult. We choose to parameterize $b_K/a_K = c_K K \log K$.

We can show that if $K > 24$, then $1/4 < c_K < 3/4$. We can also express a_K, b_K directly as:

$$a_K = 4^{1/3}(c_K K \log K + 1)^{-1/3} \tag{7.122}$$

$$b_K = 4a_K^{-2} - a_K\,. \tag{7.123}$$

See Appendix C.6 for this.

Now that we have the worst density p_K^*, we can use Theorem 17 to determine the best compander. We need to combine (7.121) with Theorem 17. The normalization constant (denominator of (7.16)) is

$$\int_0^{x'} x^{-1/3}\left(a_K x^{1/3} + b_K x^{4/3}\right)^{-1/2}\,dx = \frac{2\text{ArcSinh}\left(\sqrt{\frac{b_K x'}{a_K}}\right)}{\sqrt{b_K}} \tag{7.124}$$

which implies that

$$f_K^*(x) = \frac{\text{ArcSinh}\left(\sqrt{\frac{b_K x}{a_K}}\right)}{\text{ArcSinh}\left(\sqrt{\frac{b_K}{a_K}}\right)} = \frac{\text{ArcSinh}\left(\sqrt{c_K K \log K\, x}\right)}{\text{ArcSinh}\left(\sqrt{c_K K \log K}\right)}\,. \tag{7.125}$$

To compute the loss, we have

$$L^\dagger(p_K^*, f_K^*) = \frac{1}{24}\left(\frac{2\text{ArcSinh}\left(\sqrt{c_K K \log K}\right)}{\sqrt{b_K}}\right)^3 \tag{7.126}$$

The second term in (7.123) is negligible compared to the first. Ignoring it we get

$$L^\dagger(p_K^*, f_K^*) = \frac{1}{24}\frac{8(\log\left(\sqrt{c_K K\log K} + \sqrt{c_K K\log K + 1}\right)^3}{2c_K K\log K} \tag{7.127}$$

$$\leq \frac{1}{24}\frac{(\log 4(c_K K\log K+1))^3}{2c_K K\log K} \tag{7.128}$$

$$= \frac{1}{24}\left(\frac{\log^2 K}{c_K K} + \frac{O(\log K)}{2c_K K}\right) \tag{7.129}$$

$$= \Theta\left(\frac{1}{24c_K}\frac{\log^2 K}{K}\right). \tag{7.130}$$

This shows (7.21).

7.6.3 Saddle Point

The function $L^\dagger(p, f)$ is concave (actually linear) in p and convex in f', and we can show that the pair (f_K^*, p_K^*) forms a saddle point, thus proving (7.20) from Theorem 18.

We can compute that

$$(f_K^*)'(x) \propto (p_K^*(x)x^{-1})^{1/3} \tag{7.131}$$

$$= x^{-1/3}(a_K x^{1/3} + b_K x^{4/3})^{-1/2} \tag{7.132}$$

$$= (x^{2/3})^{-1/2}(a_K x^{1/3} + b_K x^{4/3})^{-1/2} \tag{7.133}$$

$$= \frac{1}{\sqrt{a_K x + b_K x^2}} \tag{7.134}$$

Assume we set a_K and b_K to the appropriate value for K. For any $p \in \mathcal{P}_K$,

$$L^\dagger(p, f_K^*) = \int_0^1 p(x)x^{-1}((f_K^*)'(x))^{-2}dx \tag{7.135}$$

$$= \int_0^1 p(x)x^{-1}(a_K x + b_K x^2)dx \tag{7.136}$$

$$= \int_0^1 p(x)(a_K + b_K x)dx \tag{7.137}$$

$$= a_K + b_K\frac{1}{K} \tag{7.138}$$

which implies that

$$\sup_{p\in\mathcal{P}_K} L^\dagger(p, f_K^*) = \sup_{p\in\mathcal{P}_K}\inf_{f\in\mathcal{F}} L^\dagger(p, f). \tag{7.139}$$

Since it is always true that

$$\sup_{p\in\mathcal{P}_{1/K}}\min_{f\in\mathcal{F}} L^\dagger(p, f) \leq \min_{f\in\mathcal{F}}\sup_{p\in\mathcal{P}_{1/K}} L^\dagger(p, f), \tag{7.140}$$

this completes showing that (f_*, p_K^*) is a saddle point.

Furthermore, $f_K^* \in \mathcal{F}^\dagger$ (specifically it behaves as a multiple of $x^{1/2}$ near 0), so $\widehat{L}(p, f_K^*) = L^\dagger(p, f_K^*)$ for all p, thus showing that f_K^* performs well against any $p \in \mathcal{P}_{1/K}$. Using (7.12) with the expressions for p_K^* and f_K^* gives (7.21).

7.6.4 Prior on Simplex

While p_K^* is the worst single-symbol distribution in $\mathcal{P}_{1/K}$, it is unclear whether a prior P^* on $\triangle_{K-1}$ exists with marginals p_K^*. However, it is possible to construct a prior P^* whose marginals are as hard to quantize, up to a constant, as p_K^*.

Lemma 23. *Let $p \in \mathcal{P}_{1/K}$. Then there exists a joint distribution of $(X_1, \ldots, X_K)$ such that (i) $X_i \sim p$ for all $i \in [K]$ and (ii) $\sum_{i\in[K]} X_i \leq 2$, guaranteed.*

Proof. Let F be the cumulative distribution function of p. We define the quantile function F^{-1} as

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}. \tag{7.141}$$

We break $[0, 1]$ into K uniform sub-intervals $I_i = ((i-1)/K, i/K]$ (let $I_1 = [0, 1/K]$). We then generate $X_1, X_2, \ldots, X_K$ jointly by the following procedure:

1. Choose a permutation $\sigma : [K] \to [K]$ uniformly at random (from $K!$ possibilities).
2. Let $U_k \sim \text{Unif}_{I_{\sigma(k)}}$ independently for all k.
3. Let $X_k = F^{-1}(U_k)$.

Now we consider $\sum_k X_k$. Let $b_i = F^{-1}(i/k)$ for $i = 0, 1, \ldots, K$. Note that if $\sigma(k) = i$ then $U_k \in ((i-1)/K, i/K]$ and hence $X_k = F^{-1}(U_k) \in [b_{i-1}, b_i]$. Therefore $X_{\sigma^{-1}(i)} \in [b_{i-1}, b_i]$ and thus for any permutation σ,

$$\sum_{i=1}^{K} b_{i-1} \leq \sum_{i=1}^{K} X_{\sigma^{-1}(i)} \leq \sum_{i=1}^{K} b_i \tag{7.142}$$

$$= \Big(\sum_{i=1}^{K} b_{i-1}\Big) + b_K - b_0 \tag{7.143}$$

$$\leq \Big(\sum_{i=1}^{K} b_{i-1}\Big) + 1 \leq 2 \tag{7.144}$$

□

We use Lemma 23 to determine a joint distribution on $K-1$ symbols with $X_i \sim p_K^*$. Then we scale all the symbol values by $1/2$, so their sum is guaranteed to be less than 1. Then the last symbol is $X_K = 1 - \sum_{i=1}^{K-1} X_i \geq 0$ so the sum of all K variables is exactly one. For this prior P^* we get that

$$\inf_{f\in\mathcal{F}} \widehat{\mathcal{L}}_K(P^*, f) \geq (K-1) \inf_{f\in\mathcal{F}} \widehat{L}(2p_K^*(2x), f) \tag{7.145}$$

$$= (K-1)\frac{1}{2} L^\dagger(p_K^*, f_K^*) \tag{7.146}$$

$$\geq \frac{1}{2}\frac{K-1}{K} \sup_{P\in\mathcal{P}_K^\triangle} \widehat{\mathcal{L}}_K(P, f_K^*) \tag{7.147}$$

where the last inequality holds because p_K^* is the worst-case single-symbol density (under expectation constraints). To make P^* symmetric, we can permute the symbol indices randomly without affecting the raw loss; thus we get Corollary 6.

To show how to get (7.146) from (7.145), we have

$$\inf_{f\in\mathcal{F}} \widehat{L}(2p_K^*(2x), f) = \frac{1}{24}\left(\int_0^1 (2p_K^*(2x)x^{-1})^{1/3} dx\right)^3 \tag{7.148}$$

$$= \frac{1}{24}\left(\int_0^1 (2p_K^*(u)2u^{-1})^{1/3} \frac{1}{2} du\right)^3 \tag{7.149}$$

$$= \frac{1}{2} L^\dagger(p_K^*, f_*) \tag{7.150}$$

In Figure 7.6, we validate the distribution P^* by showing the performance of each compander when quantizing random distributions drawn from P^*. For the minimax compander, the KL divergence loss on the worst-case prior looks to be within a constant of that for the other datasets.

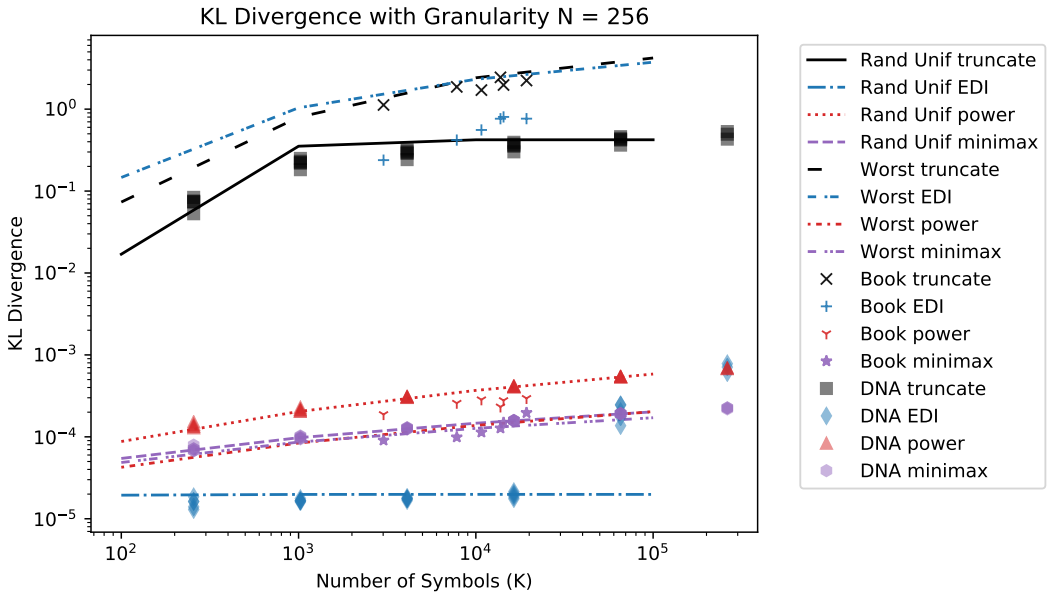


Figure 7.6: Each compander (or quantization method) is used on random distributions drawn for the prior P^*.

7.7 Worst-Case Analysis

In this section, we prove Theorem 19 which applies both to the minimax compander and the power compander.

Proof of Theorem 19. For notation, let $\delta_i = \widehat{y}_i - x_i$.

$$D_{\mathrm{KL}}\left(\boldsymbol{x}\|\boldsymbol{y}\right) = \sum_i x_i \log \frac{x_i}{y_i} \tag{7.151}$$

$$= \sum_i x_i \log \frac{x_i}{\widehat{y}_i} + \log\left(\sum_i \widehat{y}_i\right) \tag{7.152}$$

$$= \sum_i (\widehat{y}_i - \delta_i) \log \frac{\widehat{y}_i - \delta_i}{\widehat{y}_i} + \log\left(1 + \sum_i \delta_i\right) \tag{7.153}$$

Next we use that $\log(1+x) \le x$.

$$D_{\mathrm{KL}}\left(\boldsymbol{x}\|\boldsymbol{y}\right) \le \sum_i (\widehat{y}_i - \delta_i) \frac{-\delta_i}{\widehat{y}_i} + \sum_i \delta_i \tag{7.154}$$

$$= \sum_i -\delta_i + \sum_i \frac{\delta_i^2}{\widehat{y}_i} + \sum_i \delta_i \tag{7.155}$$

$$= \sum_i \frac{(\widehat{y}_i - x_i)^2}{\widehat{y}_i} \tag{7.156}$$

where in (7.154) we used the inequality $\log(1+x) \le x$ on both appearances of the logarithm.

Since there is no prior in the worst-case problem, we use midpoint decoding. Let $\bar{y}_i$ be the midpoint value of the interval x_i is mapped to. For interval j, let $\acute{y}_j$ be the value associated with the mean value of the interval. For a continuous and smooth companding function $f(x)$, and interval j with boundaries b_j and

b_{j+1}, let x be such that

$$f'(x) = \frac{f(b_{j+1}) - f(b_j)}{b_{j+1} - b_j} = \frac{1}{N} \frac{1}{b_{j+1} - b_j} \tag{7.157}$$

Let $\acute{y}_j$ be this x.

Thus for any x in an interval associated with mean value given by $\acute{y}$, the length of the interval of x falls in can be given by

$$\ell(x) \triangleq b_{j+1} - b_j = \frac{1}{Nf'(\acute{y})} \,. \tag{7.158}$$

For any $\acute{y}$ in an interval with midpoint $\bar{y}$

$$\frac{\acute{y}}{2} \leq \bar{y} \,. \tag{7.159}$$

Then

$$D_{\text{KL}}\left(\boldsymbol{x} \| \boldsymbol{y}\right) \leq \sum_i \frac{(\bar{y}_i - x_i)^2}{\bar{y}_i} \tag{7.160}$$

$$\leq \sum_i \frac{\ell(x)^2}{\bar{y}_i} \tag{7.161}$$

$$= \sum_i \frac{1}{N^2(\acute{y}_i/2)(f'(\acute{y}_i))^2} \tag{7.162}$$

$$= \frac{2}{N^2} \sum_i \frac{1}{\acute{y}_i (f'(\acute{y}_i))^2} \tag{7.163}$$

Going further, we simply let $y_i = \acute{y}_i$.

First, we analyze the worst-case performance of the power compander (which is easier to analyze). The first part of the analysis will be similar to the average analysis.

$$f(x) = x^r \tag{7.164}$$

$$f'(x) = rx^{r-1} \tag{7.165}$$

which leads to

$$D_{\text{KL}}\left(\boldsymbol{x} \| \boldsymbol{y}\right) \leq \frac{2}{N^2 r^2} \sum_i \frac{1}{y_i y_i^{2r-2}} \tag{7.166}$$

$$\leq \frac{2}{N^2 r^2} \sum_i y_i^{1-2r} \tag{7.167}$$

For power compander, we pick $r = \frac{1}{\log K}$. So long as $r < 1/2$, the function y_i^{1-2r} is concave in y_i. The maximum value is achieved when each of y_i is the same for each i. We do not know what $\sum_i y_i$ is, but we can divide this value into K equal parts for an upper bound.

$$D_{\text{KL}}\left(\boldsymbol{x} \| \boldsymbol{y}\right) \leq \frac{2}{N^2 r^2} K \left(\frac{\sum_i y_i}{K} \right)^{1-2r} \tag{7.168}$$

$$\leq \frac{2}{N^2 r^2} K^{2r} \left(\sum_i y_i \right)^{1-2r} \tag{7.169}$$

$$\leq \frac{2}{N^2 r^2} e^2 \max\left\{ 1, \sum_i y_i \right\} \tag{7.170}$$

Next, we need to bound $\max\{1, \sum_i y_i\}$. Assume $\sum_i y_i > 1$.

$$\sum_i |y_i - x_i| \leq \sum_i \ell(y_i) \tag{7.171}$$

$$\sum_i y_i \leq \sum_i x_i + \frac{1}{Nr} \sum_i y_i^{1-r} \tag{7.172}$$

$$\sum_i y_i \leq \sum_i x_i + \frac{1}{Nr} K \left(\frac{\sum_i y_i}{K} \right)^{1-r} \tag{7.173}$$

$$\sum_i y_i \leq \sum_i x_i + \frac{e}{Nr} \left(\sum_i y_i \right)^{1-r} \tag{7.174}$$

$$\sum_i y_i \leq \sum_i x_i + \frac{e}{Nr} \left(\sum_i y_i \right) \tag{7.175}$$

We can combine terms with $\sum_i y_i$.

$$\left(1 - \frac{e}{Nr}\right) \sum_i y_i \leq 1 \tag{7.176}$$

$$\implies \sum_i y_i \leq \frac{1}{1 - \frac{e}{Nr}} \tag{7.177}$$

$$\implies \sum_i y_i \leq \frac{N}{N - e \log K} = 1 + \frac{\log K}{N - e \log K} \tag{7.178}$$

For the bound to make sense $N > e \log K$. For the bound to be a constant, we need roughly that $N > 2e \log K$. Assuming this and combining (7.170) with (7.178), we have

$$\begin{aligned} D_{\text{KL}}(\boldsymbol{x} \| \boldsymbol{y}) &\leq 2e^2 \frac{(\log K)^2}{N^2} \max\left\{1, \left(1 + \frac{e \log K}{N - e \log K}\right)\right\} \\ &\leq 2e^2 \frac{(\log K)^2}{N^2} \left(1 + \frac{e \log K}{N - e \log K}\right) \end{aligned} \tag{7.179}$$

Now we will do the analysis for the minimax compander. To start, we will use $C = cK \log K$

$$f(x) = \frac{\text{ArcSinh}(\sqrt{Cx})}{\text{ArcSinh}(\sqrt{C})} \tag{7.180}$$

$$f'(x) = \frac{1}{2\text{ArcSinh}(\sqrt{C})} \frac{\sqrt{C}}{\sqrt{x}\sqrt{1 - Cx}} \tag{7.181}$$

$$\frac{1}{f'(x)} = 2\text{ArcSinh}(\sqrt{C}) \sqrt{\frac{x}{C} + x^2} \tag{7.182}$$

$$D_{\text{KL}}(\boldsymbol{x} \| \boldsymbol{y}) \leq \frac{(2\text{ArcSinh}(\sqrt{C}))^2}{N^2} \sum_i \frac{\frac{y_i}{C} + y_i^2}{y_i} \tag{7.183}$$

$$\leq \frac{(2\text{ArcSinh}(\sqrt{C}))^2}{N^2} \left(\frac{K}{C} + \sum_i y_i \right) \tag{7.184}$$

Assume that $\sum_i y_i > 1$.

$$\sum_i |y_i - x_i| \leq \sum \ell(y_i) \tag{7.185}$$

$$\sum_i y_i \leq \sum_i x_i + \frac{2\text{ArcSinh}(\sqrt{C})}{N} \sum_i \sqrt{\frac{y_i}{C} + y_i^2} \tag{7.186}$$

We will first bound the second sum in the equation above

$$\sum_i \sqrt{\frac{y_i}{C} + y_i^2} \leq \sum_i \sqrt{\frac{y_i}{C}} + \sqrt{y_i^2} \tag{7.187}$$

$$\leq \left(K \left(\frac{\sum_i y_i}{K(cK \log K)} \right)^{1/2} + \sum_i y_i \right) \tag{7.188}$$

$$\leq \left(\left(\frac{\sum_i y_i}{c \log K} \right)^{1/2} + \sum_i y_i \right) \tag{7.189}$$

$$\leq \left(\left(\frac{\sum_i y_i}{(c \log K)^{1/2}} \right) + \sum_i y_i \right) \tag{7.190}$$

$$\leq \left(\sum_i y_i \right) \left(1 + \frac{1}{(c \log K)^{1/2}} \right) \tag{7.191}$$

$$\leq 2 \left(\sum_i y_i \right) \tag{7.192}$$

Then (7.186) becomes

$$\sum_i y_i \leq 1 + \frac{4\text{ArcSinh}(\sqrt{C})}{N} \left(\sum_i y_i \right) \tag{7.193}$$

Since we have $\sum_i y_i$ on both sides of the equation, we can combine these terms like before.

$$\left(1 - \frac{4\text{ArcSinh}(\sqrt{C})}{N} \right) \sum_i y_i \leq 1 \tag{7.194}$$

$$\implies \sum_i y_i \leq \frac{N}{N - 4\text{ArcSinh}(\sqrt{C})} \tag{7.195}$$

Combining things together and using $\text{ArcSinh}(\sqrt{x}) = \log(\sqrt{x+1} + \sqrt{x}) \leq \log(2\sqrt{x} + 1)$ we get

$$D_{\text{KL}}(\boldsymbol{x} \| \boldsymbol{y}) \leq 2 \frac{(2\text{ArcSinh}(\sqrt{C}))^2}{N^2} \left(\frac{K}{C} + \frac{N}{N - 4\text{ArcSinh}(\sqrt{C})} \right)$$

$$= 2 \frac{(2\text{ArcSinh}(\sqrt{cK \log K}))^2}{N^2} \left(\frac{K}{cK \log K} + \frac{N}{N - 4\text{ArcSinh}(\sqrt{cK \log K})} \right) \tag{7.196}$$

$$= 2 \frac{(2 \log(2\sqrt{cK \log K} + 1))^2}{N^2} \left(\frac{1}{c \log K} + \frac{N}{N - 4 \log(2\sqrt{cK \log K} + 1)} \right) \tag{7.197}$$

We will assume that $N > 8 \log(2\sqrt{cK \log K} + 1)$ so the second term in the parenthesis is a constant close to 1. We can then bound the entire term in the parenthesis by 2. Then,

$$D_{\text{KL}}(\boldsymbol{x} \| \boldsymbol{y}) \leq \frac{4(\log cK \log K + 2 \log 3)^2}{N^2} \tag{7.198}$$

$$= \Theta \left(\frac{(\log K)^2}{N^2} \right) \tag{7.199}$$

□

We note that both the minimax compander and the power compander are achievable solutions for worst-case divergence covering. Theoretically, these companders show that as $\varepsilon \to 0$, we get that

$$\log M(K, \varepsilon) \leq \left(\frac{(\log K)^2}{\varepsilon} \right)^{K/2}. \tag{7.200}$$

However, the advantage is that both the power compander and minimax compander are explicit quantization methods which are simple to use. In conclusion, the minimax compander (better than the power compander) is both a practical and effective solution.

Appendix A

Permutation Channel: Other Tools

A.1 Berry-Esseen

Here is an alternative method to Petrov's calculation of bounding sums of independent random variables.

Theorem 20. *(Berry-Esseen[89])*
For $X_1, X_2, ..., X_n$ which are independent random variables with $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = \sigma_i^2 < \infty$ and $\mathbb{E}|X_i|^3 = \gamma_i < \infty$, let

$$Y = \frac{X_1 + X_2 + ... + X_n}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \tag{A.1}$$

and let $F_n(x)$ be the cdf of W. Then

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq c \frac{\sum_{i=1}^n \gamma_i}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} \tag{A.2}$$

where $\Phi(x)$ is a standard normal and c is some constant.

Suppose we have zero-mean random variables which might not be identical, but where $|X_i| < 1$. Since X_i takes values in $[-1, 1]$, $\mathbb{E}|X_i|^3 \leq \mathbb{E}|X_i|^2$, and therefore $\sum_{i=1}^n \gamma_i \leq \sum_{i=1}^n \sigma_i^2$. Thus,

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq c \frac{\sum_{i=1}^n \sigma_i^2}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} \leq \frac{c}{\sqrt{\sum_{i=1}^n \sigma_i^2}}\,. \tag{A.3}$$

Let $W = X_1' + ... + X_n'$ where X_i' is an independent Bernoulli with some parameter. Each takes values in 0 or 1 and the expected value is not 0. Let $F_n'(\cdot)$ be the cdf of $S = W/\sqrt{\sum_{i=1}^n \sigma_i^2}$ and let $G(x)$ be the cdf of a Gaussian with variance 1 shifted to match the mean of S.

In this shifted version, if we want to find the probability that $W = w$ for some integer w. We can find the pdf $P_W(w)$ by integrating over $G(x)$ with an interval of length $1/\sqrt{\sum_{i=1}^n \sigma_i^2}$ corresponding to where w occurs. The result of the integral will at most be $1/\sqrt{\sum_{i=1}^n \sigma_i^2}$ times $1/\sqrt{2\pi}$ (since the maxmium of the pdf of a Gaussian with unit variance is $1/\sqrt{2\pi}$), so

$$P_W(w) \leq \frac{1/\sqrt{2\pi}}{\sqrt{\sum_{i=1}^n \sigma_i^2}} + \frac{c}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \leq \frac{c'}{\sqrt{\sum_{i=1}^n \sigma_i^2}} = \frac{c'}{\sqrt{\mathrm{Var}(W)}} \tag{A.4}$$

for any constant w.

We will restate this in the following corollary.

Corollary 7. *For independent (possibly non-identical) Bernoulli random variables X_i, if $W = \sum_i^n X_i$ then for some integer w,*

$$\mathbb{P}[W = w] \leq \frac{c}{\sqrt{Var(W)}} \tag{A.5}$$

where c is a constant that does not depend on n.

Appendix B

More on Average-Case Divergence

B.1 Gamma Distribution

There are two ways to parameterize the Gamma distribution.

Definition 28 (Gamma Distribution)**.**
The α, β ***parameterization (shape-rate):*** *Random variable* $X \sim Gamma(\alpha, \beta)$ *has pdf*

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{B.1}$$

for $x > 0$ *and* $\alpha, \beta > 0$.
The k, θ ***parameterization (shape-scale):*** *Random variable* $X \sim Gamma(k, \theta)$ *has pdf*

$$f(x) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \tag{B.2}$$

for $x > 0$ *and* $k, \theta > 0$.

We can switch between the parameterization using $k = \alpha$ and $\beta = \frac{1}{\theta}$. The mean of a Gamma random variable is $k\theta$ (or $\frac{\alpha}{\beta}$). The variance is $k\theta^2$ (or $\frac{\alpha}{\beta^2}$).

Some examples:

- If k (or α) is a positive integer, the Gamma Distribution gives the Erlang Distribution.
- If $k = 1$, we get an exponential distribution with parameter $\frac{1}{\theta}$ or α.
- If $X \sim \text{Gamma}(v/2, 2)$, then X is a chi-squared distribution with v degrees of freedom.

Fact 11 (Scaling)**.** *If* $X \sim Gamma(k, \theta)$ *then for any* $c > 0$, $cX \sim Gamma(k, c\theta)$.

Fact 12 (Summation)**.** *If* $X_i \sim Gamma(k_i, \theta)$ *and are independent with the same* θ*, where* $i = 1, ..., N$*, then* $\sum_{i=1}^N X_i \sim Gamma\left(\sum_{i=1}^N k_i, \theta\right)$.

For integer k_i, this makes sense since the sum of k exponentials random variables with the same θ are Erlang distributions with k. It follows sums of Erlang random variables are Erlang.

The Dirichlet distribution is the normalized sum of independent Gamma distribution.

Fact 13 (Normalized Sum)**.** *Let $X_i \sim$ Gamma(k_i, θ), where $i = 1, ..., N$, be N independent Gamma distributions. Let*

$$V = \sum_{i=1}^{N} X_i \,. \tag{B.3}$$

Then

$$Y = (Y_1, \ldots, Y_N) = \left(\frac{X_1}{V}, \ldots, \frac{X_N}{V} \right) \tag{B.4}$$

is distributed as Dirichlet$(k_1, ..., k_m)$. The random variable Y is independent of X.

If we want to get points sampled according to Jeffreys' Prior on the simplex, we can do so by fixing any θ, and sampling N independent values from $X \sim \text{Gamma}\left(\frac{1}{2}, \theta\right)$. If we normalize the values in the vector so they become 1, this vector is distributed Dirichlet.

In particular if we have $X \sim \text{Gamma}(\alpha, \theta)$ and $Y \sim \text{Gamma}(\beta, \theta)$ and X and Y are independent, then $\frac{X}{X+Y}$ is a beta distribution with parameters α and β.

To find the entropy of a the Gamma distribution, we need some more definitions.

Definition 29. *The digamma function is*

$$\psi(x) = \frac{d}{dx} \ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)} \,. \tag{B.5}$$

The trigamma function is

$$\psi_1(x) = \frac{d^2}{dx^2} \ln \Gamma(x) \,. \tag{B.6}$$

Definition 30. *Let γ is the Euler-Mascheroni constant, that is*

$$\gamma = \lim_{n \to \infty} \left(-\log n + \sum_{k=1}^{n} \frac{1}{k} \right) = 0.57721... \tag{B.7}$$

The Euler-Mascheroni constant is the limiting difference between the Harmonic Series and natural logarithm. It appears frequently in integrals involving the exponential function. The log used is natural log.

Some properties of the digamma function are

- $\psi(z+1) = \psi(z) + \frac{1}{z}$
- $\psi(n) = \sum_{k=1}^{n-1} \frac{1}{k} - \gamma$
- $\psi(1) = -\gamma$
- $\psi(1/2) = -2\log 2 - \gamma$
- $\psi(n+1/2) = -\gamma - 2\log 2 + \sum_{k=1}^{n} \frac{2}{2k-1}$

Fact 14 (Logarithmic Expectations)**.** *If $X \sim$ Gamma(k, θ) then*

- $\mathbb{E}[\log(X)] = \psi(k) + \log(\theta)$
- *Var*$[\log(X)] = \psi_1(k)$

Fact 15 (Entropy). *If $X \sim Gamma(k, \theta)$ then $H(X) = k + \log(\theta) + \log(\Gamma(k)) + (1-k)\psi(k)$.*

B.2 Proofs

B.2.1 D_{max} for Exponential

Proposition 28. *If $p_X(x) = e^{-x}$, $R(p_x, D)$ intersects 0 at $D = 1 - \gamma$, where γ is the Eulor-Mascheroni constant.*

Proof. When there is only 1 reconstruction one, that point occurs at $y = 1$. We can integrate to determine what distortion level occurs. The integral to find is

$$\int_0^\infty e^{-x} x \log x dx = 1 - \gamma \,. \tag{B.8}$$

It is known that $\int_0^\infty e^{-x} \log x \, dx = -\gamma$. We can find the result using integration by parts on $\int_0^\infty e^{-x} \log x dx$. □

B.2.2 Proof of Proposition 16

This is a variation on a proof in [63] with changes for the divergence distortion.

Proof. If we first ignore the constraint $\mathbb{E}[X] = \mathbb{E}[Y]$, then finding the minimum in (6.5) is equivalent to minimizing

$$F(q) = -\frac{1}{\beta} \int p_X(x) \log \left(\int q_Y(y) e^{\beta d(x,y)} \, dy \right) dx \tag{B.9}$$

over the distribution $q_Y(y)$.

(Optimizing (B.9) for a chosen β gives an equivalent set of solutions as optimizing (6.5) without $\mathbb{E}[X] = \mathbb{E}[Y]$ for a corresponding D). Instead of searching for an optimal $q_Y(y)$, define $\hat{y} : [0, 1] \to Y$. We can replace y in equation (B.9) with $\hat{y}(u)$ and take the integral over u instead of y. Using calculus of variations by perturbing $\hat{y}(u)$ with $\epsilon\eta(u)$, a conditional for optimality is that

$$\int \eta(u) \left(\int p_X(x) \frac{e^{-\beta d(x,\hat{y}(u))}}{\int e^{-\beta d(x,\hat{y}(u'))} du'} \frac{\partial}{\partial y} d(x, \hat{y}(u)) \, dx \right) du = 0 \,. \tag{B.10}$$

(We will skip the derivation of this part as it is presented in [63].) Define $Z(x) = \int e^{-\beta d(x,\hat{y}(u))} \, du$. We will add the constraint $\mathbb{E}[X] = \mathbb{E}[Y]$ back into the problem by only allowing $\eta(u)$ such that $\int \eta(u) \, du = 0$. Hence, as a function in u, we must have that the expression

$$\int \frac{p_X(x)}{Z(x)} e^{-\beta d(x,\hat{y}(u))} \frac{\partial}{\partial y} d(x, \hat{y}(u)) \, dx \tag{B.11}$$

is constant. Thus there is a constant c so that if value y_0 is in the support of the optimal function $\hat{y}$, we must have

$$\int \frac{p_X(x)}{Z(x)} e^{-\beta d(x,y_0)} \frac{\partial}{\partial y} d(x, y_0) \, dx = c \,. \tag{B.12}$$

Suppose that the optimal $q(\hat{y}(u))$ has a continuous interval I in its support. Then its derivatives in terms of y_0 on the left-hand side of (B.12) must be zero when evaluated at any point $y_0 \in I$. Noting that $\frac{\partial}{\partial y} d(x, y_0) = -x/y_0$, we can evaluate the conditions

$$0 = \int \frac{p_X(x)}{Z(x)} \frac{\partial}{\partial y_0} e^{-\beta d(x,y_0)} \frac{\partial}{\partial y} d(x, y_0) \, dx \tag{B.13}$$

$$= \int \frac{p_X(x)}{Z(x)} \left[(-\beta) \left(-\frac{x}{y_0} \right) e^{-\beta d(x,y_0)} \left(-\frac{x}{y_0} \right) - e^{-\beta d(x,y_0)} \left(\frac{-x}{y_0^2} \right) \right] dx \tag{B.14}$$

$$= \int \frac{p_X(x)}{Z(x)} e^{-\beta d(x,y_0)} \frac{x}{y_0^2} (-\beta x + 1) \, dx \,. \tag{B.15}$$

We can continue to compute higher order derivatives.

$$0 = \int \frac{p_X(x)}{Z(x)} \frac{\partial^2}{\partial y_0^2} e^{-\beta d(x,y_0)} \frac{\partial}{\partial y} d(x,y_0)\, dx \tag{B.16}$$

$$= \int \frac{p_X(x)}{Z(x)} \frac{\partial}{\partial y_0} e^{-\beta d(x,y_0)} \frac{x}{y_0^2}(-\beta x + 1)\, dx \tag{B.17}$$

$$= \int \frac{p_X(x)}{Z(x)} e^{-\beta d(x,y_0)} \frac{x}{y_0^3}(-\beta x + 1)\left(\beta x - \frac{1}{2}\right)\, dx \tag{B.18}$$

Formally, we can use induction to get the nth order derivative.

$$0 = \int \frac{p_X(x)}{Z(x)} \frac{\partial^n}{\partial y_0^n} e^{-\beta d(x,y_0)} \frac{\partial}{\partial y} d(x,y_0)\, dx \tag{B.19}$$

$$= \int \frac{p_X(x)}{Z(x)} e^{-\beta d(x,y_0)} \frac{x}{y_0^{n+1}}(-1) \prod_{k=1}^{n}\left(\beta x - \frac{1}{k}\right)\, dx\,. \tag{B.20}$$

We now complete the proof by showing it is not possible for find a continuous interval where these constraints are met. Let $f(x,y_0) = \frac{p_X(x)}{Z(x)} e^{-\beta d(x,y_0)} x(\beta x - 1)$ and $g_n(x) = \prod_{k=1}^{n-1}\left(\beta x - \frac{1}{k+1}\right)$ where $g_1(x) = 1$. For y_0 to be in a continuous interval, for all n we must have

$$\int dx\, f(x,y_0) g_n(x) = 0 \tag{B.21}$$

Linear combinations of g_n (which are polynomials of every possible order) are sufficient to span to space of all functions so we can write $f(x,y_0) = \sum_{n=1}^{\infty} c_n g_n(x)$ for some choices of c_i. We get $\int dx\, f(x,y_0)^2 = \int dx\, f(x,y_0) \sum_{n=1}^{\infty} c_n g_n(x) = 0$, and thus $f(x,y_0)$ must be the zero function. This is not possible for any distribution $p_X(x)$. □

B.2.3 Proof of Proposition 18

Proof. Suppose $Q_{Y|X}(y|x)$ is a conditional probability density meeting constraints for a specific D (i.e. $Q_{Y|X}(y|x) \in \mathcal{B}$). Below, are the calculations where the distribution of X, Y is according to $P_X Q_{Y|X}$. For any $\alpha_\lambda \in A_\lambda$,

$$I(X;Y) + \lambda \mathbb{E}[d(X,Y)] - \int_x \int_y Q_{Y|X}(y|x) p_X(x) \log \alpha_\lambda(x,y) dy dx \tag{B.22}$$

$$= \int_x \int_y p_X(x) Q_{Y|X}(y|x) \log \frac{Q_{Y|X}(y|x)}{\alpha_\lambda(x,y) Q(y) e^{-\lambda d(x,y)}} dx dy \tag{B.23}$$

$$\geq \int_x \int_y p_X(x) Q_{Y|X}(y|x) \left[1 - \frac{\alpha_\lambda(x,y) Q(y) e^{-\lambda d(x,y)}}{Q_{Y|X}(y|x)}\right] dx dy \tag{B.24}$$

$$= 1 - \int_y Q_Y(y) \int_x p_X(x) \alpha_\lambda(x,y) e^{-\lambda d(x,y)} dx dy \tag{B.25}$$

$$= 1 - \int_y Q_Y(y) c(y) dy \tag{B.26}$$

$$\geq 0 \tag{B.27}$$

where we used that $\log(1/x) \geq 1 - x$. This gives for any $Q_{Y|X} \in \mathcal{B}$

$$I(X;Y) \geq \sup_{\lambda \geq 0, \alpha_\lambda \in A_\lambda} -\lambda D + \int_x \int_y Q_{Y|X}(y|x) p(x) \log \alpha_\lambda(x,y)\, dx\, dy\,. \tag{B.28}$$

Taking the infimum over all $Q_{Y|X} \in \mathcal{B}$ gives the result. □

B.2.4 Proof of Proposition 19

Proof. Since Y is discrete on n values, we have

$$I(X;Y) = H(Y) - H(Y|X) \leq \log(n) - H(Y|X) \leq \log(n) \tag{B.29}$$

Since f is strictly decreasing and only has nonnegative inputs, it has a nonnegative and strictly decreasing inverse f^{-1} and Theorem 11 shows that f can decay at most quadratically in n. This is because, per Theorem 11, for some constant c_X (dependent on X but not on n or D), we have

$$c_X + \frac{1}{2}\log\left(\frac{1}{D}\right) \leq R(p_X, D) \leq I(X;Y) \leq \log(n) \tag{B.30}$$

Letting $D = D(p_X, n)$ (the distortion achievable with n centers), we get that $c_X + \frac{1}{2}\log\left(\frac{1}{D(p_X,n)}\right) \leq \log(n)$ which implies $e^{c_X} D(p_X, n)^{-1/2} \leq n$ and hence $D(p_X, n) \geq e^{2c_X} n^{-2}$. Since f is such that $D(p_X, n) \leq c\, f(n)$, for constant c, we get a contradiction if $f(n) = o(n^{-2})$, and so f cannot decay faster than quadratically.

Let $D \stackrel{\triangle}{=} D(p_X, n)$. Then for sufficiently large n,

$$\begin{aligned} I(X;Y) &\leq \log(n) = \log(f^{-1}(f(n))) && \text{(B.31)}\\ &\leq \log(f^{-1}(D/c)) && \text{(B.32)}\\ &\leq \log\left(c^{1/b}(1+o(1))f^{-1}(D)\right) && \text{(B.33)}\\ &= \log(f^{-1}(D)) + \frac{1}{b}\log(c) + o(1) && \text{(B.34)} \end{aligned}$$

If $f(n) = n^{-b}$ then $f^{-1}(n) = n^{-1/b}$ and $\log(f^{-1}(D/c)) = \frac{1}{b}\log\left(\frac{1}{D}\right) + \frac{1}{b}\log(c)$ exactly. □

B.2.5 Proof of Lemma 16

Proof. Part i) (vertical scaling): $y^{(cp,I)} = y^{(p,I)}$ because y can be thought of as the expectation of a normalized p over I, and so scaling p does not change it. So:

$$\begin{aligned} D^{(cp,I)} &= \int_I cp(x)x\log\left(\frac{x}{y^{(p,I)}}\right)dx && \text{(B.35)}\\ &= c\int_I p(x)x\log\left(\frac{x}{y^{(p,I)}}\right)dx = cD^{(p,I)} && \text{(B.36)} \end{aligned}$$

Part ii) (horizontal scaling): First we note that $y^{(p_{\times c}, I_{\times c})} = cy^{(p,I)}$ (it scales with the interval). This is trivial from the 'conditional expectation' interpretation of the definition ('conditional expectation' in quotes because we can evaluate it even if p integrates to more than 1 over the interval in question). Then we can use a change of variables $x/c \to x$ to put everything back into I:

$$\begin{aligned} D^{(p_{\times c}, I_{\times c})} &= \int_{I_{\times c}} p_{\times c}(x)x\log\left(\frac{x}{y^{(p_{\times c}, I_{\times c})}}\right)dx && \text{(B.37)}\\ &= \int_{I_{\times c}} p(x/c)x\log\left(\frac{x}{cy^{(p,I)}}\right)dx && \text{(B.38)}\\ &= \int_I p(x)cx\log\left(\frac{cx}{cy^{(p,I)}}\right)c\,dx && \text{(B.39)}\\ &= c^2\int_I p(x)x\log\left(\frac{x}{y^{(p,I)}}\right)dx && \text{(B.40)}\\ &= c^2 D^{(p,I)} && \text{(B.41)} \end{aligned}$$

Part iii) (translation): First we will scale p so that p is a probability distribution on I (we can always undo the scale at the end). We note that $y^{(p_{+c}, I_{+c})} = y^{(p,I)} + c$ and use a change of variables $x - c \to x$ to

put everything back into i, giving us

$$D^{(p_{+c},I_{+c})} = \int_{I_{+c}} p_{+c}(x)x \log\left(\frac{x}{y^{(p_{+c},I_{+c})}}\right) dx \tag{B.42}$$

$$= \int_I p(x)(x+c) \log\left(\frac{x+c}{y^{(p,I)}+c}\right) dx \tag{B.43}$$

For ease of writing we will let $y := y^{(p,I)}$ for the remainder of this proof.

We will use a variational method. Let

$$f(c) = yD^{(p,I)} - (y+c)D^{(p_{+c},I_{+c})} \tag{B.44}$$

Note that $f(0) = 0$. If $f(c)$ is nonnegative, this proves our result; since $f(0) = 0$, if $\frac{d}{dc}f(c) \geq 0$ for all c, that proves $f(c)$ is nonnegative for all $c > 0$. We calculate the derivative of $f(c)$. First we will compute the derivative of $D^{(p_{+c},I_{+c})}$.

$$\frac{d}{dc}\log\left(\frac{x+c}{y+c}\right) = \frac{y+c}{x+c}\left(-\frac{x-y}{(y+c)^2}\right) \tag{B.45}$$

$$= \frac{-1}{x+c}\frac{x-y}{y+c}. \tag{B.46}$$

$$\frac{d}{dc}D^{(p_{+c},I_{+c})} = \frac{d}{dc}\int_I p(x)(x+c)\log\left(\frac{x+c}{y+c}\right) dx \tag{B.47}$$

$$= \int_I p(x)\log\left(\frac{x+c}{y+c}\right) dx - \int_I p(x)\frac{x-y}{y+c}\,dx \tag{B.48}$$

$$= \int_I p(x)\log\left(\frac{x+c}{y+c}\right) dx \tag{B.49}$$

since $y = \int_I p(x)x\,dx$. Note that $\mathbb{E}[\log(x+c)] - \mathbb{E}[\log(y+c)] \leq 0$ due to Jensen's inequality. We can conclude that shifting by c always decreases the distortion.

$$\frac{d}{dc}f(c) = -(y+c)\int_I p(x)\log\left(\frac{x+c}{y+c}\right) dx - D^{(p_{+c},I_{+c})} \tag{B.50}$$

$$\frac{\frac{d}{dc}f(c)}{y+c} = -\int_I p(x)\left(1+\frac{x+c}{y+c}\right)\log\left(\frac{x+c}{y+c}\right) \tag{B.51}$$

$$= -\int_I p\left((y+c)z - c\right)(1+z)\log z\,dz \tag{B.52}$$

$$= -\mathbb{E}_Z[(Z+1)\log Z] \tag{B.53}$$

where $z = \frac{x+c}{y+c}$, and Z is the random variable produced by generating X according to p and letting $Z = \frac{X+c}{y+c}$ (since we're only dealing with one interval, y is fixed). Since p (governing X) is uniform, Z is also uniform (since Z is just a shifted and scaled X); since $\mathbb{E}[X] = y$ by definition, $\mathbb{E}[Z] = 1$. Thus, Z is uniform and centered at 1, supported on $[c/(y+c), (2y+c)/(y+c)]$. Therefore, we need to show that $-\mathbb{E}_Z[(Z+1)\log Z] \geq 0$, or equivalently that $\mathbb{E}_Z[(Z+1)\log Z] \leq 0$.

Letting $\theta = \frac{c}{y+c} \in [0,1)$, we note that Z is uniform on $[\theta, 2-\theta]$ (centered at 1). Therefore, we aim to show that for all θ,

$$\mathbb{E}_{Z\sim\mathrm{Unif}_{[\theta,2-\theta]}}[(Z+1)\log Z] \leq 0 \tag{B.54}$$

We do this by writing out the Taylor expansion of $(z+1)\log(z)$ at $z = 1$:

$$(z+1)\log(z) = (2z-1) + \sum_{n\geq 2}\frac{(-1)^{n+1}(z-1)^n(n-2)}{(n-1)n} \tag{B.55}$$

This Taylor expansion converges for all $|z-1| < 1$, which is exactly what we need. By linearity of expectations, to get the expected value of $(Z+1)\log(Z)$, we can take the expectation of each term of the sum separately and add those up.

Taking the expectation over an interval centered at 1, we note that the terms for odd n (including the $2z-1$ term) all wind up with expectation 0 by symmetry (since $Z-1$ will be symmetric about 0, so $\mathbb{E}[(Z-1)^n] = 0$ for all odd n). For even n, we get that $(-1)^{n+1} = -1$, and all the other terms are nonnegative (including $\mathbb{E}[(Z-1)^n]$).

Thus, we can conclude that

$$\mathbb{E}[(Z+1)\log(Z)] = \sum_{\text{even } n \geq 2} \frac{-\mathbb{E}[(Z-1)^n](n-2)}{(n-1)n} < 0 \tag{B.56}$$

whenever the expectation is taken over a uniform distribution on $[\theta, 2-\theta]$. Thus, we have shown that $\frac{d}{dc}f(c) = -\mathbb{E}[(Z+1)\log(Z)] > 0$ and thus (since $f(0) = 0$) that $f(c) \geq 0$ for all c, thus showing our result. □

B.2.6 Full Proof of Lemma 17

Proof. We will show that for any p, continuously adding mass to p will never decrease the distortion (even as the conditional expected value y changes according to the mass added). Since p_2 can be produced from p_1 by continuously adding $p_2 - p_1$ (i.e. we can define $p^{(\lambda)} = p_1 + \lambda(p_2 - p_1)$ and let λ smoothly increase from 0 to 1, and if $D^{(p^{(\lambda)})}$ is increasing the whole time then $D^{(p_2)} \geq D^{(p_1)}$).

So let $p : I \to \mathbb{R}_{\geq 0}$ such that $\int_I p(x)\,dx = 1$ WLOG (since scaling p will scale the distortion as it does not change y, and we can scale ε to match) and $z : I \to \mathbb{R}_{\geq 0}$ such that $\int_I z(x)\,dx = 1$ as well (z indicates the "direction" in which we increase p, so as long as $\int_I z(x)\,dx < \infty$, we can scale it WLOG to have total mass 1 on I). We then define $p_\varepsilon \triangleq p + \varepsilon z$ and make our claim:

$$\Big[\frac{d}{d\varepsilon} D^{(p_\varepsilon)}\Big]_{\varepsilon=0} \geq 0 \text{ for all such } p, z$$

If we prove this, then, as discussed above, we have proved the lemma.

First, defining $y^* \triangleq y^{(z)} - y^{(p)}$, we get

$$y^{(p_\varepsilon)} = \frac{\int_I p_\varepsilon(x) x\,dx}{\int_I p_\varepsilon(x)\,dx} \tag{B.57}$$

$$= \frac{\int_I (p(x) + \varepsilon z(x)) x\,dx}{\int_I (p(x) + \varepsilon z(x))\,dx} \tag{B.58}$$

$$= \frac{y^{(p)} + \varepsilon y^{(z)}}{1+\varepsilon} \tag{B.59}$$

$$= y^{(p)} + \frac{\varepsilon}{1+\varepsilon} y^* = y^{(p)}\Big(1 + \frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}\Big) \tag{B.60}$$

Therefore, we can write

$$D^{(p_\varepsilon)} = \int_I p_\varepsilon(x) x \log\Big(\frac{x}{y^{(p_\varepsilon)}}\Big)\,dx \tag{B.61}$$

$$= \int_I p(x) x \log\Big(\frac{x}{y^{(p_\varepsilon)}}\Big)\,dx + \varepsilon \int_I z(x) x \log\Big(\frac{x}{y^{(p_\varepsilon)}}\Big)\,dx \tag{B.62}$$

Let's isolate the first term $\int_I p(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx$. We can break it down further as

$$\int_I p(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx = \int_I p(x)x\left(\log\left(\frac{x}{y^{(p)}}\right)+\log\left(\frac{y^{(p)}}{y^{(p_\varepsilon)}}\right)\right)dx \tag{B.63}$$

$$= \int_I p(x)x\left(\log\left(\frac{x}{y^{(p)}}\right)-\log\left(1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}\right)\right)dx \tag{B.64}$$

$$= D^{(p)} - \int_I p(x)x\log\left(1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}\right)dx \tag{B.65}$$

Now we can take the derivative with respect to ε of this first part, which yields

$$\frac{d}{d\varepsilon}\int_I p(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx = -\frac{d}{d\varepsilon}\int_I p(x)x\log\left(1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}\right)dx \tag{B.66}$$

$$= -\int_I p(x)x\frac{d}{d\varepsilon}\log\left(1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}\right)dx \tag{B.67}$$

$$= -\int_I p(x)x\frac{1}{1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}}\frac{1}{(1+\varepsilon)^2}\frac{y^*}{y^{(p)}}\,dx \tag{B.68}$$

(switching the integral and derivative via the Leibniz rule). Evaluating at $\varepsilon = 0$ (which is the case we need as that is when we are adding a tiny bit of mass to p) gives

$$\left[-\int_I p(x)x\frac{1}{1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}}\frac{1}{(1+\varepsilon)^2}\frac{y^*}{y^{(p)}}\,dx\right]_{\varepsilon=0} = -\int_I p(x)x\frac{y^*}{y^{(p)}}\,dx = -y^* \tag{B.69}$$

because $\int_I p(x)x\,dx = y^{(p)}$.

Now we can isolate the second term $\varepsilon\int_I z(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx$ and take the derivative:

$$\frac{d}{d\varepsilon}\varepsilon\int_I z(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx = \int_I z(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx + \varepsilon\int_I z(x)x\frac{d}{d\varepsilon}\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx \tag{B.70}$$

From before we saw that

$$\frac{d}{d\varepsilon}\log\left(\frac{x}{y^{(p_\varepsilon)}}\right) = -\frac{1}{1+\frac{\varepsilon}{1+\varepsilon}\frac{y^*}{y^{(p)}}}\frac{1}{(1+\varepsilon)^2}\frac{y^*}{y^{(p)}} \tag{B.71}$$

(noting that the x in the numerator can be ignored). Noting that we need to evaluate at $\varepsilon = 0$, we observe that the ε in front of the integral eliminates that term entirely (and that $p_0 = p$) and we get

$$\left[\frac{d}{d\varepsilon}\varepsilon\int_I z(x)x\log\left(\frac{x}{y^{(p_\varepsilon)}}\right)dx\right]_{\varepsilon=0} = \int_I z(x)x\log\left(\frac{x}{y^{(p)}}\right)dx \tag{B.72}$$

We can therefore combine both parts to get

$$\left[\frac{d}{d\varepsilon}D^{(p_\varepsilon)}\right]_{\varepsilon=0} = \int_I z(x)x\log\left(\frac{x}{y^{(p)}}\right)dx - y^* \tag{B.73}$$

To prove this is nonnegative, we use the fact that z is actually a probability distribution over I (since it is nonnegative and $\int_I z(x)\,dx = 1$). Noting that $x\log(x/c)$ (for any constant $c > 0$) is a convex function and

$\mathbb{E}_{X\sim z}[X] = y^{(z)}$, we can use Jensen's inequality to get:

$$\int_I z(x)x\log\left(\frac{x}{y^{(p)}}\right)dx = \mathbb{E}_{X\sim z}\left[X\log\left(\frac{X}{y^{(p)}}\right)\right] \tag{B.74}$$

$$\geq \mathbb{E}_{X\sim Z}[X]\log\left(\frac{\mathbb{E}_{X\sim z}[X]}{y^{(p)}}\right) \tag{B.75}$$

$$= y^{(z)}\log\left(\frac{y^{(z)}}{y^{(p)}}\right) \tag{B.76}$$

$$= -y^{(z)}\log\left(\frac{y^{(p)}}{y^{(z)}}\right) \tag{B.77}$$

$$= -y^{(z)}\log\left(1+\frac{y^{(p)}-y^{(z)}}{y^{(z)}}\right) \tag{B.78}$$

$$\geq -y^{(z)}\frac{y^{(p)}-y^{(z)}}{y^{(z)}} \tag{B.79}$$

$$= -\left(y^{(p)}-y^{(z)}\right) \tag{B.80}$$

$$= y^* \tag{B.81}$$

(since $y^{(z)} > 0$ the $\log(1+t) \leq t$ inequality gets reversed). Thus, we get

$$\left[\frac{d}{d\varepsilon}D^{(p_\varepsilon)}\right]_{\varepsilon=0} \geq 0 \tag{B.82}$$

as desired, and the proof is complete. □

B.2.7 Proof of Theorem 12

Proof of Theorem 12. Let X_I denote a uniform source over interval I.

We first give a bound on $D(p_{X_{[a,b]}}, n)$ by scaling and translating from $D(p_{X_{[0,1]}}, n)$. By Theorem 15, we know $D(p_{X_{[0,1]}}, n) \leq (9/32)/n^2$; then we scale it out to a width of $b-a$, which requires a horizontal scaling by $b-a$ and a vertical scaling by $1/(b-a)$ (since the way we define horizontal scaling in Lemma 16 does not adjust the densities). This means that

$$D(p_{X_{[0,b-a]}}, n) \leq (b-a)\frac{9}{32}\frac{1}{n^2}\,. \tag{B.83}$$

Then we add a to the interval to go from $[0, b-a]$ to $[a,b]$, which multiplies the distortion of an interval centered at y by a factor of (at most) $y/(y+a)$. Noting that each interval is in $[0, b-a]$, we have

$$y \leq b-a \tag{B.84}$$

$$yb + ay \leq yb + ab + a^2 \tag{B.85}$$

$$\frac{y}{y+a} \leq \frac{b-a}{b} \tag{B.86}$$

so multiplying the distortion by this is an upper bound. Hence

$$D(p_{X_{[a,b]}}, n) \leq \frac{(b-a)^2}{b}\frac{9}{32}\frac{1}{n^2}\,. \tag{B.87}$$

□

B.2.8 Proof of Lemma 20

Proof. Choose $\gamma = 3/(2+s)$ and set the interval boundaries at $a_j = j^\gamma/n^\gamma$.

We will first compute the distortion on interval $[a_0, a_1] = [0, 1/n^\gamma]$. Select

$$y_1 = \mathbb{E}[X|X \in [0, a_1]] = \frac{\int_0^{a_1}(1+s)xx^s\,dx}{\int_0^{a_1}(1+s)x^s\,dx} \tag{B.88}$$

$$= \frac{(a_1^{2+s})/(2+s)}{(a_1^{1+s})/(1+s)} = \frac{1+s}{2+s}a_1\,. \tag{B.89}$$

$$D^{(p,[0,a_1])} = \int_0^{a_1}(1+s)x^s x\log\frac{x}{y_1}dx \tag{B.90}$$

$$= (1+s)\int_0^{a_1} x^s x\log x\,dx - \int_0^{a_1} x^s x\log\left(\frac{1+s}{2+s}a_1\right)\,dx \tag{B.91}$$

$$= a_1^{2+s}\left(\frac{1+s}{2+s}\log\frac{2+s}{1+s} - \frac{1+s}{(2+s)^2}\right) \tag{B.92}$$

$$= \left(\frac{1}{n^\gamma}\right)^{2+s}\left(\frac{1+s}{2+s}\log\frac{2+s}{1+s} - \frac{1+s}{(2+s)^2}\right) \tag{B.93}$$

$$= \frac{1}{n^3}C(s)\,. \tag{B.94}$$

We use $C(s)$ to denote that terms are a constant which do not depend on n. For $-1 < s < 0$, maximizing over s gives $C(s) < 0.162$.

Next we will compute the distortion on the remaining intervals. We will use Lemma 17 and bound each of the intervals by a uniform density and use Lemma 19 to compute an upper bound on the distortion.

Let u_j be the density where $u_j(x) = (1+s)a_j^s$. Then on $[a_j, a_{j+1}]$ for $j \geq 1$,

$$p(x) = (1+s)x^s \leq (1+s)a_j^s = u_j(x)\,. \tag{B.95}$$

$$D^{(p,[a_j,a_{j+1}])} \leq D^{(u_j,[a_j,a_{j+1}])} \tag{B.96}$$

$$= \int_{a_j}^{a_{j+1}}(1+s)a_j^s x\log\frac{x}{y^{(u_j,[a_j,a_{j+1}])}}\,dx \tag{B.97}$$

$$\leq (1+s)a_j^s\frac{1}{12}\frac{(a_{j+1}-a_j)^3}{\frac{1}{2}(a_{j+1}+a_j)} \tag{B.98}$$

$$= (1+s)\left(\frac{j^\gamma}{n^\gamma}\right)^s\frac{1}{12}\frac{\left(\frac{(j+1)^\gamma}{n^\gamma} - \frac{j^\gamma}{n^\gamma}\right)^3}{\frac{1}{2}\left(\frac{(j+1)^\gamma}{n^\gamma} + \frac{j^\gamma}{n^\gamma}\right)} \tag{B.99}$$

$$\leq (1+s)\frac{j^{\gamma s}}{n^{\gamma s}}\frac{1}{12}\frac{\left(\frac{\tilde{C}(\gamma)j^{\gamma-1}}{n^\gamma}\right)^3}{\frac{j^\gamma}{n^\gamma}} \tag{B.100}$$

$$= (1+s)\frac{(\tilde{C}(\gamma))^3}{12}\frac{j^{2\gamma+\gamma s-3}}{n^{2\gamma+\gamma s}} \tag{B.101}$$

$$= (1+s)\frac{(\tilde{C}(\gamma))^3}{12}\frac{1}{n^3}\,. \tag{B.102}$$

We use $\tilde{C}(\gamma)$ to denote a constant which only depends on γ. In order to find a bound on $\tilde{C}(\gamma)$ in terms of γ, we can first see that j^γ/n^γ is a convex function in j. This is because $-1 < s < 0$ and $\gamma = 3/(2+s)$, $3/2 < \gamma < 3$. We upper-bound the difference by approximating both points using the tangent line at

$(j+1)^\gamma/n^\gamma$. Recall that $j \geq 1$.

$$\frac{(j+1)^\gamma}{n^\gamma} - \frac{j^\gamma}{n^\gamma} \leq \gamma \frac{(j+1)^{\gamma-1}}{n^\gamma} \tag{B.103}$$

$$\leq \gamma \frac{(2j)^{\gamma-1}}{n^\gamma} \tag{B.104}$$

$$= \gamma 2^{\gamma-1} \frac{j^{\gamma-1}}{n^\gamma} \tag{B.105}$$

$$= \frac{\tilde{C}(\gamma) j^{\gamma-1}}{n^\gamma}\,. \tag{B.106}$$

Maximizing over s gives

$$(1+s)(\tilde{C}(\gamma))^3/12 \leq 6.332\,. \tag{B.107}$$

$$D(p,n) \leq \sum_{j=0}^{n-1} D^{(p,[a_j,a_{j+1}])} \tag{B.108}$$

$$\leq \frac{1}{n^3} C(s) + (1+s) \sum_{j=1}^{n-1} \frac{(\tilde{C}(\gamma))^3}{12} \frac{1}{n^3} \tag{B.109}$$

$$\leq \frac{19/3}{n^2} \tag{B.110}$$

□

B.2.9 Proof of Lemma 21

Fact 16 (Summation)**.** *If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ and are independent with the same β, where $i = 1, ..., k$, then $\sum_{i=1}^k X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$.*

Fact 17. *For $x \in \mathbb{R}_{>0}$, we have*

$$\frac{d^n}{dx^n}\Gamma(x) = \int_0^\infty (\log t)^n t^{x-1} e^{-t} dt \tag{B.111}$$

Fact 18 (ψ-function)**.** *Define*

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}\,. \tag{B.112}$$

where $\Gamma'(x) = \frac{d}{dx}\Gamma(x)$. Then

1. $\psi(x+1) = \psi(x) + \frac{1}{x}$
2. $\frac{1}{2x} \leq \log x - \psi(x) \leq \frac{1}{x}$

Proof. We want to compute a bound on $\mathbb{E}[S \log S]$ where $S \sim \text{Gamma}(\alpha, \alpha)$. First, consider when $S \sim$

Gamma(α, β).

$$\mathbb{E}[S \log S] = \int_0^\infty t \log t \frac{\beta^\alpha t^{\alpha-1} e^{-\beta t}}{\Gamma(\alpha)} dt \tag{B.113}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty (\log t) t^\alpha e^{-\beta t} dt \tag{B.114}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \log \frac{u}{\beta} \left(\frac{u}{\beta}\right)^\alpha e^{-u} \frac{1}{\beta} du \tag{B.115}$$

$$= \frac{1}{\beta\Gamma(\alpha)} \left(\int_0^\infty (-\log \beta)(u)^\alpha e^{-u} du + \int_0^\infty (\log u)(u)^\alpha e^{-u} du \right) \tag{B.116}$$

In the above, we substituted βt with u. Next, we will use Fact 17 to compute rightmost integral.

$$\mathbb{E}[S \log S] = \frac{1}{\beta\Gamma(\alpha)} \left(\Gamma(\alpha+1) \log \frac{1}{\beta} + \Gamma'(\alpha+1) \right) \tag{B.117}$$

$$= \frac{\alpha}{\beta} \log \frac{1}{\beta} + \frac{1}{\beta} \frac{\Gamma'(\alpha+1)}{\Gamma(\alpha)} \tag{B.118}$$

$$= \frac{\alpha}{\beta} \log \frac{1}{\beta} + \frac{1}{\beta} \frac{\Gamma'(\alpha+1)}{\Gamma(\alpha+1)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \tag{B.119}$$

$$= \frac{\alpha}{\beta}(-\log \beta + \psi(\alpha+1)) \tag{B.120}$$

In the last line, we used the definition of ψ from Fact 18. Next we will continue to use two statements in Fact 18 and set $\alpha = \beta$.

$$\mathbb{E}[S \log S] = \frac{\alpha}{\beta} \left(-\log \beta + \psi(\alpha) + \frac{1}{\alpha} \right) \tag{B.121}$$

$$= -\log \alpha + \psi(\alpha) + \frac{1}{\alpha} \tag{B.122}$$

$$\leq -\frac{1}{2\alpha} + \frac{1}{\alpha} \tag{B.123}$$

$$= \frac{1}{2\alpha}. \tag{B.124}$$

□

While we used an inequality on $\psi(\alpha)$ to show that $\mathbb{E}[S \log S] \leq 1/(2\alpha)$, asymptotically,

$$\psi(\alpha) \sim \log \alpha - 1/(2\alpha). \tag{B.125}$$

Thus we expect as α grows, $\mathbb{E}[S \log S] \sim 1/(2\alpha)$.

Appendix C

Compander Appendix

C.1 Experimental Data Overview

The purpose of our experiments is to demonstrate that our methods using companders do get good results. They also demonstrate that some of our asymptotic theoretical results hold non-asymptotically. Our experiments include one synthetic dataset on randomly generated distributions, and two datasets from based on *real-world* data.

Random Synthetic Distributions In the absence of any other information, it is typical to use a uniform prior. For our synthetic dataset, we draw random distributions from a uniform prior on the probability simplex.

The uniform prior is a member of the class of Dirichlet priors which are a commonly used class of priors for modeling probability. We say that the Dirichlet distribution is *symmetric* when the marginal distribution on each symbol is the same. This distribution can be parameterized by one value α. Choosing $\alpha = 1$ gives the uniform prior over the probability simplex. (In some other experiments, we change the value of α.)

For our experiments, we draw 1000 random distributions $\boldsymbol{x} = (x_1, ..., x_K)$. The KL divergence loss we get from these 1000 samples are average together in our results. Unless specified, we draw this from the uniform ($\alpha = 1$) prior.

Word Frequencies A natural source of distributions over large alphabets are distributions on words. The study of natural language processing historically has used many large alphabet distributions. While there are fairly elaborate models that exists, such as n-grams, which are order n Markov models of appearances of words in text, for our experiments, we stick to using the frequency of words which occur in a particular book as a probability distribution.

These frequencies are computed from text available on the Natural Language Toolkit (NLTK) libraries for Python. For each text, we get tokens (single words or punctuation) from each text and simply count the occurrence of each token.

DNA k-mers Large alphabet distributions also occur in biological applications, especially with DNA. DNA sequences from many different sources are stored for research purposes. The shear number of stored sequences strain storage resources. A particular statistic of DNA information which is relevant for research is k-mers. For a given sequence of DNA, the set of k-mers are the set of length k substrings which appear in the sequence. These frequency naturally form a probability distribution. The alphabet size of this distribution grows exponentially in k. For commonly used values of k such as 9, the alphabet size is already prohibitively large.

To get k-mer frequencies, our data source is the human genome sequences uploaded on December 2013 (on UCSC website). We select certain chromosomes and list the frequencies of k-mers from the sequences recorded. Parts of the sequences are marked as repeats (where viruses might have inserted DNA) and these are removed when we compute frequencies.

C.2 Power Compander

One of our original interests in the EDI compander was how simple the function $x^{2/3}$ (appears as the scale for the intervals near 0) is as a compander function. This led us to explore compander functions of the form $f(x) = x^s$ for some value $s \in (0, 1)$. We call this the power compander. The results were, very surprisingly, good. We discuss experiments using the power compander on the synthetic and real-world data. We apply this compander function so that the reconstruction point $\bar{\boldsymbol{y}}$ uses the midpoint of $I^{(n)}$.

Experiments Varying Power

For our first set of experiments, we plot the value of the KL divergence, $\mathcal{L}_K(W, x^s, N)$ for different values of s. When we do this, we get a rather interesting observation. The curve as a function of s empirically behaves similarly regardless of the distribution. In particular, for frequencies of words in books, plotted in Figure 7.4 regardless of the book, for the same choice of N, the curve corresponding to each book has nearly the exact same pattern.

Generally, the KL divergence is large for large s. If we set $s = 1$, this is equal to truncating. Decreasing s makes the intervals $I^{(n)}$ near 0 (small values of n) smaller relative to the intervals near 1 (larger values of n). Hence, lowering the value of s creates less KL divergence loss since smaller values are being quantized into smaller intervals, making the reconstruction points much closer to the original value. However, at some point while decreasing s, the KL divergence goes up. We can see this in Figure C.1 where we plot the performance of the power compander against synthetic distributions with different parameters. This also occurs for our real-world distributions. These are shown in Figure C.2 and Figure C.3 where we compare them to synthetic distributions with different α (parameter of the symmetric Dirichlet distribution). We show the comparison of the same chromosome but with different values of k (length of k-mer strings) in Figure C.5

In all these plots, there is a clearly a value of s which gives the minimum KL Divergence. The natural question to ask is what value of s achieves this. Empirically, this value can be approximated by

$$s = \frac{1}{\log K}\,. \tag{C.1}$$

This is most clearly observed in Figure C.4. We will show in the next section that (C.1) is indeed the correct power s to use for minimizing KL divergence.

C.2.1 Analysis of Power Compander

Starting with Theorem 17, we can use the asymptotic analysis to understand why the power compander works well for all distributions.

Proposition 29. *Let single-symbol density p be the marginal probability of one symbol on any symmetric probability distribution W over K symbols. For the power compander $f(x) = x^s$ where $s \le \frac{1}{2}$,*

$$\widehat{L}(p, f) \le \frac{1}{K}\frac{1}{24}s^{-2}K^{2s} \tag{C.2}$$

and for any symmetric prior W,

$$\widehat{\mathcal{L}}_K(W, x^s) \le \frac{1}{24}s^{-2}K^{2s}\,. \tag{C.3}$$

Optimizing over s gives

$$\widehat{\mathcal{L}}_K(W, x^s) \le \frac{e^2}{24}(\log K)^2 \tag{C.4}$$

Proof. Since $f(x) = x^s$ we have that $f'(x) = sx^{s-1}$. Using Theorem 17, this gives

$$\widehat{L}(p, f) = \frac{1}{24}s^{-2}\int_0^1 x^{1-2s}p(x)dx = \frac{1}{24}s^{-2}\mathbb{E}_{X\sim p}[X^{1-2s}]\,. \tag{C.5}$$

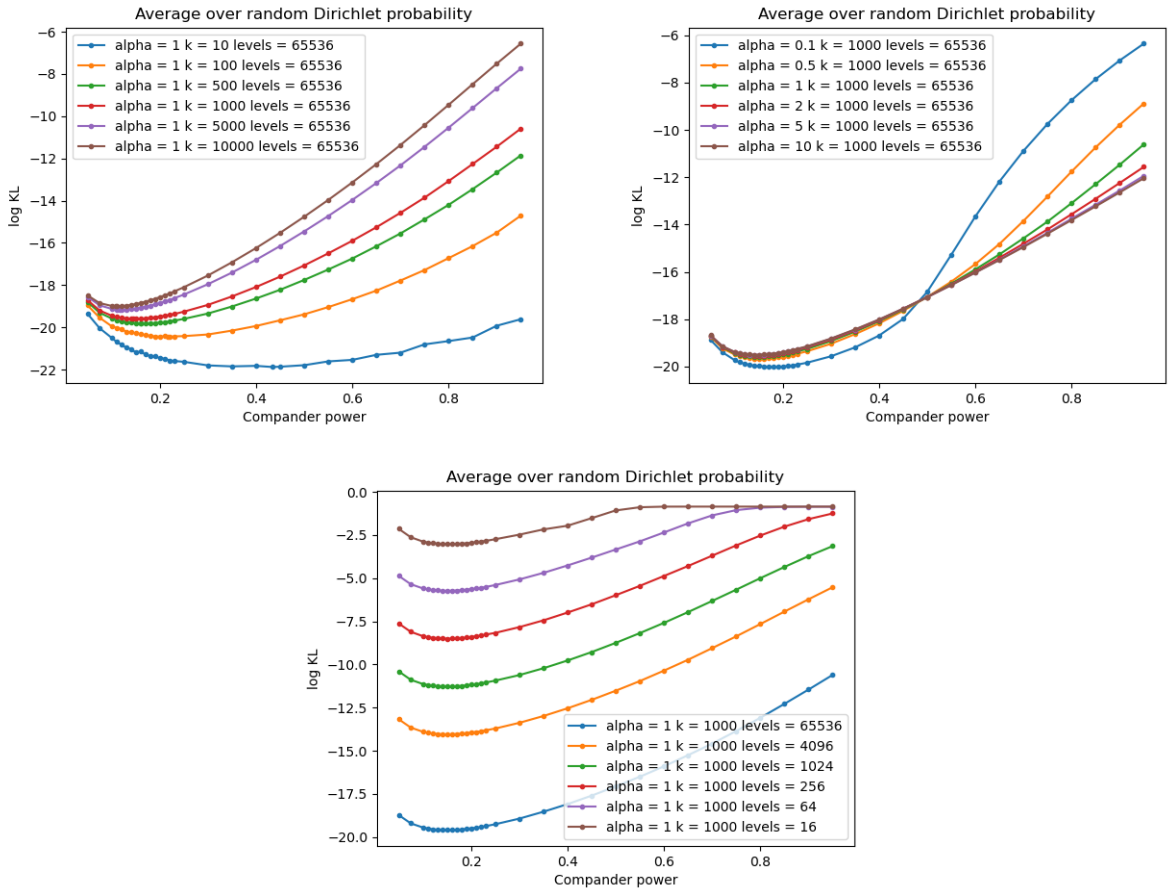


Figure C.1: Plots varying compander power while testing different α (Dirichlet parameter), K (alphabet size) and N (number of levels or granularity). Plots show logarithm of KL divergence on y-axis. We tested s at intervals of 0.05 over $[0, 1]$ with addition values between $[.10, .25]$ at intervals of .01

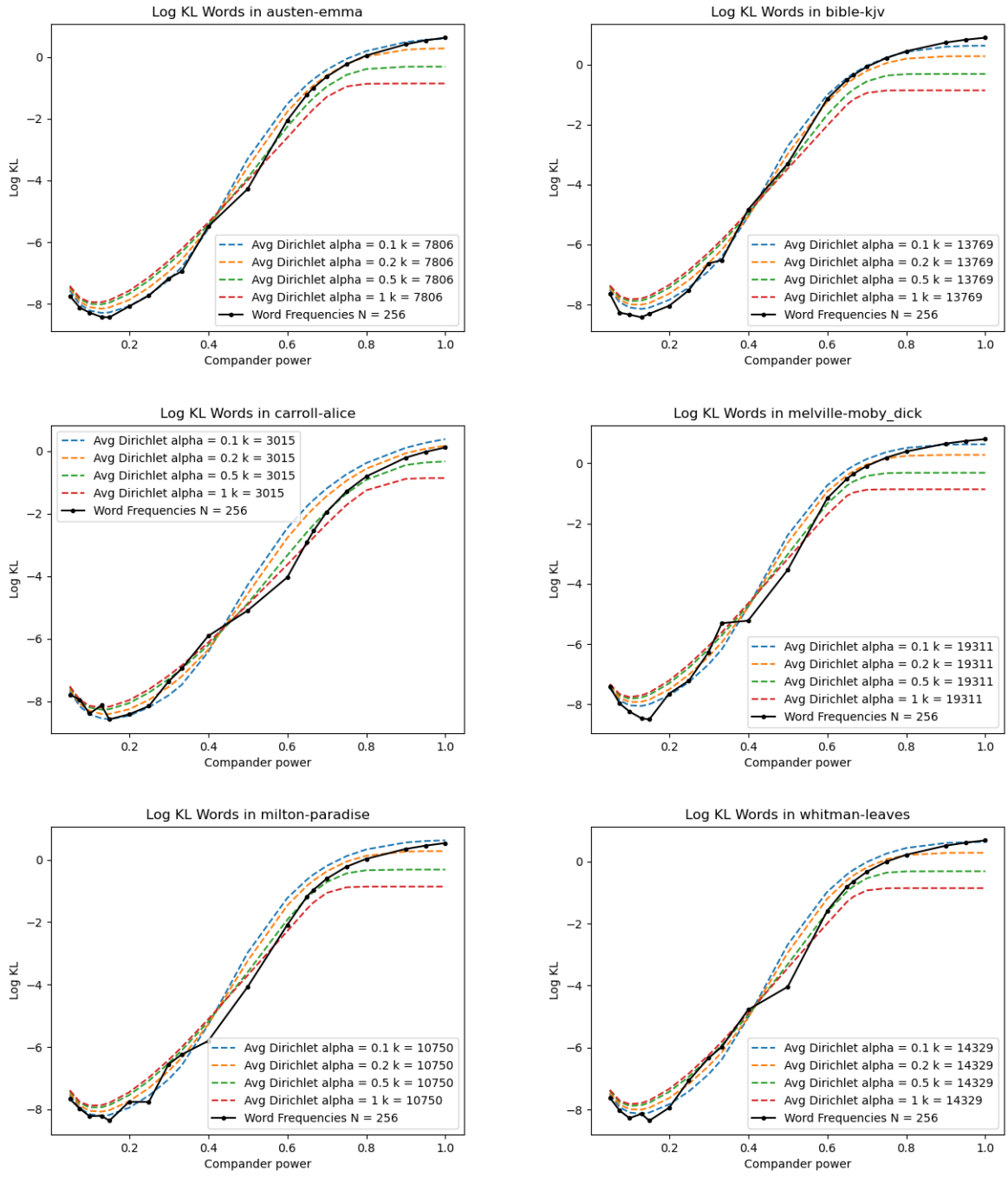


Figure C.2: Comparing the power compander used on word frequencies from different text to the power compander on random distributions drawn with symmetric Dirichlet prior with different α. Alphabet size K is chosen to match number of distinct words in the text and $N = 2^8$.

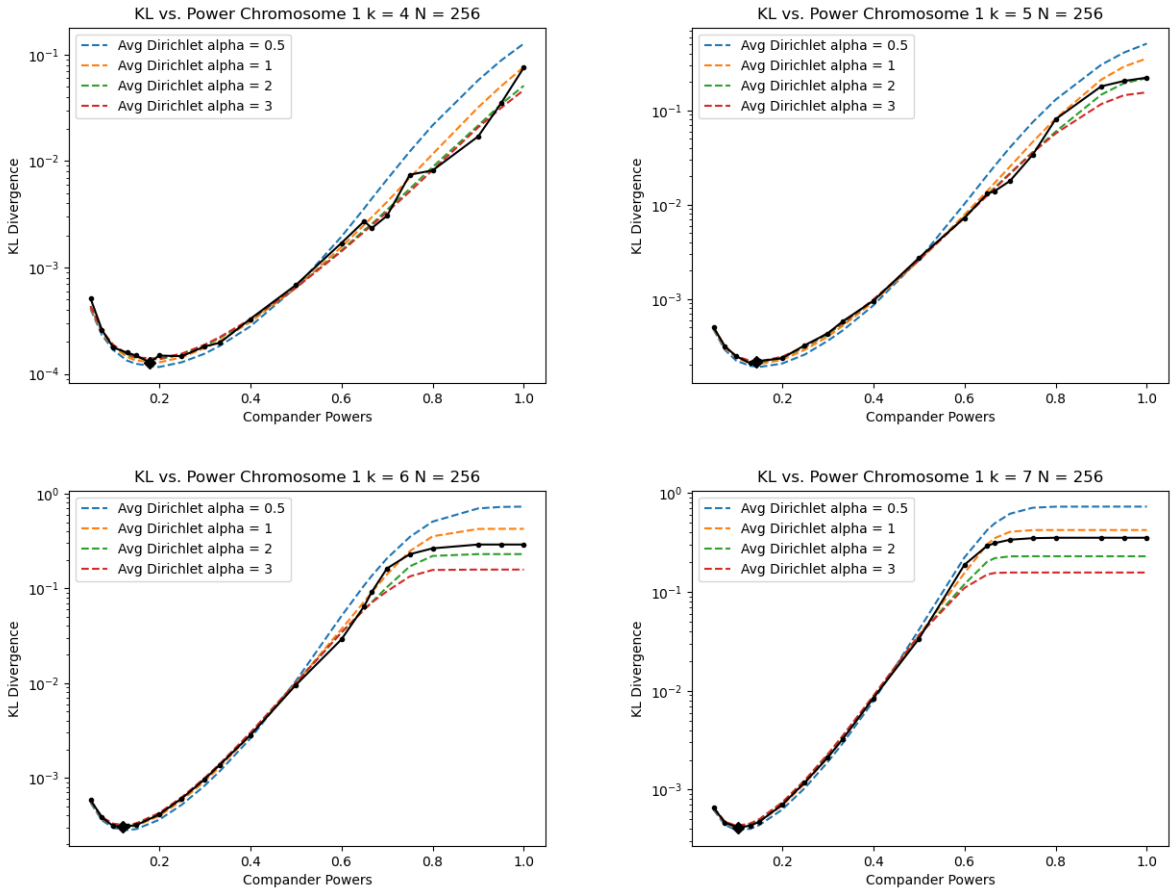


Figure C.3: Plots varying the compander power for quantizing frequencies of k-mers. The data is for chromosome 1. The plots range cover the range for $k = 4, 5, 6, 8$. Plotted against the KL divergence results are plots for synthetic Dirichlet drawn distributions with the same number of symbols. We show also the point corresponding to $s = \frac{1}{\log K}$ with the larger dot.

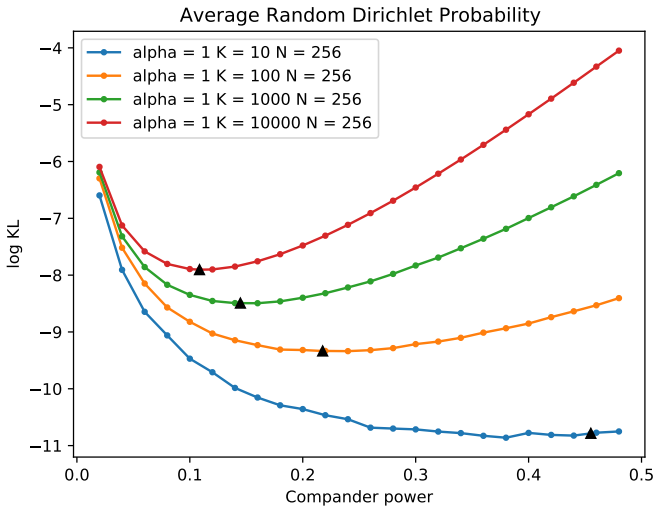


Figure C.4: Plots of synthetic distribution with different values of K. The triangle marks where $s = \frac{1}{\log K}$ occurs. This value of s seems to give the minimum KL divergence.

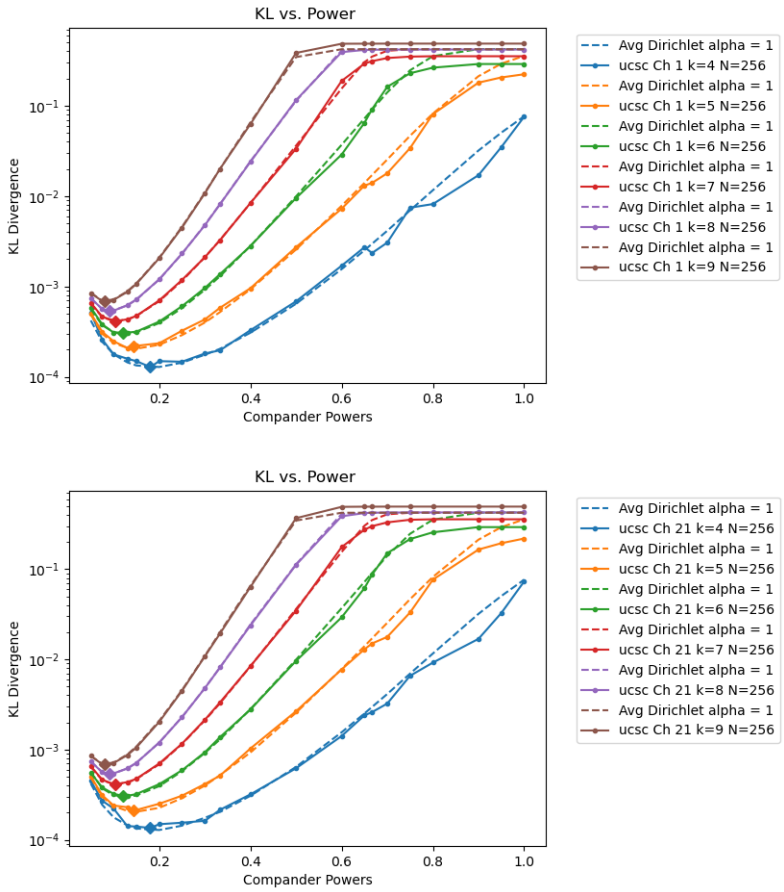


Figure C.5: Varying the compander power and the value of k (length of k-mers). For chromosomes 1 and 21.

Note that the function x^{1-2s} is increasing and also a concave function.

We want to find the worst-case prior distribution $W = (p_1, p_2, ..., p_K)$ over the simplex. The constraints are that

$$\sum_i \mathbb{E}_{X_i \sim p_i}[X_i] = 1 \tag{C.6}$$

$$\mathbb{E}_{(X_1,...,X_K)\sim p}\left[\sum_i X_i\right] = 1 \tag{C.7}$$

(another constraint is that values of $p(x)$ must sum to one, but we give a weaker constraint here).

We want to choose W to maximize

$$\sum_i \mathbb{E}_{X_i \sim p_i}[X_i^{1-2r}] = \mathbb{E}_{(X_1,...,X_K)\sim W}\left[\sum_i X_i^{1-2s}\right] \tag{C.8}$$

By concavity, the maximum solution is given when $X_1 = \cdots = X_K$. Therefore, the worst-case W is such that

$$p_1 = p_2 = \cdots = p_K = p = \frac{1}{K} \tag{C.9}$$

The density $p = \frac{1}{K}$ is a limit point of a continuous density which is of the form

$$p(x) = \frac{1}{2\varepsilon} \text{ on } \left[\frac{1}{K} - \varepsilon, \frac{1}{K} + \varepsilon\right]. \tag{C.10}$$

since we are restricting to continuous probability distribution (the cumulative distribution function is continuous).

Evaluating gives

$$\widehat{L}(p, f) = \frac{1}{24} s^{-2} \mathbb{E}_{X\sim p}[X^{1-2s}] \tag{C.11}$$

$$\leq \frac{1}{24} s^{-2} \left(\frac{1}{K}\right)^{1-2s} \tag{C.12}$$

$$= \frac{1}{K}\frac{1}{24} s^{-2} K^{2s} \tag{C.13}$$

which shows (C.4). Multiplying by K gives $\hat{\mathcal{L}}_K(p, f)$.

Finding the s which minimizes $\frac{1}{24} s^{-2} K^{2s}$ is equivalent to finding s which minimizes $s \log K - \log s$.

$$0 = \frac{d}{ds} s \log K - \log s = \log K - \frac{1}{s} \tag{C.14}$$

$$\implies s = \frac{1}{\log K} \tag{C.15}$$

We can plug this back into our equation, using the fact that $e^{\log K} = K$ implies that $K^{1/\log K} = e$.

Our final result is that using $f(x) = x^{1/\log K}$ gives a worst-case bound for any prior W is

$$\widehat{\mathcal{L}}_K(W, f) \leq \frac{e^2}{24} (\log K)^2 \tag{C.16}$$

□

The power compander turns out to give guarantees bounds on the value on $\widehat{\mathcal{L}}_K(W, f)$ when f is chosen so that $s = 1/\log K$.

C.3 Optimal Compander for One Value

As alluded to in Section 7.6.2, finding the worst-case density $p(x)$ that maximizes $\int (p(x)x^{-1})^{1/3}dx$ leads to an interesting corollary.

Corollary 8. The single-symbol density p which produces the largest value of $\min_f L^\dagger(p,f)$ (i.e. the source p which is most difficult to quantize accurately with a compander) is

$$p(x) = \frac{1}{2}x^{-1/2}\,. \tag{C.17}$$

and $f(x) = x^{1/2}$ is the optimal compander. Then

$$\widehat{L}(p,f) = \frac{1}{24}\bigg(\int_0^1 \big(p(x)x^{-1}\big)^{1/3}\,dx\bigg)^3 = \frac{1}{24}(2^{2/3})^3 = \frac{1}{6} \tag{C.18}$$

Proof. The optimal compander f satisfies $f'(x) \propto (p(x)x^{-1})^{1/3} \propto x^{-1/2}$. Normalizing gives $f(x) = x^{1/2}$. □

Remark 11. *Furthermore, on $\triangle_2$ (the simplex on 3 variables), the Dirichlet prior $Dir_3(1/2)$ has $p(x) = \frac{1}{2}x^{-1/2}$ as the marginal on all three symbols, making it automatically the most difficult prior to compand on $\triangle_2$.*

While this is interesting, we note that $\mathbb{E}_{X\sim p}[X] = 1/3$ when $p(x) = \frac{1}{2}x^{-1/2}$. This means that for large-alphabet distributions (over the simplex $\triangle_{K-1}$ where K is large), the prior cannot have this p over many different symbols since the sum of the values must be equal to 1.

In [90], the authors computed the optimal quantization function for a binary alphabet using worst-case KL divergence loss. When this quantization function is described as a compander function f, then

$$f(x) = \frac{\cos^{-1}(1-2x)}{2\pi}\,. \tag{C.19}$$

This function does not generalize well to more than 2 symbols, however it turns how that if we were to use it anyways and test it on our experimental dataset, its average KL loss is very similar to when the compander function is $x^{1/2}$. This makes sense since for small x, the functions $x^{1/2}$ and (C.19) look similar when scaled by some multiplicative constant.

C.4 Beta Companders for Symmetric Dirichlet Priors

In this section, we want to derive the optimal compander f_p for symmetric Dirichlet distributions. The following fact is necessary:

Fact 19. *For $\boldsymbol{x} \sim Dir(\alpha_1,\dots,\alpha_K)$, the marginal distribution on x_k is $x_k \sim \mathrm{Beta}(\alpha_k,\beta_k)$, where $\beta_k = \sum_{j\neq k}\alpha_k$ (of course, the x_k are not independent). When the prior is symmetric with parameter α, we get that all x_k are distributed according to $\mathrm{Beta}(\alpha,(K-1)\alpha)$.*

As a corollary to Theorem 17 and Fact 19, we get the following:

Corollary 9. *When $\boldsymbol{x} \sim Dir_K(\alpha)$, let $p(x)$ be the associated single-symbol density (same for all elements due to symmetry). The optimal compander for p satisfies*

$$f'(x) = B\Big(\frac{\alpha+1}{3},\frac{(K-1)\alpha+2}{3}\Big)^{-1}x^{(\alpha-2)/3}(1-x)^{((K-1)\alpha-1)/3} \tag{C.20}$$

where $B(a,b)$ is the Beta function. Therefore, $f(x)$ is the normalized incomplete Beta function $I_x((\alpha+1)/3,((K-1)\alpha+2)/3)$.

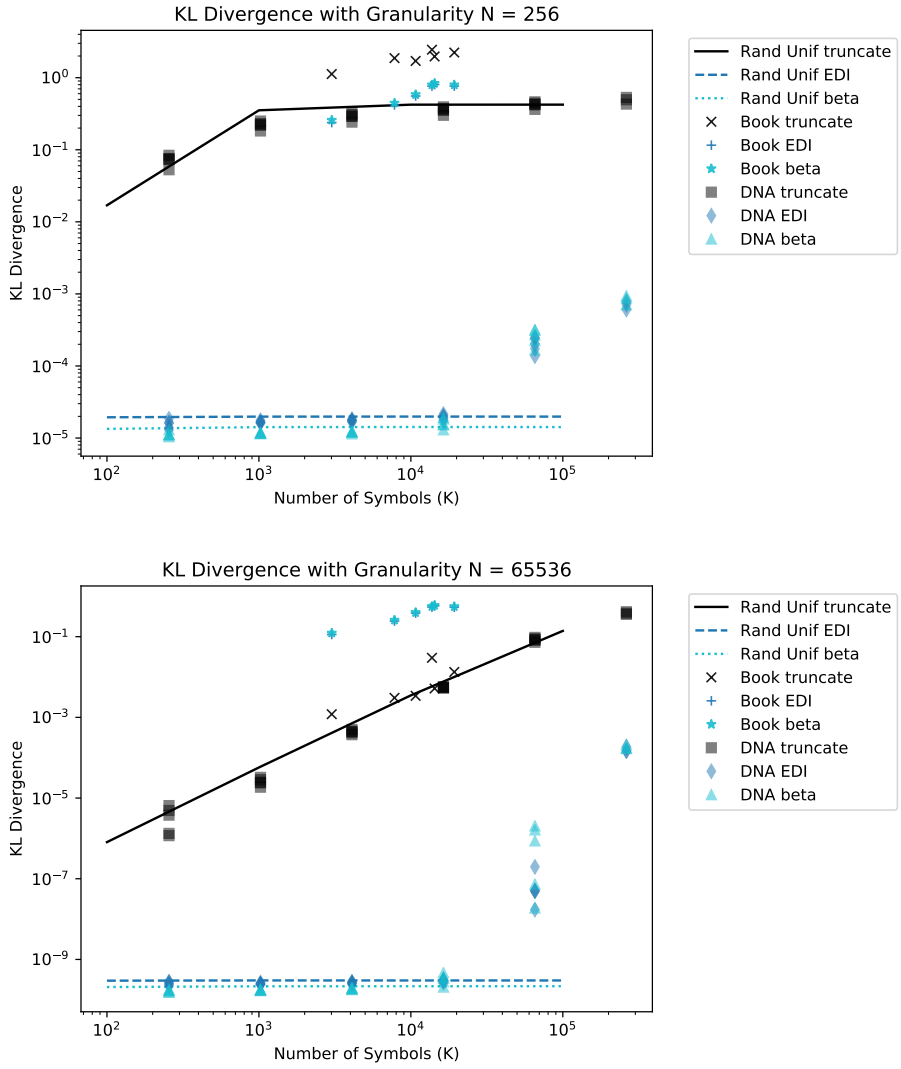


Figure C.6: Comparing the beta compander and the EDI method.

Then

$$\widehat{L}(p,f) = \frac{1}{2}B\Big(\frac{\alpha+1}{3}, \frac{(K-1)\alpha+2}{3}\Big)^3 B(\alpha,(K-1)\alpha)^{-1} \tag{C.21}$$

Remark 12. *Since* (C.21) *scales with* K^{-1}*, this means that* $\widehat{\mathcal{L}}_K(Dir_K(\alpha), f)$ *and* $\mathcal{L}_K(Dir_K(\alpha), f)$ *are constants with respect to* K*. This is consistent with what we get with the EDI compander* (6.4)*.*

We will call this compander the *beta compander* since the companding function is defined as an incomplete beta function. The beta compander is most similar to the EDI method, which also has $\mathcal{L}_K(\mathrm{Dir}_K(\alpha), f)$ which is constant with K. The beta compander naturally performs better than the EDI method since this beta compander is optimized to do so (the EDI method is primarily a tool for deriving bounds). We can see the comparison in Figure C.6, where the beta compander is better than the EDI method by a constant amount for all K.

The beta compander is not the easiest algorithm to implement however. It is necessary to compute an incomplete beta function in order to find the compander function f. (In contrast, applying the minimax compander does not require an integral step.)

C.5 Analysis of Truncate Companding, μ-Law and A-Law

Using Theorem 16, we can look at how various companding functions $f(x)$ perform with different $p(x)$. In this section, we want to assess how well the truncate compander (this is when $f(x) = x$; we can also call it the identity compander) and the μ-law and A-law companders work.

Definitions of μ-law and A-law The companding function for μ-law for $-1 \leq x \leq 1$ is

$$f(x) = \operatorname{sgn}(x) \frac{\log(1 + \mu|x|)}{\log(1 + \mu)} \tag{C.22}$$

where μ is generally set to 255.

The companding function for A-law is

$$f(x) = \operatorname{sgn}(x) \begin{cases} \frac{A|x|}{1+\log A} & \text{if } |x| < \frac{1}{A} \\ \frac{1+\log(A|x|)}{1+\log A} & \text{if } \frac{1}{A} \leq |x| \leq 1 \end{cases} \tag{C.23}$$

where A is generally set to 87.6.

Infinite Asymptotic Normalized Expected Divgerence Starting for f the truncate (identity) compander, we have

$$\widehat{L}(p, f) = \frac{1}{24} \int_0^1 \frac{p(x)}{x f'(x)} dx = \int_0^1 \frac{p(x)}{x} dx \,. \tag{C.24}$$

If the value $p(x)$ does not go to zero sufficiently quickly at $x = 0$, then the integral will not be finite. This is the major problem with the truncate compander. Distributions with marginals that have mass at zero will not quantize well.

The same thing occurs with the μ-law and A-law companders. For f which is the compander function for the μ-law,

$$\widehat{L}(p, f) = \frac{1}{24} \int_0^1 \frac{p(x)}{x f'(x)} dx \tag{C.25}$$

$$= \frac{1}{24} \int_0^1 \left(\frac{\log(1 + \mu)}{\mu} \right)^2 p(x) \frac{(1 + \mu x)^2}{x} dx \tag{C.26}$$

$$= \frac{1}{24} \left(\frac{\log(1 + \mu)}{\mu} \right)^2 \int_0^1 p(x) \left(\frac{1}{x} + 2\mu + \mu^2 x \right) dx \,. \tag{C.27}$$

Similarly if we let f be the A-law, we have

$$\widehat{L}(p, f) = \frac{1}{24} \int_0^1 \frac{p(x)}{x f'(x)} dx \tag{C.28}$$

$$= \frac{1}{24} \int_0^{1/A} \left(\frac{1 + \log A}{A} \right)^2 \frac{p(x)}{x} dx + \frac{1}{24} \int_{1/A}^1 (1 + \log A)^2 x^2 \frac{p(x)}{x} dx \,. \tag{C.29}$$

Both the μ-law and A-law have a problematic term $1/x$ which needs to be integrated. This will cause the integral to be infinite if $\lim_{x \to 0} p(x)$ is not 0 (or if it does not go to zero quickly enough.)

For instance, the value of $\widehat{L}(p, f)$ is infinite for these companders if

$$p(x) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)\Gamma((K-1)\alpha)} x^{\alpha - 1} (1 - x)^{(K-1)\alpha - 1} \,. \tag{C.30}$$

for $\alpha \leq 1$. This $p(x)$ has a beta distribution, which is the marginal distribution which occurs for one symbol under a probability which is generated from a symmetric Dirichlet prior with alphabet size K and parameter α.

If $\alpha \leq 1$, then $\widehat{L}(p,f)$ is infinite; however, if $\alpha > 1$, then $\widehat{L}(p,f)$ can be finite if we adjust the value of μ or A. For the rest of the analysis, we will focus on the μ-law (since the A-law is similar).Suppose $\alpha > 1$, then with f as the compander function from the μ-law, we have

$$\widehat{L}(p,f) = \frac{1}{24}\left(\frac{\log(1+\mu)}{\mu}\right)^2 \left(\int_0^1 \frac{\Gamma(K\alpha)}{\Gamma(\alpha)\Gamma((K-1)\alpha)} x^{\alpha-2}(1-x)^{(K-1)\alpha-1}dx + 2\mu + \mu^2\mathbb{E}[x]\right) \tag{C.31}$$

$$= \frac{1}{24}\left(\frac{\log(1+\mu)}{\mu}\right)^2 \left(\frac{K\alpha-1}{\alpha-1} + 2\mu + \mu^2\frac{1}{K}\right) \tag{C.32}$$

$$= \frac{1}{24}\left(\log(1+\mu)\right)^2 \left(\frac{K\alpha-1}{\mu^2(\alpha-1)} + 2\frac{1}{\mu} + \frac{1}{K}\right). \tag{C.33}$$

For such a case, we can set $\mu = K$ (this is roughly optimal up to constant). This gives that $\widehat{L}(p,f) \approx \frac{1}{K}(\log K)^2$. This is the distortion for one symbol. For K symbols, the distortion is $(\log K)^2$. This illustrates that even for distributions p where the integral for $\widehat{L}(p,f)$ is not infinite, there is still a need to optimize μ in terms of K in order to get reasonable distortion.

Behavior of μ-Law with N Plots comparing the performance of the truncate, μ-law and A-law companders are in Figure C.7. We notice that though we computed $\widehat{L}(p,f)$ is infinite for the μ-law, the behavior of μ-law when $\mu = K$ resembles the behavior of the minimax compander even when testing on a the synthetic distributions with $\alpha = 1$ (the case which $\widehat{L}(p,f)$ is infinite). The performance seems to only be off by a constant as K increases. This is because $\widehat{L}(p,f)$ being infinite comes into affect as N increases, not necessarily as K increases.

In the case where the integral $\widehat{L}(p,f)$ is infinite, the performance of the compander does not behave as $O(1/N^2)$. However, it can still have a performance which is close to $O(1/N^2)$, like say $O((\log N)/N^2)$. For small values of N, this performance will not differ so much from $O(1/N^2)$. This is in fact what happens with μ-law companding.

Starting at the beginning, we will create intervals $I^{(n)}$ according to the companding given by the μ-law. Let y_n be the midpoint of these intervals. Then, we have that the expected divergence is

$$\mathbb{E}\widehat{L}(p,f,N) = \sum_{n=1}^{N} \int_{I^{(n)}} p(x) x \log \frac{x}{y_n} dx \tag{C.34}$$

$$\leq \sum_{n=1}^{N} \int_{I^{(n)}} p(x) \frac{(x-y_n)^2}{y_n} dx \tag{C.35}$$

$$\tag{C.36}$$

We approximate that $p(x)$ is uniform over each interval and is close to $p(y_n)$. The interval length is approximated by $1/(Nf'(y_n))$. Let $\delta = x - y_n$. This gives

$$\mathbb{E}\widehat{L}(p,f,N) \approx \sum_{n=1}^{N} \int_{I^{(n)}} p(y_n) \frac{\delta^2}{y_n} d\delta \tag{C.37}$$

$$= \sum_{n=1}^{N} \int_{-\frac{1}{2}\frac{1}{Nf'(y_n)}}^{\frac{1}{2}\frac{1}{Nf'(y_n)}} p(y_n) \frac{\delta^2}{y_n} d\delta \tag{C.38}$$

$$= \sum_{n=1}^{N} p(y_n) \frac{1}{y_n} \frac{1}{24} \frac{1}{N^3 (f'(y_n))^3} \tag{C.39}$$

$$\tag{C.40}$$

We now will compute approximately what the values of y_n are under the μ-law companding. We know that the boundaries of the intervals, which we will express as $b_0, ..., b_N$ are given as $b_0 = 1$ and b_j for $j > 0$

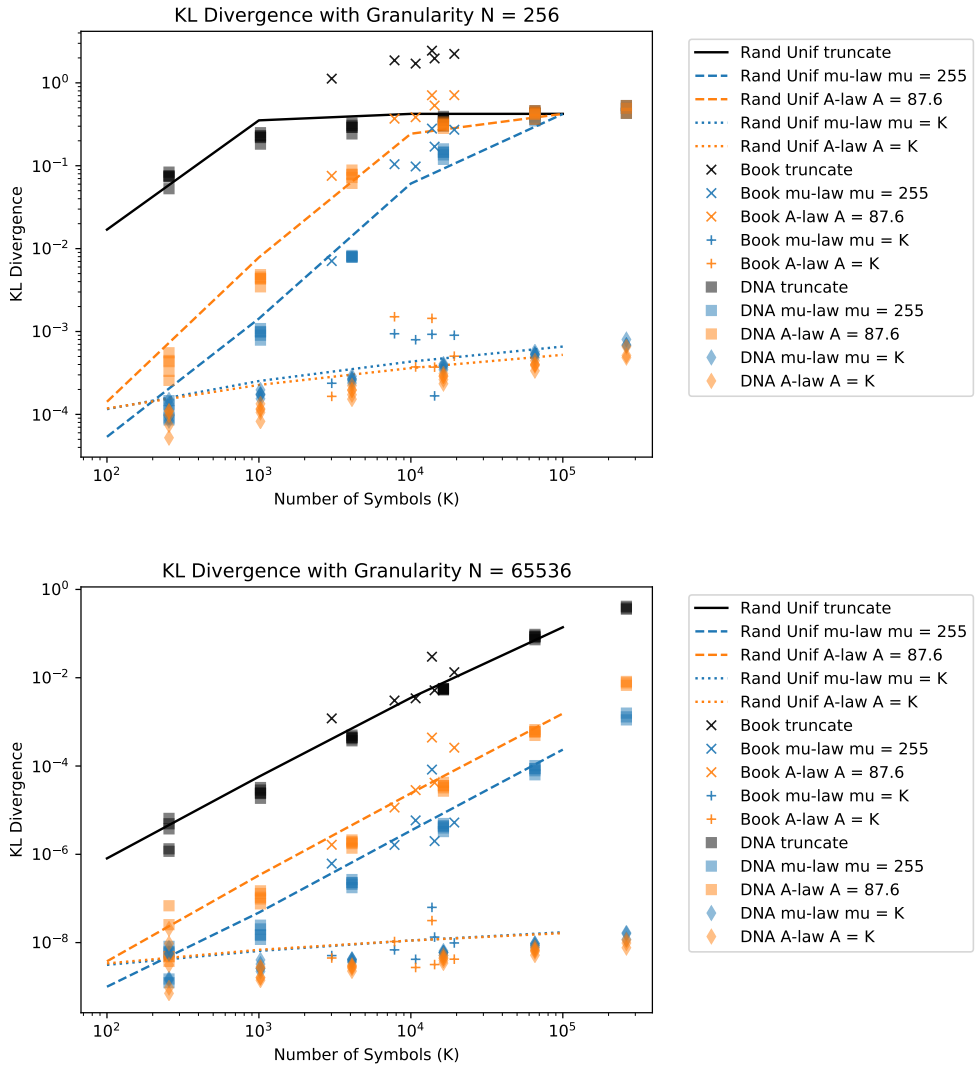


Figure C.7: Plots comparing performance of the truncate, μ-law and A-law companders with standard values for μ and A, and μ-law and A-law with K as the value for μ and A.

can be expressed by

$$f(b_j) = \frac{\log(1+\mu b_j)}{\log(1+\mu)} = \frac{j}{N} \tag{C.41}$$

For small values of x, like say $x \leq 1/\mu$, we can use the approximation that $\log(1+\mu x) \approx \mu x$. This gives that

$$b_j = \frac{\log(1+\mu)}{\mu}\frac{j}{N} \tag{C.42}$$

If we want to be more precise, we can try to bound b_j precisely. We have for when $b_j \leq \frac{1}{\mu}$

$$\log 2 \frac{\mu b_j}{\log(1+\mu)} \leq \frac{\log(1+\mu b_j)}{\log(1+\mu)} \leq \frac{\mu b_j}{\log(1+\mu)} \tag{C.43}$$

and hence we get that

$$\log 2 \frac{\log(1+\mu)}{\mu}\frac{j}{N} \leq b_j \leq \frac{\log(1+\mu)}{\mu}\frac{j}{N} \tag{C.44}$$

We can use this when $j \leq \frac{N}{\log(1+\mu)}$.

When $b_j \geq \frac{1}{\mu}$ (and μ is large, say $\mu > 1$), we can use the approximation that

$$\frac{\log(1+\mu b_j)}{\log(1+\mu)} \approx \frac{\log(\mu x)}{\log \mu} \tag{C.45}$$

which gives

$$b_j \approx u^{1-j/N} \tag{C.46}$$

For the purposes of determining the right behavior, we can use that $y_n = b_n$. Under this assumption, we have that

$$\sum_{n=1}^{N} \frac{1}{y_n} \approx \sum_{n=1}^{\frac{N}{\log(1+\mu)}} \frac{\mu}{\log(1+\mu)}\frac{N}{j} + \sum_{n=\frac{N}{\log(1+\mu)}}^{N} \frac{1}{u^{1-j/N}} \tag{C.47}$$

$$\leq \frac{\mu}{\log(1+\mu)} N \sum_{n=1}^{\frac{N}{\log(1+\mu)}} \frac{1}{j} + \sum_{n=\frac{N}{\log(1+\mu)}}^{N} 1 \tag{C.48}$$

$$\approx \frac{\mu}{\log(1+\mu)} N \log\left(\frac{N}{\log(1+\mu)}\right) + \left(N - \frac{N}{\log(1+\mu)}\right) \tag{C.49}$$

$$\tag{C.50}$$

We can see the $N \log N$ term appearing. The behavior will still need to depend on $p(x)$. (As we showed above, if $p(x)$ corresponds to a symmetric Dirichlet with $\alpha > 1$, we will get a behavior in $O(1/N^2)$.) For symmetric Dirichlet with $\alpha = 1$, we have

$$p(x) = \frac{K\alpha - 1}{\alpha - 1}(1-x)^{(K-1)\alpha - 1} \leq \frac{K\alpha - 1}{\alpha - 1} \tag{C.51}$$

If we will treat $p(x)$ as a uniform function, then $\widehat{L}(p, f, N)$ scales as $O((\log N)/N^2)$.

We can see from Figure C.8 that indeed the other companders where $\widehat{L}(p, f)$ is finite have divergences which are constant when multiplied by N^2. However, the curve for the μ-law increasing as N increases.

Next, we will consider what is the worst-case $p(x)$ that gets the largest performance in terns of N. Recall the issue with the μ-law compander is that we are left with a term which depends only on the probability $p(x)$ and $1/y_n$.

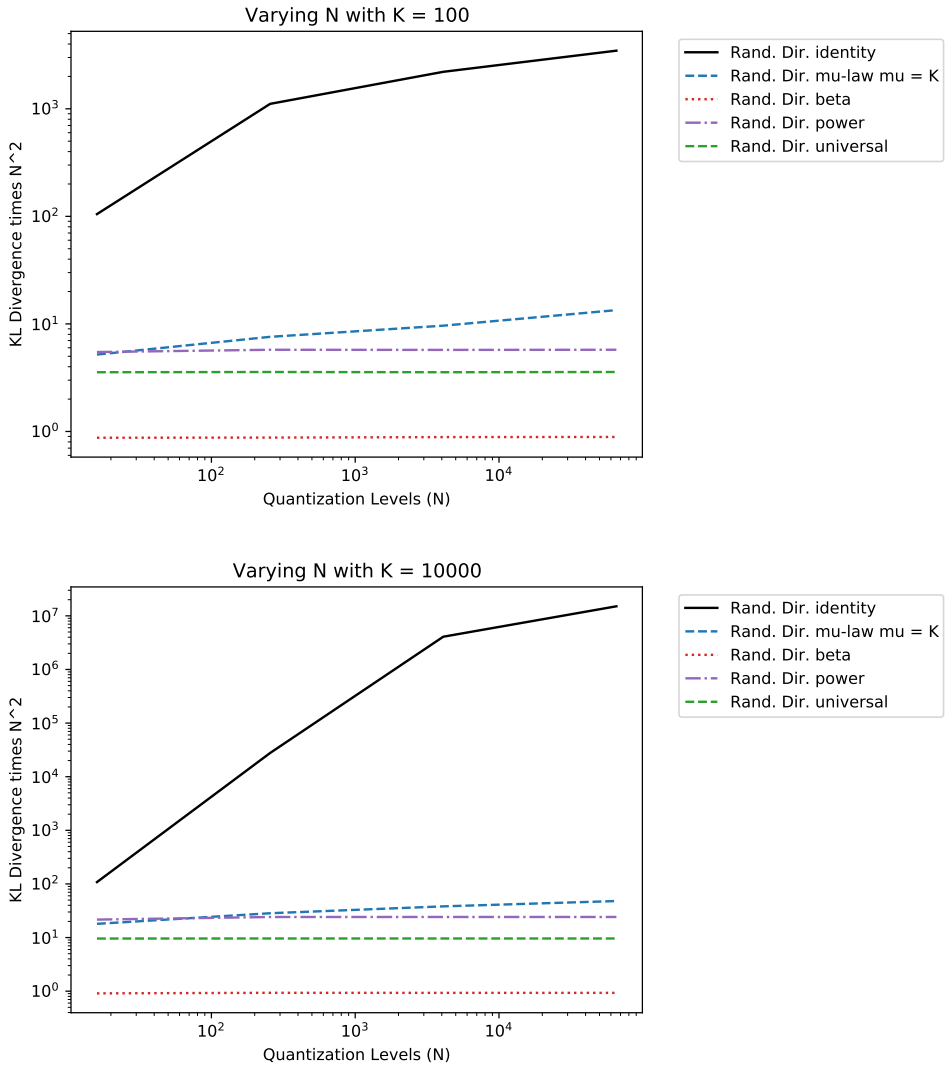


Figure C.8: Fixing the alphabet size and then varying the quantization levels N. The plot gives the average KL divergences times N^2 for synthetic distributions under uniform prior.

Writing this slightly different than above, we have

$$\widehat{L}(p,f,N) = \sum_{n=1}^{N} \int_{I^{(n)}} p(x) \frac{\left(\frac{1}{Nf'(y_n)}\right)^2}{y_n} dx \tag{C.52}$$

$$\approx \sum_{n=1}^{N} \left(\int_{I^{(n)}} p(x)dx \right) \frac{1}{y_n} \frac{1}{N^2 (g'(y_n))^2} \tag{C.53}$$

(See above to note that $1/(g'(y_n))^2$ contributes a constant term.)

Given a value of N, we want to pick a $p(x)$ so that

$$\int_{I^{(n)}} p(x)dx \propto \frac{1}{y_n} \tag{C.54}$$

We should then set

$$\int_{I^{(n)}} p(x)dx = \frac{1}{n} \frac{1}{\log N} \tag{C.55}$$

which will match the values of y_n for the smaller values (the larger values do affect the sum as much.)

Evaluating gives that

$$\widehat{L}(p,f,N) = \sum_{n=1}^{N} \frac{1}{n} \frac{1}{\log N} \frac{\log(1+\mu)}{\mu} \frac{N}{n} \frac{1}{N^2 (f'(y_n))^2} = O\left(\frac{1}{N \log N} \right) \tag{C.56}$$

since $\sum_{i=1}^{N} \frac{1}{n^2}$ is finite.

C.6 Analysis of Minimax Companding Constant

Determining bounds on c_K If $a_K, b_K \geq 0$, then $p(x)$ is well-behaved (and bigger than 0) and we are done. Now we set

$$a_K = 4b_K^{-2} - a_K \tag{C.57}$$

We must meet the condition that

$$\mathbb{E}_{X\sim p}[X] = \int_0^1 x \left(a_K x^{1/3} + b_K x^{4/3} \right)^{-3/2} dx \tag{C.58}$$

$$= \frac{-2}{b_K \sqrt{a_K + b_K}} + \frac{2 \sinh^{-1}\left(\sqrt{\frac{b_K}{a_K}} \right)}{b_K^{3/2}} \tag{C.59}$$

The constraint that $\int_0^1 p(x)\,dx = 1$ requires that $a_K \sqrt{a_K + b_K} = 2$. We can use this to get

$$\mathbb{E}_{X\sim p}[X] = \frac{-a_K}{b_K} + \frac{a_K \sqrt{\frac{a_K}{b_K} + 1} \sinh^{-1}\left(\sqrt{\frac{b_K}{a_K}} \right)}{b_K} \tag{C.60}$$

$$= \frac{-1}{r} + \frac{\sqrt{\frac{1}{r} + 1} \sinh^{-1}\left(\sqrt{r} \right)}{r} \tag{C.61}$$

$$= \frac{-1}{r} + \frac{\sqrt{\frac{1}{r} + 1} \log\left(\sqrt{r} + \sqrt{r+1} \right)}{r} \tag{C.62}$$

where we use $r = b_K/a_K$. We will find upper and lower bounds in order to approximate what r should be.

$$\mathbb{E}_{X\sim p}[X] \leq \frac{1}{2}\frac{\log r}{r} \tag{C.63}$$

so long as $r > 3$. If we choose $r = c_K K \log K$ and set $c_K = .75$, then

$$\mathbb{E}_{X\sim p}[X] \leq \frac{1}{2}\frac{\log(c_K K \log K)}{c_K K \log K} \leq \frac{1}{2c_K K} + \frac{\log\log K}{2c_K K \log K} + \frac{\log c_K}{2c_K K \log K} \leq \frac{1}{K} \tag{C.64}$$

so long as $K > 4$. We also have

$$\mathbb{E}_{X\sim p}[X] \geq \frac{1}{3}\frac{\log r}{r} \tag{C.65}$$

for all r. If we choose $r = c_K K \log K$ and set $c_K = .25$, then

$$\mathbb{E}_{X\sim p}[X] \geq \frac{1}{3}\frac{\log(c_K K \log K)}{c_K K \log K} \geq \frac{1}{K} \tag{C.66}$$

so long as $K > 24$.

Changing the value of c_K changes the value of $\mathbb{E}_{X\sim p}[X]$ continuously. Hence, for each $K > 24$, there exists a c_K so that if $r = c_K K \log K$, then

$$\mathbb{E}_{X\sim p}[X] = \frac{1}{K}\,. \tag{C.67}$$

Limiting value of c_K

Lemma 24. *In the limit, $c_K \to 1/2$.*

Proof. We start with $r = \frac{b_K}{a_K} = c_K K \log K$, and we need to meet the condition that

$$\frac{-1}{r} + \frac{\sqrt{\frac{1}{r}+1}\log\left(\sqrt{r}+\sqrt{r+1}\right)}{r} = \frac{1}{K}\,. \tag{C.68}$$

Substituting we get

$$\frac{-1}{c_K K \log K} + \frac{\sqrt{\frac{1}{c_K K \log K}+1}\log\left(\sqrt{c_K K \log K}+\sqrt{c_K K \log K+1}\right)}{c_K K \log K} = \frac{1}{K} \tag{C.69}$$

$$\frac{-1}{\log K} + \frac{\sqrt{\frac{1}{c_K K \log K}+1}\log\left(\sqrt{c_K K \log K}+\sqrt{c_K K \log K+1}\right)}{\log K} = c_K \tag{C.70}$$

$$\tag{C.71}$$

Let $c = \lim_{K\to\infty} c_K$. Since c_K is bounded, we know that $\lim_{K\to\infty} c_K K \log K \to \infty$ since c_K is bounded above by $3/4$.

$$c = \lim_{K\to\infty} \frac{-1}{\log K} + \frac{\sqrt{\frac{1}{c_K K \log K}+1}\log\left(\sqrt{c_K K \log K}+\sqrt{c_K K \log K+1}\right)}{\log K} \tag{C.72}$$

$$= 0 + 1 \lim_{K\to\infty} \frac{\log\left(2\sqrt{c_K K \log K}\right)}{\log K} \tag{C.73}$$

$$= \lim_{K\to\infty} \frac{\log 2 + \frac{1}{2}\log c_K + \frac{1}{2}\log K + \frac{1}{2}\log\log K}{\log K} \tag{C.74}$$

$$= \frac{1}{2} \tag{C.75}$$

□

Bibliography

[1] T.M. Cover and J.A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, USA, 2006.

[2] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Tight bounds for universal compression of large alphabets," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2875–2879.

[3] Anuran Makur, "Coding theorems for noisy permutation channels," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 672–6748, Nov 2020.

[4] V. M. Tikhomirov A. N. Kolmogorov, "ε-entropy and ε-capacity of sets in function spaces," *Uspekhi Mat. Nauk*, vol. 14, no. 2(86), pp. 3–86, 1959.

[5] D. Donoho, "Wald lecture I: Counting bits with Kolmogorov and Shannon," 2000.

[6] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *The Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[7] P.A. Chou, M. Effros, and R.M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1109–1138, 1996.

[8] Yuriy A. Reznik, "An algorithm for quantization of discrete probability distributions," in *2011 Data Compression Conference*, 2011, pp. 333–342.

[9] I. Csiszar and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[10] W. Rudin, *Principles of Mathematical Analysis*, International series in pure and applied mathematics. McGraw-Hill, 1976.

[11] Qun Xie and A.R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[12] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[13] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[14] AN Kolmogorov, "Three approaches to the definition of notion ?quantity of information?," *The Selectas*, vol. 3, 1965.

[15] B.M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Problems Inform. Transmission*, vol. 2, pp. 1–7, 1966.

[16] B.M. Fitingof, "The compression of discrete information," *Problems Inform. Transmission*, vol. 3, pp. 28–36, 1967.

[17] L. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 783–795, 1973.

[18] R. Rice and J. Plaunt, "Adaptive variable-length coding for efficient compression of spacecraft television data," *IEEE Transactions on Communication Technology*, vol. 19, no. 6, pp. 889–897, 1971.

[19] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[20] B.S. Clarke and A.R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[21] Bertrand S. Clarke and Andrew R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[22] Q. Xie and A. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[23] David Haussler and Manfred Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.

[24] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Transactions on Information Theory*, vol. 24, no. 6, pp. 674–682, 1978.

[25] A. Orlitsky and N.P. Santhanam, "Speaking of infinity [i.i.d. strings]," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2215–2230, 2004.

[26] Y. Shtarkov, T. Tjalkens, and F. Willems, "Multialphabet universal coding of memoryless sources," *Problems of Information Transmission*, vol. 31, pp. 114–127, 1995.

[27] Y. Shtarkov, "Universal sequential coding of single messages," *Probl. Peredachi Inf.*, vol. 23, pp. 3–17, 1987.

[28] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2686–2707, 2004.

[29] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for markov sources," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1393–1402, 2004.

[30] Jayadev Acharya, Hirakendu Das, and Alon Orlitsky, "Tight bounds on profile redundancy and distinguishability," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. 2012, vol. 25, Curran Associates, Inc.

[31] Ulf Grenander, "On the theory of mortality measurement," *Scandinavian Actuarial Journal*, vol. 1956, no. 2, pp. 125–153, 1956.

[32] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Tight bounds for universal compression of large alphabets," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2875–2879.

[33] A. J. Stam, "Distance between sampling with and without replacement," *Statistica Neerlandica*, vol. 32, pp. 81–91, 1978.

[34] P. Diaconis and D. Freedman, "Finite exchangeable sequences," *The Annals of Probability*, vol. 8, no. 4, pp. 745–764, 1980.

[35] John MacLaren Walsh, Steven Weber, and Ciira wa Maina, "Optimal rate delay tradeoffs for multipath routed and network coded networks," in *2008 IEEE International Symposium on Information Theory*, 2008, pp. 682–686.

[36] Mladen Kovacevic and Dejan Vukobratovic, "Subset codes for packet networks," *IEEE Communications Letters*, vol. 17, no. 4, pp. 729–732, Apr 2013.

[37] Mladen Kovacevic and Dejan Vukobratovic, "Perfect codes in the discrete simplex," *Designs, Codes and Cryptography*, vol. 75, no. 1, pp. 81–95, Nov 2013.

[38] S. M. Hossein Tabatabaei Yazdi, Han Mao Kiah, Eva Garcia-Ruiz, Jian Ma, Huimin Zhao, and Olgica Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.

[39] Yaniv Erlich and Dina Zielinski, "DNA fountain enables a robust and efficient storage architecture," *bioRxiv*, 2016.

[40] Reinhard Heckel, Ilan Shomorony, Kannan Ramchandran, and David N. C. Tse, "Fundamental limits of DNA storage systems," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 3130–3134.

[41] Mladen Kovacevic and Vincent Y. F. Tan, "Codes in the space of multisets-coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, Jul 2018.

[42] Ilan Shomorony and Reinhard Heckel, "Capacity results for the noisy shuffling channel," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 762–766.

[43] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *MIT (6.441), UIUC (ECE 563)*, 2013-2016.

[44] K. Marton, "A simple proof of the blowing-up lemma (corresp.)," *IEEE Transactions on Information Theory*, vol. 32, no. 3, pp. 445–446, 1986.

[45] Herbert Robbins, "A remark on stirling's formula," *The American Mathematical Monthly*, vol. 62, no. 1, pp. 26–29, 1955.

[46] V.V. Petrov, *Sums of Independent Random Variables*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 2. Folge. Springer Berlin Heidelberg, 2012.

[47] John Duchi, "Lecture notes for statistics 311/electrical engineering 377," *Stanford (Statistics 311/EE 377)*, 2021.

[48] Frantisek Matus, "Urns and entropies revisited," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1451–1454.

[49] Roman Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[50] Terrance Tao, "254a, notes 2: The central limit theorem," Jan 2010.

[51] D. Haussler and A. Barron, "How well do Bayes methods work for on-line prediction of ± 1 values?," in *In Proceedings of the Third NEC Symposium on Computation and Cognition. SIAM*, 1992, pp. 74–100.

[52] T.S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[53] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993?1022, January 2003.

[54] H. Steck and T. Jaakkola, "On the Dirichlet prior and Bayesian regularization," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2002, NIPS'02, p. 713?720, MIT Press.

[55] S. Schober, "Some worst-case bounds for Bayesian estimators of discrete distributions," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2194–2198.

[56] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2007.

[57] K. Varshney and L. Varshney, "Quantization of prior probabilities for hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4553–4562, 2008.

[58] Y. Zhu and J. Lafferty, "Quantized estimation of Gaussian sequence models in euclidean balls," 2014.

[59] Y. Zhu and J. Lafferty, "Quantized minimax estimation over Sobolev ellipsoids," *Information and Inference: A Journal of the IMA*, vol. 7, no. 1, pp. 31–82, 06 2017.

[60] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, vol. 14, no. 3, pp. 1080–1100, 1986.

[61] B. Clarke, "Asymptotic cumulative risk and Bayes risk under entropy loss, with applications," 1989.

[62] S.L. Fix, "Rate distortion functions for squared error distortion measures," *Proc. 16th Annu. Allerton Conf. Commun., Contr., Comput.*, Oct. 1978.

[63] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.

[64] K. Watanabe and S. Ikeda, "Rate-distortion function for gamma sources under absolute-log distortion measure," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2557–2561.

[65] T. Berger, *Rate Distortion Theory: Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[66] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2026–2031, 1994.

[67] T. Koch, "The Shannon lower bound is asymptotically tight," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6155–6161, 2016.

[68] T. Linder and R. Zamir, "High-resolution source coding for non-difference distortion measures: the rate-distortion function," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 533–547, 1999.

[69] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[70] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol. 14, no. 5, pp. 676–683, 1968.

[71] A. Buzo, F. Kuhlmann, and C. Rivera, "Rate-distortion bounds for quotient-based distortions with application to Itakura-Saito distortion measures," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 141–147, 1986.

[72] H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute-magnitude criterion," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 59–64, 1975.

[73] J. Rissanen, "Universal coding, information, prediction, and estimation," *Information Theory, IEEE Transactions on*, vol. 30, pp. 629 – 636, 08 1984.

[74] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[75] W. R. Bennett, "Spectra of quantized signals," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 446–472, 1948.

[76] P.F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proceedings of the IRE*, vol. 39, no. 1, pp. 44–48, 1951.

[77] P. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 139–149, 1982.

[78] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 373–380, 1979.

[79] Michele Lewis and SC MTSA, "A-law and mu-law companding implementations using the tms320c54x," 1997.

[80] Bernard Smith, "Instantaneous companding of quantized signals," *The Bell System Technical Journal*, vol. 36, no. 3, pp. 653–710, 1957.

[81] Noam Slonim and Naftali Tishby, "Agglomerative information bottleneck," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 1999, NIPS'99, p. 617?623, MIT Press.

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

[83] Fernando Pereira, Naftali Tishby, and Lillian Lee, "Distributional clustering of English words," in *Proceedings of the ACL*, 1993, pp. 183–190.

[84] Bin Jiang, Jian Pei, Yufei Tao, and Xuemin Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 751–763, 2013.

[85] Jie Cao, Zhiang Wu, Junjie Wu, and Wenjie Liu, "Towards information-theoretic k-means clustering for image indexing," *Signal Processing*, vol. 93, no. 7, pp. 2026–2037, 2013.

[86] Inderjit Dhillon and Subramanyam Mallela, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of machine learning research*, vol. 3, pp. 1265?1287, 04 2003.

[87] Frank Nielsen, "Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 657–660, 2013.

[88] R. Veldhuis, "The centroid of the symmetrical kullback-leibler distance," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 96–99, 2002.

[89] Carl-Gustav Esseen, "On the remainder term in the central limit theorem," *Arkiv fr Matematik*, vol. 8, no. 1, pp. 7–15, 1969.

[90] D. Rajwan and M. Feder, "Universal finite memory machines for coding binary sequences," in *Proceedings DCC 2000. Data Compression Conference*, 2000, pp. 113–122.