UNIVERSITY OF CALIFORNIA, SAN DIEGO

Estimation and Compression Over Large Alphabets

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Jayadev Acharya

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Sanjoy Dasgupta
Professor Massimo Franceschetti
Professor Alexander Vardy
Professor Jacques Verstraete
Professor Kenneth Zeger

2014

The dissertation of Jayadev Acharya is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

____
Chair

University of California, San Diego

2014

DEDICATION

To my late grandfather, Mahendra Acharya

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I am extremely fortunate to have Alon Orlitsky as my advisor. It must have been some good deeds in a past life of mine, that I got the opportunity to meet him, know him and learn from him. He has been very supportive of me through the ups and downs of my studies. He has patiently listened to my crazy thoughts, academic and non-academic. I can never thank him enough! For example, when we wrote the first paper, he stayed up whole night working with me. Without an advisor like him, I would perhaps have never been able to write this thesis.

I am thankful to my committee members Professors Sanjoy Dasgupta, Massimo Franceschetti, Alexander Vardy, Jacques Verstraete, and Kenneth Zeger for agreeing to be on my committee and providing valuable suggestions and feedback. I thank Professors Young-Han Kim and Ramamohan Paturi for showing interest in my work, and attending my defense talk.

A good portion of this thesis is a result of the collaborations I had with my lab-mates, Hirakendu Das, Ashkan Jafarpour, Shengjun Pan, and Ananda Theertha Suresh. I am fortunate to have such smart people to bounce ideas around, and learn a lot of things in the process. I thank Narayana Santhanam, and Krishnamurthy Viswanathan for their thoughts on my research and for their advice.

I thank Dr. Jaideva Goswami for being an amazing mentor, and friend throughout. I would like to thank late Ramachandra Hota, who even at 70 agreed to teach me geometry! I still do not know how he did the constructions with ruler and compass. I thank Dr. Chandra Kishore Mohapatra, and Dr. Swadhin pattanayak for their encouragement that opened many avenues for me.

I thank all the friends at UCSD and from before, for all the encouragement, and fun times. This space is clearly not adequate to capture even a fraction of them.

I thank my wife Snigdha for her immense support, encouragement, understanding, and for the sacrifices, both personal and professional, that she had to make during the course of my studies. The day before my defense, I reach home at 2am, and found her waiting to warm the food, so we can have dinner together. I thank her for all this and much more. Every evening, I feel overjoyed to receive the widest possible smile from my daughter Saanvi when she sees me return home. It acts like an instant re-energizer. Finally, I am deeply indebted to my parents for their unconditional love and support throughout my life, for teaching me that knowledge is more important than most other things. I hope that some day I am half the engineer that my father is, and some day I can prove trigonometric identities as fast as my mother. I thank my brother Sukadeb for all the fun we have had. I hope we find time soon to do fun things, playing cricket in a 5ftx10ft balcony, and video games with bollywood songs in the background.

Chapter 4 is partially adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Efficient compression of monotone and m-modal distributions", submitted to *IEEE International Symposium on Information Theory (ISIT)*, 2014.

Chapter 5 is partially adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Poissonization and universal compression of envelope classes", submitted to *IEEE International Symposium on Information Theory (ISIT)*, 2014.

Chapter 6 is partially adapted from Jayadev Acharya, Hirakendu Das, Alon Orlitsky, "Tight bounds on profile redundancy and distinguishability", *Neural Information Processing Systems (NIPS)*, 2012, and Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Tight Bounds for Universal Compression of Large Alphabets", *IEEE International Symposium on Information Theory (ISIT)*, 2013.

VITA

2007	B. Tech. in Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur
2007-2013	Graduate Student Researcher, University of California, San Diego
2009	M. S. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego
2014	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

Jayadev Acharya, Alon Orlitsky, Shengjun Pan, "The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2009.

Jayadev Acharya, Hirakendu Das, Olgica Milenkovic, Alon Orlitsky, Shengjun Pan, "Recent results on pattern maximum likelihood", *IEEE Information Theory Workshop (ITW)*, 2009.

Jayadev Acharya, Hirakendu Das, Olgica Milenkovic, Alon Orlitsky, Shengjun Pan, "String reconstruction using substring compositions", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1238-1242, 2010.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, Narayana Prasad Santhanam, "Classification using pattern maximum likelihood", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1493-1497, 2010.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, "Exact calculation of pattern probabilities", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1498-1502, 2010.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, "Competitive closeness testing", *Conference on Learning Theory (COLT)*, 47-68, 2011.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, "Algebraic computation of pattern maximum likelihood", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 400-404, 2011.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, "Estimating multisets of Bernoulli processes", *IEEE Symposium on Information Theory (ISIT)*, 2012.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh, "Competitive classification and closeness testing", *Conference on Learning Theory (COLT)*, 2012.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, "Tight bouunds on Profile redundancy and distinguishability", *Proceedings of the Neural Information Processing Systems (NIPS 2012)*, 2012.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Tight Bounds for Universal Compression of Large Alphabets", *IEEE Symposium on Information Theory (ISIT)*, 2013.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "A competitive test for uniformity of monotone distributions", *Artificial Intelligence and Statistics (AISTATS)*, 2013.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Efficient compression of monotone and m-modal sources", Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2014.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Efficient compression of monotone and m-modal distributions", Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2014.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Poissonization and universal compression of envelope classes", Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2014.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Sorting with adversarial comparators and application to density estimation", Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2014.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Sublinear algorithms for outlier detection and generalized closeness testing", Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2014.

ABSTRACT OF THE DISSERTATION

Estimation and Compression Over Large Alphabets

by

Jayadev Acharya

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California, San Diego, 2014

Professor Alon Orlitsky, Chair

Compression, estimation, and prediction are basic problems in information theory, statistics and machine learning. These problems have been extensively studied in all these fields, though the primary focus in a large portion of the work has been on understanding and solving the problems in the asymptotic regime, *i.e.,* the alphabet size is fixed and the length of observations grow. Compression of long *i.i.d.* sequences over a small alphabet has been studied extensively. Kieffer, Davisson, and others showed that there is no good scheme for compression of *i.i.d.* distributions over an infinite alphabet. With the advent of data with larger underlying alphabet size over the past decade, researchers have considered various methods/models for which efficient compression is possible.

We use redundancy, the extra number of bits beyond the optimal as the performance metric. We consider three general models to address compression of large alphabets.

The first model considers sources with only a few modes. Most natural distributions over the integers consists of only a few modes. Moreover, mixture of a few simple distributions also satisfies this property. However, even the class $\mathcal{M}$ of all monotone distributions over $\mathbb{N}$ also has infinite redundancy. In spite of this, Foster, Stine and Wyner constructed encoding schemes that have diminishing redundancy for any monotone distribution with a finite entropy. We restrict our attention to $\mathcal{M}_k$, the class of monotone distributions over k alphabets. The precise redundancy of this class of distributions is characterized by Shamir for the range $k = O(n)$, *i.e.,* for block-length at most linear in the alphabet size. We extend the characterization and in fact show that as long as the underlying alphabet size is sub-exponential in the block-length, it is possible to compress monotone distributions with diminishing per-symbol redundancy. We extend these results to distributions with a constant number of modes, whose locations are unknown.

A second elegant approach proposed by Boucheron, Garivier and Gassiat considers distributions that are bounded by an envelope. They provide characterization of the redundancy of such classes and in particular, find bounds on the redundancy of power-law and exponential envelopes. Bontemps and later Bontemps, Boucheron and Gassiat consider the class of sub-exponential envelopes and characterize its redundancy precisely. However, these methods do not work for distributions with a heavy tail, *e.g.,* the power-law distributions. Poisson sampling is a widely used method to introduce independence among the number of symbol appearances, and thereby simplifying the analysis of many algorithms. We show that for discrete distributions, the redundancy of Poisson sampled sequences is sufficient to characterize the redundancy of fixed length sequences. Using this, we provide simple bounds on the redundancy of envelope classes. We then demonstrate the efficacy of these bounds by proving tight bounds on the redundancy of power-law classes, answering an open question of Boucheron et al.

The third approach, proposed initially by Aberg, Shtarkov and Smeets, and studied extensively by Orlitsky, Santhanam, Zhang, and Shamir, consider compressing the structure (called pattern) and dictionary of the sequence separately. In particular, they show that patterns can be compressed with redundancy that grows as $O(n^{1/2})$ with the block-length n, independent of the underlying alphabet size. This problem can also be interpreted as studying the Good-Turing probability estimation problem under logarithmic loss. We develop several simple and useful tools to bound redundancy of distribution classes and use them with Poisson sampling to show that the redundancy of patterns grows as $O(n^{1/3})$.

Chapter 1

Introduction

Information theory, machine learning and statistics are closely related disciplines. One of their main intersection areas is compression redundancy, online estimation and learning, and hypothesis testing.

It is well known since Shannon proved in 1948 [1] that the number of bits required to compress a distribution P over $\mathcal{A}$ is its entropy $H(P)$. This is achieved by an encoding that uses roughly $\log(1/P(x))$ bits to encode x. However in most practical applications, the distribution generating the data is unknown. In such problem a general assumption is that the data is generated by an *unknown* distribution from a *known* class $\mathcal{P}$ of distributions, for example the collection of all *i.i.d.* distributions or all Markov distributions. This uncertainty in the underlying distribution raises the number of bits above the entropy of some distribution in the class. This is studied in *Universal compression* [2, 3, 4, 5, 6]. Any encoding corresponds to some distribution Q over the encoded symbols. Hence the increase in the expected number of bits used to encode P is $\mathbb{E}_P \log(1/Q(x)) - H(P) = D(P||Q)$, the KL divergence between P and Q. Typically one is interested in the highest increase for any distribution $P \in \mathcal{P}$, and finds the encoding that minimizes it. The resulting quantity called the *(expected) redundancy* of $\mathcal{P}$, *e.g.,* [7] is given by the following min-max expression

$$\overline{R}(\mathcal{P}) \stackrel{\text{def}}{=} \inf_Q \sup_{P \in \mathcal{P}} D(P||Q),$$

where the infimum is over all possible distributions over $\mathcal{A}$.

1.1 Log-loss prediction and redundancy

The same quantity arises in online-learning, *e.g.,* [8, Ch. 9], where the probabilities of random elements $X_1, \ldots, X_n$ are sequentially estimated. One of the most popular measures for the performance of an estimator Q is the per-symbol *log loss* $\frac{1}{n}\sum_{i=1}^{n} \log Q(X_i|X^{i-1})$. As in compression, for underlying distribution $P \in \mathcal{P}$, the expected log loss is $E_P \log 1/Q(X)$, and the *log-loss regret* is $E_P \log 1/Q(X) - H(P) = D(P||Q)$. The maximal expected regret for any distribution in $\mathcal{P}$, minimized over all estimators Q is again the KL minimax, namely, redundancy.

1.2 Packing and distinguishability

In statistics, redundancy arises in multiple hypothesis testing. Consider the largest number of distributions that can be distinguished from their observations. For example, the largest number of topics distinguishable based on text of a given length. This is a form of packing of distributions. For a class $\mathcal{P}$ of dsitributions over $\mathcal{A}$. As in [9], a sub-collection $\mathcal{S} \subseteq \mathcal{P}$ of the distributions is ϵ-*distinguishable* if there is a mapping $f : \mathcal{X} \to \mathcal{S}$ such that if X is generated by a distribution $S \in \mathcal{S}$, then $P(f(X) \neq S) \leq \epsilon$. This is a much stronger condition than the related *packing* number of $\mathcal{P}$, which is the largest number of subset of distributions in a class with all pairwise ℓ_1 distances $\geq \epsilon$.

In Chapter 2 we discuss the notations, and give an overview of the notions involved. In particular, we will consider universal compression in greater detail, and discuss the stronger notion of worst case redundancy in greater detail. We will relate the problems of compression and prediction, describe patterns and profiles, and finally discuss some mathematical results that will be used.

To prove the results on redundancy we use several basic results on redundancy. While the results themselves are not hard to prove, we believe that these are interesting, and to the best of our knowledge not been presented together before. Therefore, we discuss these general results, which will serve as tools for proving redundancy results in Chapter 3.

1.3 Large alphabets

Most of the work done in compression and prediction are in the following regime. $\mathcal{A}$ is typically sequences over an underlying alphabet $\mathcal{X}$ of a fixed (usually small) size $k = |\mathcal{X}|$, and the results are guarantees that are a function of k. However, in a wide range of applications the natural underlying alphabet could be comparable or even larger than the length of sequences/number of observations. The number of possible pixels of an image is 2^{24} ($\approx 16MP$), which is roughly the size of a typical picture from a digital camera. The number of words in an article is a few hundred, which is a miniscule fraction of the number of all possible words. Most of the results are vacuous in this regime. Consider another well known example where the underlying alphabet is unknown. We collect butterfly species in the wild and after collecting a few butterflies, the objective is to predict the distribution of butterflies, not only the species that have been observed, but also the elements that have not been observed. This framework is similar to the Good-Turing estimator where we have to assign probabilities to the elements observed and to the symbols that have not been observed.

1.4 Approaches for studying large alphabets

Two differenct approaches have been pursued to study the redundancy of such classes of distributions.

One of the methods is to consider classes of distributions with certain tail properties. For example, [10] consider the class of all monotone distributions over the integers. They design a class of efficient codes such that length n *i.i.d.* sequences from any fixed monotone distribution P can be encoded using $nH(P) + o_n(nH(P))$ bits. However, the class of monotone distributions over $\mathbb{N}$ is not universally compressible. This led to other researchers to consider classes with slightly stricter conditions. [11] considered the class of monotone distributions over a finite alphabet alphabet of size k, *i.e.*, $|\mathcal{X}| = k$. For this class of distributions they characterize the redundancy and characterize the redundancy as a function of k and n in the regime $n = \mathcal{O}(k)$. [12] consider universal compression of sequences from distributions over the integers. Among other results, they consider the two specific classes of distributions, that obey the power law, and exponential law, *i.e.,* there are constants C and α such that any distribution satisfies for any i, $P(i) \le C/i^\alpha$ for power law, and $P(i) \le Ce^{-\alpha i}$ for exponential law. They prove upper and lower bounds on the redundancy of these classes, however their bounds are not optimal, *e.g.,*, for power law envelope they mention, "We are not in a position to claim that one of the two bounds is tight, let alone which one is tight". [13] solved the exponential envelope problem by finding the growth rate up to first order terms.

We first consider the well known "Poisson sampling" technique from the balls and bins set-up to provide an alternate bounds on the redundancy. This technique yields explicit upper and lower bounds on the redundancy of any envelope class. We prove that these simple expressions are able to find the redundancy of both the power law and exponential class up to the first order term. Furthermore, our result improves the second order term of [13]. In chapter 5 we consider the class of envelope distributions and prove these results.

Another line of work initiated by [14] and studied extensively by [15]. Motivated by language modeling for speech recognition and machine learning applications this approach separates the symbols and structure of the sequence, called its *pattern*. For example, the length-9 sequence *defendant* has the pattern 123241546, which represents the ordering of the sequence, disregarding the actual symbols. For a wide range of problems such as estimating the support size, entropy, or other *symmetric properties* of distributions, the patterns are a sufficient statistic [16, 17, 18]. This notion can also be extended to multiple sequences for the problems of closeness testing and classification [19, 20, 21, 22].

[23] considered the problem of probability estimation in the celebrated Good-Turing set-up. Relating the problem of density estimation in this setting to compression of patterns of *i.i.d.* distributions they study the performance of the Good-Turing estimator and propose a variant of the original Good-Turing estimator that works well under their criterion. Their work considers a stringent metric of worst-case performance of the algorithm and they provide sub-linear

bounds on the *log-loss* of the estimator. This implies that per-symbol loss goes to zero with the sequence length. This and the fact that the result is independent of the underlying alphabet size generated much interest about this result. They also considered the results from a purely compression framework in [15] and proved similar results. More precisely, they proved that with the block length n, the worst-case redundancy of patterns is between $n^{1/3}$ and $\sqrt{n}$.

The slightly less stringent criterion of average redundancy was studied by [24] who showed an upper bound of $n^{0.4}$. We essentially solve the problem by determining the pattern redundancy up to logarithmic factors. We prove the lower bound by designing a *larger* class of distinguishable distributions based on binary codes with a given minimum distance. We obtain the upper bound by constructing a *smaller* class of distributions that form a *covering* over the patterns. We use Poisson sampling as before, to simplify the computation. In chapter 6 we consider pattern redundancy in detail. We discuss relation of the pattern redundancy to other problems concerning large alphabet learning.

Chapter 2

Preliminaries

2.1 Standard notation

Most of the symbols we use throughout the thesis are given in the following table. Beyond these, we use capital P, Q to denote distributions, $\mathcal{P}$, to denote a collection of distributions.

Table 2.1: Notation

notation	description
μ	multiplicity
τ	type
$\overline{\psi}$	pattern
$\bar{\varphi}$	profile
$\mathcal{X}$	underlying alphabet
$\mathcal{A}$	final alphabet
k	alphabet size
n	sequence lengths
R	redundancy
$D(\cdot\|\|\cdot)$	KL divergence

2.2 Universal compression, prediction and redundancy

Let $\mathcal{A}$ be a discrete alphabet. An *encoding* is a prefix-free 1-1 map from $\mathcal{A}$ to $\{0,1\}^*$. Let $l(a)$ be the length of the code of symbol a. It can be shown that evey encoding corresponds to an implied distribution Q over $\mathcal{A}$ such that the length of symbol a is approximately $\log(1/Q(a))$, *i.e.,* $Q(a) \approx 2^{-l(a)}$.

The length of a code with implied distribution Q under a distribution P is

$$\sum_{a \in \mathcal{A}} l(a)P(a) = \sum_{a \in \mathcal{A}} P(a) \log \frac{1}{Q(a)}.$$

Source coding theorem of Shannon states that the smallest length of a code with respect to a distribution P is its entropy

$$H(P) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} P(a) \log \frac{1}{P(a)},$$

achieved by a code with the same implied distribution as the underlying distribution.

In most practical applications the underlying distribution is unknown and one has to compress data from these unknown distributions. The extra number of bits to encode a when Q is used to encode P is

$$\log \frac{1}{Q(a)} - \log \frac{1}{P(a)} = \log \frac{P(a)}{Q(a)}.$$

The expected extra number of bits to encode P using a code Q is

$$D(P||Q) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)},$$

the KL-divergence between P and Q.

A natural method to deal with unknown distributions is to assume that the distribution belongs to a known class $\mathcal{P}$ of distributions and then design codes such that for any distribution in the class the extra number of bits beyond its entropy is not large. This notion is captured in the definition of (average) redundancy of the class $\mathcal{P}$ given by

$$\overline{R}(\mathcal{P}) \stackrel{\text{def}}{=} \inf_{Q} \sup_{P \in \mathcal{P}} D(P||Q), \tag{2.1}$$

where the infimum is over all possible distributions over $\mathcal{A}$, and not necessarily an element of $\mathcal{P}$. This quantity is the minimum expected extra number of bits used over all possible distributions in $\mathcal{P}$.

A stronger concept is that of worst case redundancy given by,

$$\hat{R}(\mathcal{P}) \stackrel{\text{def}}{=} \inf_{Q} \sup_{P\in\mathcal{P}} \sup_{a\in\mathcal{A}} \log \frac{P(a)}{Q(a)}, \tag{2.2}$$

the minimum extra number of bits used over the worst symbol in the worst distribution. By its definition the worst case redundancy is always larger than the average redundancy and hence is a more stringent criterion on the quality of codes. Furthermore, let $\hat{P}(a) \stackrel{\text{def}}{=} \sup_{P\in\mathcal{P}} \hat{P}(a)$, then the Shtarkov sum [25] of $\mathcal{P}$ is

$$S(\mathcal{P}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \hat{P}(a).$$

Shtarkov showed that the worst case redundancy is

$$\hat{R}(\mathcal{P}) = \log S(\mathcal{P}),$$

achieved by the *Normalized Maximum Likelihood* [9] distribution assigning probability $\hat{P}(a)/S(\mathcal{P})$ to symbol a.

2.2.1 Compression of *i.i.d.* sequences

In most of this thesis, our objective is compression/prediction of *i.i.d.* samples from unknown distributions. In the later part, we consider patterns of *i.i.d.* sequences. For a distribution P over an underlying alphabet $\mathcal{X}$, let P^n be the product distribution $P \times P \ldots \times P$ over $\mathcal{X}^n$, *i.e.,* for a sequence $x_1^n \in \mathcal{X}^n$,

$$P^n(x_1^n) = \prod_{i=1}^{n} P(x_i).$$

Let P belong to a known class $\mathcal{P}$. Let

$$\mathcal{P}^n \stackrel{\text{def}}{=} \{P^n : P \in \mathcal{P}\}$$

be the class of all distributions P^n with $P \in \mathcal{P}$. The notation restricts each element to be the product distribution of the same distribution.

Coding length-n sequences over $\mathcal{X}$ equivalently is encoding symbols from $\mathcal{A} = \mathcal{X}^n$. Such block compression can be treated as a one-shot coding when we treat each $x_1^n \in \mathcal{X}^n$ as a symbol a. As before we consider the class of all distributions (not only *i.i.d.*) Q_n over $\mathcal{X}^n$ to define

$$\overline{R}(\mathcal{P}^n) \overset{\text{def}}{=} \inf_{Q_n} \sup_{P^n} D\left(P^n || Q_n\right),$$

$$\hat{R}(\mathcal{P}^n) \overset{\text{def}}{=} \inf_{Q_n} \sup_{P^n} \sup_{x_1^n \in \mathcal{X}^n} \log \frac{P^n(x_1^n)}{Q_n(x_1^n)},$$

as the worst case redundancy, or minimax regret of $\mathcal{P}^n$. The class $\mathcal{P}$ is said to be universal if $R(\mathcal{P}) = o(n)$, *i.e.,* redundancy is sublinear in the block-length.

For $\overline{x} \in \mathcal{X}^n$, let

$$\hat{P}^n(\overline{x}) \overset{\text{def}}{=} \sup_{P^n \in \mathcal{P}^n} P^n(\overline{x}),$$

be the maximum likelihood (ML) probability of $\overline{x}$ and $\hat{P^n}$ is the ML distribution. Since we are compressing length-n sequences, $\mathcal{A} = \mathcal{X}^n$,

$$S(\mathcal{P}^n) = \sum_{\overline{x} \in \mathcal{X}^n} \hat{P}^n(\overline{x}),$$

where $\hat{P}^n(\overline{x})$ is the distribution in $\mathcal{P}^n$ that assigns the highest probability to $\overline{x}$.

The most widely studied class of distributions is $\mathcal{I}_k$, the class of distributions over k elements, *e.g.,* $[k]$. By the previous definition, $\mathcal{I}_k^n$ is the collection of *i.i.d.* distributions over sequences of length n over an alphabet $\mathcal{X}$ of size k. Under our notation the $\mathcal{A} = [k]^n$, since the we are interested in compressing length-n sequences. A succession of papers [26, 27, 28, 29, 30, 31, 32] show that for $k = o(n)$

$$\overline{R}(\mathcal{I}_k^n) + f_1(k) = \hat{R}(\mathcal{I}_k^n) + f_2(k) = \frac{k-1}{2} \log \frac{n}{k} \tag{2.3}$$

and for $n = o(k)$

$$\overline{R}(\mathcal{I}_k^n) + g_1(n) = \hat{R}(\mathcal{I}_k^n) + g_2(n) = n \log \frac{k}{n},$$

where f's are independent of n and g's independent of k.

An encoding is called *universal* if the per-symbol redundancy → 0 with n. For a given alphabet size k as length n grows there exist universal codes for $\mathcal{I}_k^n$.

2.2.2 Prediction and redundancy

Redundancy is closely related to the problem of prediction under logarithmic loss.

We describe the problem of prediction. An observer sees a sequence of observations $x_1, x_2, \ldots, x_t, \ldots$ over alphabet $\mathcal{X}$. At each time t the observer has to provide a distribution $Q_{t+1}(\cdot|x_1^t)$ over $\mathcal{X}$. If the underlying distribution was P_{t+1} the observer incurs a loss which is the KL divergence between the P_t and Q_t. When the underlying distribution is *i.i.d.* then $P_t = P$ for all t. The cumulative loss of the observer up to time n is

$$\begin{aligned} R_n &= \sum_{t=1}^{n} D(P||Q_t) \\ &= D(P^n||Q_1 \cdot Q_2 \cdots Q_n). \end{aligned}$$

The distribution $Q_1 \cdot Q_2 \cdots Q_n$ induces a distribution over length-n sequences on $\mathcal{X}$.

If the class of underlying distributions belongs to $\mathcal{P}$ then the cumulative loss of prediction to time n is the average redundancy of the class $\mathcal{P}^n$.

2.2.3 Large alphabets

However in many applications the natural alphabet that captures the data is very large, possibly infinite. For example, the building block of a language is words, not the letters in the alphabet. For example, a typical article consists of a few hundred words compared to the hundreds of thousands of words in the English dictionary. The natural symbols in a typical digital image are pixels, which can take 2^{24} distinct values.

A study of universal compression over large or arbitrary alphabets was undertaken by Kieffer [4]. He derived a necessary and sufficient condition for universality, and used it to show that *i.i.d.* distributions over infinite alphabets entail infinite redundancy. Further results in the case of large alphabets was done by [33, 32]. In these problems the alphabet size $k \gg n$, and if the underlying distribution is *i.i.d.*, then $\hat{R}(\mathcal{I}_k^n)$ is large and increases to infinity as k grows.

Faced with Kieffer's impossibility results, subsequent universal compression work has typically avoided general distributions over large alphabets. An approach to this was proposed in [14] and was subsequently studied by [15]. The method is to describe a sequence as its *pattern* and the associated *dictionary.* The details are proved in the next section on patterns, and the related concept of profiles.

2.3 Patterns

Patterns contain all the structural information about a sequence and discards the information about the individual symbols appearing in it. The *pattern* of a sequence $x_1^n \stackrel{\text{def}}{=} x_1 \ldots x_n$, denoted $\overline{\psi}(x_1^n)$ is the integer sequence obtained by replacing each symbol in x_1^n by the number of distinct symbols up to (and including) its first appearance. The definition of patterns is simple and is now explained with a few examples. The pattern of the length four sequence G O O D is 1223 since G is the first distinct symbol, O the second and D the third. $\overline{\psi}$(T O D O) = 1232 since here T, O and D are the three distinct symbols appearing in that order. When $\mathcal{X}$ is the set of all english words, then the pattern of the phrase *to be or not to be* is $\overline{\psi}(\textit{to be or not to be}) = 123412$.

A sequence can be described by encoding its pattern and the *dictionary* separately. For example, one can encode 12314151231 and then convey the dictionary as $1 \to G, 2 \to O, 3 \to D$.

We consider the distributions induced on Ψ^n when length$-n$ sequences are *i.i.d.* generated. Let P be a distribution over any $\mathcal{X}$. The probability of a sequence $\overline{x} \in \mathcal{X}^n$ under P is $P^n(\overline{x}) \stackrel{\text{def}}{=} \prod_{i=1}^n P(x_i)$, *i.e.,* the probability that the outcome is $\overline{x}$ when a length$-n$ sequence is generated independently according to P. The probability of a pattern is

$$P^n(\overline{\psi}) \stackrel{\text{def}}{=} \sum_{\overline{x}:\overline{\psi}(\overline{x})=\overline{\psi}} P^n(\overline{x}),$$

the probability of observing a sequence with pattern $\overline{\psi}$. For example, the probability of pattern 1232 under *i.i.d.* distribution P over $\mathcal{X} = \{A, B, \ldots, Z\}$ is

$$P^3(1232) = P^3(ABCB) + P^3(ABDB) + \ldots + P^3(ZYXY).$$

2.4 Profiles

Profiles serve the same role for patterns as types do for sequences. Pattern redundancy and prediction bounds seem easier to accomplish by considering profiles. Since we consider *i.i.d.* distributions, the order of the elements does not affect its probability. For example, for every distribution P, the probability of generating a pattern 112 is the same as that of 122.

The *profile* or *fingerprint* of a sequence is the multiset of multiplicities of all the symbols appearing in it [15, 19, 34]. For example, the sequences T O D O and G O O D both have one element with multiplicity 2 and two elements appearing once, therefore the profile of both of them is $\{1, 1, 2\}$.

Let Φ^n be the set of all profiles of length-n sequences. Clearly, $\Phi^1 = \{\{1\}\}$ and $\Phi^2 = \{\{2\}, \{1, 1\}\}$ since any sequence of length-2 either contains one symbol appearing twice or two symbols each appearing once. Similarly, $\Phi^2 = \{\{3\}, \{1, 2\}, \{1, 1, 1\}\}$. Using this notion it can be shown that there is a bijtection between the set of all profiles of length n and the integer partitions of n.

2.5 Mathematical preliminaries

2.5.1 Stirling's approximation

Stirling's approximation is a characterization of $n!$ as an exponent. We will use the following tight form of this approximation.

Lemma 1. *For any positive integer n,*

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\theta_n}, \ \textit{for some } \theta_n \in \left(\frac{1}{12n+1}, \frac{1}{12n}\right).$$

2.5.2 Poisson distribution and tail bounds

Let $\mathrm{poi}(\lambda)$ denote the Poisson distribution with mean λ, and let

$$\mathrm{poi}(\lambda, \mu) \stackrel{\mathrm{def}}{=} \frac{e^{-\lambda}\lambda^\mu}{\mu!},$$

is the probability of μ under $\mathrm{poi}(\lambda)$.

The KL divergence of a Poisson random variable has a closed form expression and can be bounded as follows.

Lemma 2. *Let* $\mathrm{poi}(\lambda)$ *and* $\mathrm{poi}(\lambda')$ *be Poisson random variables. Then*

$$D\Big(\mathrm{poi}(\lambda)||\mathrm{poi}(\lambda')\Big) = \lambda' - \lambda + \lambda \log \frac{\lambda}{\lambda'} \le \frac{(\lambda-\lambda')^2}{\lambda'}.$$

Proof. By the definition of Poisson distribution, $\mathrm{poi}(\lambda,\mu) = e^{-\lambda}\lambda^{\mu}/\mu!$,

$$\begin{aligned} D\Big(\mathrm{poi}(\lambda)||\mathrm{poi}(\lambda')\Big) = \sum_{\mu=0}^{\infty} \mathrm{poi}(\lambda,\mu) \log \frac{\mathrm{poi}(\lambda,\mu)}{\mathrm{poi}(\lambda',\mu)} &= \sum_{\mu=0}^{\infty} e^{-\lambda}\frac{\lambda^{\mu}}{\mu!} \log\left(\frac{e^{-\lambda}\lambda^{\mu}}{e^{-\lambda'}\lambda'^{\mu}}\right) \\ &= \sum_{\mu=0}^{\infty} e^{-\lambda}\frac{\lambda^{\mu}}{\mu!}\left((\lambda'-\lambda) + \mu\log\left(\frac{\lambda}{\lambda}\right)\right. \\ &= \lambda' - \lambda + \lambda\log\left(\frac{\lambda}{\lambda'}\right) \\ &\overset{(a)}{\le} \lambda' - \lambda + \lambda\Big(\frac{\lambda}{\lambda'} - 1\Big) \\ &= \frac{(\lambda-\lambda')^2}{\lambda'}, \end{aligned}$$

where (a) uses $\log(x) \le x - 1$. □

Using the Chernoff bounds, we show strong concentration of Poisson random variables around its mean. We first look at the moment generating function of a $\mathrm{poi}(\lambda)$ random variable.

Lemma 3. *Let* $X \sim \mathrm{poi}(\lambda)$*,*

$$\mathbb{E}[e^{tX}] = e^{\lambda(e^t-1)}.$$

Proof. By definition,

$$\mathbb{E}[e^{tX}] = \sum_{\mu\ge 0} \mathrm{poi}(\lambda,\mu)e^{t\mu} = \frac{e^{-\lambda}(\lambda e^t)^{\mu}}{\mu!} = e^{\lambda(e^t-1)}.$$

□

Using this we prove the following tail bounds on Poisson random variables.

Lemma 4. *([35]) Let* $X \sim \mathrm{poi}(\lambda)$*, then,*

1. *For* $x \geq \lambda$*,*

$$\Pr(X \geq x) \leq \exp(-\lambda)\left(\frac{e\lambda}{x}\right)^{x} \leq \exp\left(\frac{(x-\lambda)^2}{2x}\right),$$

2. *and for* $x \leq \lambda$*,*

$$\Pr(X \leq x) \leq \exp(-\lambda)\left(\frac{e\lambda}{x}\right)^{x} \leq \exp\left(\frac{(x-\lambda)^2}{2\lambda}\right).$$

Proof. For any $t \geq 0$ and $x \geq \lambda$,

$$\Pr(X \geq x) = \Pr(e^{tX} > e^{tx}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tx}}.$$

Substituting the moment generating function and plugging $t = \ln(x/\lambda)$ yields the first inequality in the first item. The second inequality is simple calculus. The second item can be proved similarly. □

2.5.3 Binary codes

A *binary code* $\mathcal{C}(k, d)$ of length k and minimum distance d is a collection of length-k binary strings (codewords) such that the Hamming distance between any two codewords is at least d. The size of the code, denoted $|\mathcal{C}(k, d)|$ is the number of codewords in it. For large lengths, when the minimum distance is stipulated to be a constant ($\alpha < 1/2$) fraction of the length then there exist codes with size exponential in the length. This is shown by the following Gilbert-Varshamov bound.

Lemma 5 ([36])**.** *Let* $0 < \alpha < 1/2$*. There exists* $\mathcal{C}(k, \alpha d)$ *with*

$$|\mathcal{C}(k, d)| \geq 2^{k(1-h(\alpha)-o(1))}.$$

Proof. A simple volume argument (Gilbert-Varshamov bound) shows that there exists $\mathcal{C}(k, d)$ for $d \leq k/2$ with

$$|\mathcal{C}(k, d)| \geq \frac{2^k}{d\binom{k}{d}}.$$

By Stirling's approximation, $\binom{k}{\alpha k} \leq \frac{2^{kh(\alpha)}}{\sqrt{2\pi m\alpha(1-\alpha)}}$. Plugging $d = \alpha k$, we obtain

$$|\mathcal{C}(k, \alpha k)| \geq 2^{k(1-h(\alpha)-\frac{\log k}{k})}.$$

□

2.6 Poissonization

Throughout this thesis, we would be interested in sampling distributions independently n times for some large n. We then consider the sequence, or a function of the sequence that is observed (in the case of patterns and profiles considered in Chapter 6).

Let P be a distribution over a discrete alphabet $\mathcal{X}$. and recall that P^n is the product distribution over $\mathcal{X}^n$ obtained by *i.i.d.* sampling of P is sampled n times.

When P is sampled n times then the number of appearances of different symbols are dependent (*e.g.,* they sum to n). A classic technique that removes this dependency and simplifies many arguments/computations is Poisson sampling. Instead of sampling the distribution n times consider the alternate two step process:

- Generate $n' \sim \text{poi}(n)$, the Poisson distribution with mean n.
- Sample P independently n' times to obtain a sequence in $\mathcal{X}^{n'}$.

This process defines a distribution $P^{\text{poi}(n)}$ over $\mathcal{X}^*$ where the probability of a sequence $x_1^{n'} \in \mathcal{X}^*$ is

$$P^{\text{poi}(n)}(x_1^{n'}) = \text{poi}(n, n')P^{n'}(x_1^{n'}) = \text{poi}(n, n')\prod_{i=1}^{n'} P(x_i). \tag{2.4}$$

Similar to $\mathcal{P}^n$, let

$$\mathcal{P}^{\text{poi}(n)} \stackrel{\text{def}}{=} \{P^{\text{poi}(n)} : P \in \mathcal{P}\}$$

be the class of distributions over $\mathcal{X}^*$ via sampling a distribution *i.i.d.* $\text{poi}(n)$ times.

Note that even though the length of sequence generated by $P^{\text{poi}(n)}$ is random, it is concentrated arount n as a Poisson random variable. For large n, the length n' is concentrated around the mean n with standard deviation $\sqrt{n}$. Some useful properties of Poissonization are described in the following lemma.

Lemma 6 ([35]). *Let P be a distribution sampled independently $n' \sim \text{poi}(n)$ times.*

1. *Conditioned on n', the distribution induced on $\mathcal{X}^{n'}$ is $P^{n'}$.*

2. *A symbol $x \in \mathcal{X}$ with $P(x) = p$ appears* $\text{poi}(np)$ *times* independently *of all other symbols.*

Proof. The first statement follows from the method of sampling. For notational ease suppose the distribution P is over $a_1, a_2, \ldots$ with probabilities $p(a_1), p(a_2), \ldots$ respectively. Let N_j be the number of appearances (multiplicity) of a_j in a sample $\sim P^{\text{poi}(n)}$. Then,

$$\begin{aligned}
\Pr(N_j = \mu) &= \sum_{n'=\mu}^{\infty} \text{poi}(n,n') \binom{n'}{\mu} p(a_j)^{\mu}(1-p(a_j))^{n'-\mu}\\
&= \sum_{n'=\mu}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\frac{n'!}{\mu!(n'-\mu)!}p(a_j)^{\mu}(1-p(a_j))^{n'-\mu}\\
&= \frac{e^{-n}(np(a_j))^{\mu}}{\mu!}\sum_{n'=\mu}^{\infty}\frac{(n(1-p(a_j)))^{n'-\mu}}{(n'-j)!}\\
&= \frac{e^{-n}(np(a_j))^{\mu}}{\mu!}e^{(n(1-p(a_j)))}\\
&= \text{poi}(np(a_j),\mu).
\end{aligned}$$

This proves that number of appearances are Poisson distributed. To prove independence, we find $Prob(N_1 = \mu_1, N_2 = \mu_2, \ldots)$. By part 1, conditioned on number of samples n', this corresponds to $P^{n'}$. Using the multinomial theorem,

$$\begin{aligned}
&\Pr(N_1 = \mu_1, N_2 = \mu_2, \ldots)\\
=&\Pr(n' = \mu_1 + \mu_2 + \ldots) \cdot Prob(N_1 = \mu_1, N_2 = \mu_2, \ldots | n' = \mu_1 + \mu_2 + \ldots)\\
=&\text{poi}(n,n') \cdot \binom{n'}{\mu_1, \mu_2, \ldots} \prod_{j \ge 1} p(a_j)^{\mu_j}\\
\overset{(a)}{=}&e^{-n(p(a_1)+p(a_2)+\ldots)}\frac{n^{\mu_1+\mu_2+\ldots}}{n'!}\frac{n'!}{\prod_{j\ge1}\mu_j!}\prod_{j\ge1}p(a_j)^{\mu_j}\\
=&\prod_{j\ge1}e^{-np(a_j)}\frac{(np(a_j))^{\mu_j}}{\mu_j!}\\
=&\prod_{j\ge1}Prob(N_j = \mu_j),
\end{aligned}$$

where (a) uses $p(1) + p(2) + \ldots = 1$. This proves independence of the number of multiplicities of symbols. □

We now provide a relation between the Shtarkov sums of Poisson-sampling and sampling n times.

Lemma 7. *For a class $\mathcal{P}$,*

$$S\left(\mathcal{P}^{\text{poi}(n)}\right) = \sum_{n'\geq 0} \text{poi}(n,n') S\left(\mathcal{P}^{n'}\right)$$

Proof. By the first item of the previous lemma or Equation 2.4, the Shtarkov sum conditioned on the length does not change, namely for a sequence $x_1^{n'}$ the same distribution attains maximum likelihood for both poi(n) sampling and sampling *i.i.d.* n' times, for any n. Therefore,

$$S\left(\mathcal{P}^{\text{poi}(n)}\right) = \sum_{n'\geq 0} \text{poi}(n,n') \sum_{x_1^{n'}} \sup_{P\in\mathcal{P}} P^{n'}(x_1^{n'}) = \sum_{n'\geq 0} \text{poi}(n,n') S\left(\mathcal{P}^{n'}\right),$$

where in the first step we sum maximum likelihoods of sequences by their lengths. □

We will use these results to provide simplified expressions and analyses. In Chapter 5 we obtain simple bounds on the redundancy, and in Chapter 6 we bound the redundancy of patterns using Poissonization.

Chapter 3

Basic results on redundancy

Let $\mathcal{P}$ be a class of distributions over $\mathcal{A}$. We now state a few preliminary results on redundancy that are used to prove the results in this thesis.

First in Lemma 9 we show a lower bound on the average redudancy in terms of the number of distributions in the class that form a packing in ℓ_1 distance. Such a result also shows lower bound on the worst case redundancy. We then prove three general results for both average and worst case redundancy. For any function from $\mathcal{A}$, $\mathcal{P}$ induces a class of distributions on the image space. In Lemma 11, we show that the induced class has smaller redundancy. In the average case, this results is similar to a data processing inequality. It is reasonable that a class consisting of similar distributions has a small redundancy. Our approach would be to divide the class of distributions on profiles into classes such that each class has similar distributions. Lemma 17 bounds the redundancy of a class in terms of the individual classes and the number of classes. Using Lemma 11, we show that the redundancy of profiles is upper bounded by the redundancy of a class consisting of product distributions. Lemma 14 relates the redundancy of a class of product distributions to the redundancy of the class of marginal distributions over the individual spaces.

3.1 A redundancy capacity lower bound on redundancy

One of the standard techniques to lower bound the average redundancy is to relate it to the capacity of a specific channel and using the Redundancy-Capacity theorem [7, Chapter 13]. Consider the discrete memoryless channel where the inputs of the channel correspond to the different possible distributions of the source (i.e., all elements in P). The output of the channel for a given input distribution is simply a sample from the distribution. Let $C(\mathcal{P})$ be the capacity of this channel. Then,

Theorem 8 (Redundancy Capacity Theorem)**.** *For any $\mathcal{P}$,*

$$\overline{R}(\mathcal{P}) = C(\mathcal{P}).$$

However, for some problems computing the capacity may not be easy. Therefore instead of finding by the capacity, which corresponds to block length ∞, we lower bound redundancy by the number of bits one can transmit with block length 1 and error ϵ.

A sub-collection $\mathcal{S} \subseteq \mathcal{P}$ of the distributions is $\delta-$*distinguishable* if there is a mapping $f : \mathcal{A} \to \mathcal{S}$, such that if X is distributed according to $P \in \mathcal{S}$, then $\Pr(f(X) \neq P) \leq \delta$. An equivalent criterion for $\delta-$distinguishablity of $\mathcal{S}$ is as follows. There exists a partition of $\mathcal{A}$ into $|\mathcal{S}|$ classes $\{\mathcal{A}_P : P \in \mathcal{S}\}$, such that for any $P \in \mathcal{S}$,

$$P(\mathcal{A}_P) \geq 1 - \delta.$$

Let $M(\mathcal{P}, \delta)$ be the largest number of $\delta-$distinguishable distributions in $\mathcal{P}$, *i.e.,*

$$M(\mathcal{P}, \delta) = \max_M \{\exists P_i \in \mathcal{P}, \mathcal{A}_i \in \mathcal{A} \text{ for } 1 \leq i \leq M : \mathcal{A}_i \cap \mathcal{A}_j = \emptyset, P_i(\mathcal{A}_i) \geq 1 - \delta\},$$

which is the maximum number of input distributions from $\mathcal{P}$ that can be transmitted over the (sampling) channel used once with error probability $\leq \delta$.

Lemma 9. $\overline{R}(\mathcal{P}) + 1 \geq (1 - \delta) \log [M(\mathcal{P}, \delta)]$.

Proof. Let Q be any distribution over $\mathcal{A}$. Since $\mathcal{A}_i$ are disjoint, there is an i, such that $Q(\mathcal{A}_i) \le 1/M$. Then

$$\begin{aligned}\sup_{P\in\mathcal{P}} D(P||Q) \ge& D(P_i||Q)\\ =&\sum_{a\in\mathcal{A}_i} P_i(a)\log\frac{P_i(a)}{Q(a)} + \sum_{a\in\overline{\mathcal{A}_i}} P_i(a)\log\frac{P_i(a)}{Q(a)}\\ \overset{(a)}{\ge}& P_i(\mathcal{A}_i)\log\frac{P_i(\mathcal{A}_i)}{Q(\mathcal{A}_i)} + P_i(\overline{\mathcal{A}_i})\log\frac{P_i(\overline{\mathcal{A}_i})}{Q(\overline{\mathcal{A}_i})}\\ \overset{(b)}{\ge}&(1-\delta)\log\frac{1-\delta}{1/M} + \delta\log\frac{\delta}{1-1/M}\\ \ge&(1-\delta)\log M - h(\delta),\end{aligned}$$

where the (a) follows from the log-sum inequality (which follows from the concavity of logarithms), (b) uses $P_i(\mathcal{A}_i) \ge 1-\delta$. $h(\cdot)$ is the Shannon entropy, thus bounded by 1. Taking Q to be the distribution that achieves average redundancy of $\mathcal{P}$ (Equation (2.1)) proves the Lemma. □

3.2 Redundancy of functions

Let $f : \mathcal{A} \to \mathcal{B}$ be a function. For $P \in \mathcal{P}$, let $f(P)$ be the distribution induced over $\mathcal{B}$ by P via f. In other words, the probability assigned to $b \in \mathcal{B}$ is $\sum_{f(a)=b} P(a)$. Let $f(\mathcal{P}) = \{f(P) : P \in \mathcal{P}\}$. The following Lemma shows that the number of distinguishable distributions in $f(\mathcal{P})$ cannot be larger than that of the original class.

Lemma 10 (Distinguishability of functions)**.** *For any* $\delta > 0$

$$M(\mathcal{P},\delta) \ge M(f(\mathcal{P}),\delta).$$

Proof. Consider the set of distributions in $f(\mathcal{P})$ that achieve $M(\mathcal{P},\delta)$. The images of these distributions form a class of δ-distinguishable distribtuions on $\mathcal{P}$ proving the lemma. □

The next result is a variation of states that the redundancy of $f(\mathcal{P})$ is at most the redundancy of $\mathcal{P}$. In other words, compressing functions random variables can be achieved with fewer loss than the original random variables. The result is a variation of the data processing ineuquality for average redundancy.

Lemma 11 (Redundancy of functions)**.** *$\overline{R}(f(\mathcal{P})) \le \overline{R}(\mathcal{P})$, and $\hat{R}(f(\mathcal{P})) \le \hat{R}(\mathcal{P})$.*

Proof. For any $b \in \mathcal{B}$,

$$\hat{P}(b) = \sup_{P\in\mathcal{P}} f(P)(b) = \sup_{P\in\mathcal{P}} \sum_{f(a)=y} P(a) \le \sum_{f(a)=y} \sup_{P\in\mathcal{P}} P(a) = \sum_{f(a)=b} \hat{P}(a).$$

$$S(\mathcal{P}) = \sum_{a\in\mathcal{A}} \hat{P}(a) = \sum_{b\in\mathcal{B}} \sum_{f(a)=b} \hat{P}(a) \ge \sum_{b\in\mathcal{B}} \sup_{f(P)\in f(\mathcal{P})} P(b) = S(f(\mathcal{P})).$$

Taking logarithm yields the result. □

While this result holds in general we would be interested in functions that preserve redundancy of classes since the alternate classes could allow for simpler characterization and computation of redundancy. Let $f : \mathcal{A} \to \mathcal{B}$ be a function such that if $f(a) = f(a')$ then all distributions in $\mathcal{P}$ assign the same probability to these symbols, *i.e.,*

$$f(a) = f(a') \Rightarrow P(a) = P(a')\forall P \in \mathcal{P}. \tag{3.1}$$

Lemma 12. *If f satisfies* (3.1)*, then $\overline{R}(f(\mathcal{P})) = \overline{R}(\mathcal{P})$, and $\hat{R}(f(\mathcal{P})) = \hat{R}(\mathcal{P})$.*

Proof. Consider all symbols that map to the same element. They have the same maximum likelihood distribution $\hat{P}$. By Equation (3.1) $f(\hat{P})$ assigns the largest probability to the image of these symbols, proving the lemma. □

Using this we now show that the redundancy of *i.i.d.* distributions is the same as the redundancy of the *types*.

3.2.1 Redundancy of types

The *type* of a sequence $\overline{x}$ over $\mathcal{X} = \{a_1, \ldots, \}$ is

$$\tau(\overline{x}) \overset{\text{def}}{=} (\mu(a_1), \mu(a_2), \ldots),$$

the tuple of multiplicities of the symbols in the sequence $\overline{x}$. For example, if $a_i = i$, for $i = 1, \ldots, 6$ denotes the possible outcomes of a die. Then the sequence of outcomes $2, 3, 1, 6, 1, 3, 3, 4, 6$ has type $\tau = (2, 1, 3, 1, 0, 2)$. For *i.i.d.* sampling the types are sufficient statistic of the sequence, namely all sequences with the same type have the same probability. Let $\tau(P^n)$ be the distribution induced on *types* by P^n. Let

$$\tau(\mathcal{P}^n) \stackrel{\text{def}}{=} \{\tau(P^n) : P \in \mathcal{P}\}$$

be all distributions over types from distributions of the form P^n. Similarly, let

$$\tau(\mathcal{P}^{\text{poi}(n)}) \stackrel{\text{def}}{=} \{\tau(P^{\text{poi}(n)}) : P \in \mathcal{P}\}.$$

Since type is a sufficient statistic, as a function of the sequence it satisfies Equation 3.1, we obtain the following result.

Lemma 13. *For $R \in \{\overline{R}, \hat{R}\}$,*

$$R(\mathcal{P}^n) = R\left(\tau(\mathcal{P}^n)\right) \ \text{ and } \ R(\mathcal{P}^{\text{poi}(n)}) = R\left(\tau(\mathcal{P}^{\text{poi}(n)})\right).$$

Proof. We show the result for $\hat{R}$ due to simplicity. For a sequence in $\mathcal{X}^n$, let

$$\hat{P}^n(\overline{x}) = \sum_{P \in \mathcal{P}} P^n(\overline{x}).$$

All sequences with the same type are assigned the same probability, therefore the maximum likelihood distribution is same for types and sequences.

$$S(\mathcal{P}^n) = \sum_{\overline{x} \in \mathcal{X}^n} \hat{P}^n(\overline{x}) = \sum_{\tau} \sum_{\overline{x}:\tau(\overline{x})=\tau} \hat{P}^n(\overline{x}) = \sum_{\tau} \hat{P}^n(\tau) = S(\tau(\mathcal{P}^n)).$$

Taking logarithms proves the result. For Poisson-sampling, by Lemma 7,

$$\begin{aligned} S\left(\mathcal{P}^{\text{poi}(n)}\right) =& \sum_{n' \geq 0} \text{poi}(n, n') S\left(\mathcal{P}^{n'}\right) \\ =& \sum_{n' \geq 0} \text{poi}(n, n') S\left(\tau(\mathcal{P}^{n'})\right) \\ =& S\left(\tau(\mathcal{P}^{\text{poi}(n)})\right), \end{aligned}$$

where in the last step we use the definition of Poisson sampling. □

3.3 Redundancy of product distributions

In many applications the random variable considered is a tuple or collection of random variables,*e.g.,* the type of a length-n sequence is the collection of multiplicities of all symbols. In such problems it may be easier to study the individual random variables. While it may not be always possible, in the special case when the random variables are independent we can relate the redundancy of the class to the redundancy of the classes introduced by the marginals. More formally, for a class $\mathcal{P}$ consisting of product (independent) distributions over $\mathcal{A} \times \mathcal{B}$, *i.e.,* each element in $\mathcal{P}$ is a distribution of the form $P \times Q$, where P and Q are distributions over $\mathcal{A}$ and $\mathcal{B}$ respectively. Let $\mathcal{P}_{\mathcal{A}}$ and $\mathcal{P}_{\mathcal{B}}$ be the class of marginals. The redundancy of $\mathcal{P}$ is at most the sum of the marginal redundancies.

Lemma 14 (Redundancy of products)**.** *For a collection $\mathcal{P}$ of product distributions over $\mathcal{A} \times \mathcal{B}$,*

$$\overline{R}(\mathcal{P}) \le \overline{R}(\mathcal{P}_{\mathcal{A}}) + \overline{R}(\mathcal{P}_{\mathcal{B}}), \text{ and } \hat{R}(\mathcal{P}) \le \hat{R}(\mathcal{P}_{\mathcal{A}}) + \hat{R}(\mathcal{P}_{\mathcal{B}}).$$

Proof. For any $a \times b \in \mathcal{A} \times \mathcal{B}$,

$$\sup_{(P,Q)\in\mathcal{P}} P(a)Q(b) \le \sup_{P\in\mathcal{P}_{\mathcal{A}}} P(a) \sup_{Q\in\mathcal{P}_{\mathcal{B}}} Q(b).$$

Now,

$$\begin{aligned} S(\mathcal{P}) &= \sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \sup_{(P,Q)\in\mathcal{P}} P(a)Q(b) \\ &\le \sum_{a\in\mathcal{A}} \sup_{P\in\mathcal{P}_{\mathcal{A}}} P(a) \sum_{b\in\mathcal{B}} \sup_{Q\in\mathcal{P}_{\mathcal{B}}} Q(b) \le S(\mathcal{P}_{\mathcal{A}})S(\mathcal{P}_{\mathcal{B}}), \end{aligned}$$

where the inequality follows from the equation above, and the lemma follows by taking logarithms. □

Corollary 15. *Suppose $\mathcal{P} = \mathcal{P}_{\mathcal{A}} \times \mathcal{P}_{\mathcal{B}}$, i.e., all marginals are possible, then*

$$\overline{R}(\mathcal{P}) = \overline{R}(\mathcal{P}_{\mathcal{A}}) + \overline{R}(\mathcal{P}_{\mathcal{B}}), \text{ and } \hat{R}(\mathcal{P}) = \hat{R}(\mathcal{P}_{\mathcal{A}}) + \hat{R}(\mathcal{P}_{\mathcal{B}}).$$

Proof. The inequality in the Lemma's proof becomes an equality. □

The independence requirement is crucial as follows from the following example.

Example 16. *Let $\mathcal{P} = \{P_1, P_2\}$ consist of two distributions over $\{0,1\} \times \{0,1\}$.*

$$P_1((0,0)) = P_1((1,1)) = \frac{1}{2}, \quad P_1((0,1)) = P_1((1,0)) = 0,$$
$$P_2((0,0)) = P_2((1,1)) = 0, \quad P_2((0,1)) = P_2((1,0)) = \frac{1}{2}.$$

The Shtarkov sum of $\mathcal{P}$ is

$$S(\mathcal{P}) = 4 \times \frac{1}{2} = 2 \Rightarrow \hat{R}(\mathcal{P}) = 1.$$

Note that both the distributions have the same marginal distribution. Clearly, a class consisting a single distribution has zero redundancy (Shtarkov sum equals 1). Therefore the sum of redundancies equals 0. This shows that all random variables need not satisfy the lemma.

3.4 Redundancy of unions

Suppose we can partition $\mathcal{P}$ into T collections of distributions such that the distributions within each collection are *close* in KL divergence, implying that the redundancy of each collection is small. In such cases it may be easy to bound the redundancy of the collections individually. We now show that an upper bound on the redundancies of the T collections yields an upper bound on the whole class. For classes $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_T$ of distributions, let $\mathcal{P} = \cup_{i=1}^{T} \mathcal{P}_i$.

Lemma 17 (Redundancy of unions)**.**

$$\overline{R}(\bigcup_{1\le i\le k} \mathcal{P}_i) \le \max_{1\le i\le T} \overline{R}(\mathcal{P}_i) + \log T, \text{ and } \hat{R}(\bigcup_{1\le i\le k} \mathcal{P}_i) \le \max_{1\le i\le T} \hat{R}(\mathcal{P}_i) + \log T.$$

Proof. For any $a \in \mathcal{A}$,

$$\sup_{P\in\mathcal{P}} P(a) \le \sum_{i=1}^{T} \sup_{P\in\mathcal{P}_i} P(a).$$

Summing over all x,

$$\begin{aligned} S(\mathcal{P}) = \sum_{a \in \mathcal{A}} \sup_{P \in \mathcal{P}} P(a) &\leq \sum_{a \in \mathcal{A}} \sum_{i=1}^{T} \sup_{P \in \mathcal{P}_i} P(a) \\ &= \sum_{i=1}^{T} \sum_{a \in \mathcal{A}} \sup_{P \in \mathcal{P}_i} P(a) \\ &= \sum_{i=1}^{T} S(\mathcal{P}_i) \\ &\leq T \cdot \max_{1 \leq i \leq T} S(\mathcal{P}_i). \end{aligned}$$

Taking logarithm, the results follows. □

Chapter 4

Efficient compression of distributions with a few modes

Recall that the redundancy of *i.i.d.* distributions over an infinite alphabet, *e.g.,* $\mathbb{N}$ is infinite. In this chapter, we consider the sub-class of distributions that have only a few modes. Monotone and unimodal distributions are a special case of such distributions with zero and one mode respectively.

We now define monotone and m-modal distributions.

Definition 18. *A distribution P over $\mathbb{N}$ is monotone (decreasing) if $P(i) \geq P(i+1)$ for all i.*

Let $\mathcal{M}$ be the class of all monotone distributions over $\mathbb{N}$, and $\mathcal{M}_k$ be the class of all monotone distributions over $[k]$. Let

$$\mathcal{M}_k^n \overset{\text{def}}{=} \{P^n : P \in \mathcal{M}_k\}$$

denote the distributions obtained by sampling distribution in $\mathcal{M}_k$ *i.i.d.* n times. An interval $[l_1, l_2] \overset{\text{def}}{=} \{l_1, l_1+1, \ldots, l_2\}$ is called a mode if, for all $i, j \in [l_1, l_2]$ $P(i) = P(j)$ and $(P(l_1-1) - P(l_1))(P(l_2+1) - P(l_2)) > 0$. Such sets denote the *ups and downs* of a distribution.

Definition 19. *A distribution is m-modal if it has at most m modes.*

Let $\mathcal{M}_{k,m}$ be the collection of all m-modal distributions over $[k]$, and

$$\mathcal{M}_{k,m}^n \overset{\text{def}}{=} \{P^n : P \in \mathcal{M}_{k,m}\}$$

be all distributions over $[k]^n$ obtained by sampling a distribution in $\mathcal{M}_{k,m}$ *i.i.d.* n times.

Most natural distributions have only a few modes. For example, the life expectancy of a population can be expected to be unimodal. Poisson and Binomial distributions are both unimodal. Mixture distributions have received attention in various communities over the past decade, since many real world phenomenon can be modeled as mixtures of simple distributions. Such mixtures of m simple distributions are typically m-modal. There has been enormous work in the past decade in learning mixtures of distributions.

Mixture distributions were initially studied by Pearson [37], who after observing measurements (say diameter over heights) of Naples crab population, postulated that the data was best explained as a mixture of two Gaussians, predicting the presence of more than one specie. However, most of the theoretical results beyond using EM algorithm for learning mixture distributions are relatively new.

Monotone distributions are a special case of m-modal distributions with $m = 0$. We first study monotone distributions and then apply the results to prove redundancy bounds on m-modal distributions. Monotone distributions are extremely interesting on their own. For example [38], the probability of a person being affected by an epidemic decreases with the distance from the center. Monotone distributions are those where we have prior knowledge about the relative probabilities of symbols, though we may not know the exact probabilities. In a text document we have some knowledge about word frequencies and probabilities. In such language modeling applications, Zipf distributions are common [39]. Geometric distributions over integers are useful in compressing residual signals in image compression [40].

4.1 Related Work

Codes for monotone distributions were studied initially by [41, 42, 43, 44]. They consider *per-symbol* codes, *i.e.,* codes that compress one symbol at a time. Their motivation was to design good codes for $\mathbb{N}$. If such codes are used to code a sequence of random variables, the redundancy is linear. [44] show that for the class $\mathcal{M}_k$,

$$\overline{R}(\mathcal{M}_k) = \log\left(1 + \sum_{i=2}^{k}\left(1 - \frac{1}{i}\right)^i \frac{1}{i-1}\right).$$

Since $(1 - 1/i)^i \to e^{-1}$ and $\sum_{i=1}^k 1/i \sim \log k$,

$$\overline{R}(\mathcal{M}_k) \sim \log\log k.$$

Elias codes [41] that assign codewords of length $\log i + 2\log\log i$ to the symbol i are therefore nearly optimal. This result also shows that the redundancy of $\mathcal{M}$, and therefore $\mathcal{M}^n$ is infinite.

Rissanen [42] considred compressing random variables from $\mathcal{M}_k$ with the goal of finding the following min-max quantity

$$\overline{r}(\mathcal{P}) = \inf_Q \sup_P \frac{\overline{R}(P,Q)}{H(P)}.$$

They come up with the optimal Q, as the solution to a convex optimization problem.

In a more recent work [45] consider the problem of designing codes for monotone distributions that minimize the *min-ave redundancy*. The min-ave redundancy of a code is the average value of its redundancy over a distribution *chosen at random* from the class (not the worst ditsribution as we usually consider). They show that the min-ave redundancy is constant for monotone distributions, as opposed to the average redundancy of $\log\log k$.

These papers consider codes for compressing a *single* symbol generated by a monotone distribution over $\mathbb{N}$, *i.e.,* they consider the class $\mathcal{M}$ or $\mathcal{M}_k$. We are interested in encoding sequences of random variables generated from monotone distributions, namely, the class $\mathcal{M}_k^n$. This follows a more "classical" flavor, where the distribution is fixed and a block of symbols is observed. [10] was the first work to consider the compression of distributions in $\mathcal{M}^n$. They show that even while the redundancy of this class is infinite, the set of all distributions in $\mathcal{M}$ with a finite entropy can be compressed with a diminishing relative redundancy. In particular, they prove the following result.

Theorem 20 ([10])**.** *There exists a code Q over $\mathbb{N}^n$ such that for any $P \in \mathcal{M}$ with $H(P) < \infty$*

$$\overline{R}(P^n, Q) \le nH(P)\frac{\log\log(nH(P))}{\log(nH(P))}.$$

Remark We note that the theorem holds for any distribution with a finite entropy, even with infinite support. In particular, if we let H grow as a polymonial of n, then the code is no longer universal. For example, a uniform distribution over $[\exp(\sqrt{n})]$ elements has the redundancy guarantee of at most $n^{3/2}$.

The works of Shamir, and in particular [11] is perhaps closest to our work. They consider compression of $\mathcal{M}_k^n$ as a function of both the alphabet size k and block-length n. Interestingly, they show that up to alphabet size $n^{1/3}$, the block-redundancy grows linearly with k, and behaves similar to the redundancy of the class $\mathcal{I}_k^n$. However, for the range of $O(n^{1/3})$ to $O(n)$, they show that the redundancy grows as *almost like* $n^{1/3}$ up to logarithmic factors.

They consider compression of monotone distributions restricted to $[k]$ and varying the block length n. They show tight lower and upper bounds on $\overline{R}(\mathcal{M}_k^n)$ for all $k = O(n)$. In particular, for $k = O(n)$,

$$\overline{R}(\mathcal{M}_k^n) = \tilde{O}(n^{1/3}).$$

This shows among other things that monotone distributions have diminishing per-symbol redundancy when the block length grows linearly with the alphabet size. These results can be extended by using methods from [46, 47] to show that $k = 2^{o(\sqrt{n})}$ the $\mathcal{M}_k^n$ is universally compressible.

The following result summarizes their results for various ranges of k. For clean representation we drop the constants and lower order terms from the expressions.

Monotone distributions have also been studied in statistics and theoretical computer science. Birgé [48] consider the related problem of learning monotone and unimodal distributions with least number of samples. The sample complexity of learning m-modal distributions over an alphabet of size k was considered more recently by [49]. Testing distributions for monotonicity has been considered in a variety of settings [50, 51, 52, 53].

Table 4.1: Known bounds on $\overline{R}(\mathcal{M}_k^n)$

Range	Lower bound	Upper bound
$k = o(n^{1/3})$	$\Omega(k \log \frac{n}{k})$	$O(k \log \frac{n}{k})$
$k = n^{O(1)}$	$\Omega(n^{1/3})$	$O(n^{1/3} \log n)$

4.2 Results

In this work we significantly extend the alphabet size for which $\mathcal{M}_k^n$ has diminishing per-symbol redundancy. We show that as long as alphabet size is sub-exponential in the block-length (or block length is super-logarithmic in the alphabet size), $\mathcal{M}_k^n$ is universally compressible. We complement this result by showing that for k growing exponentially with n, $\mathcal{M}_k^n$ has linear redundancy in the block-length, thereby showing a nearly tight characterization of the range of alphabet size for which universal coding is possible. We also consider the case $k = 2^{\omega(n)}$, where alphabet size is super-exponential in the block length. For this case, we show that there is essentially no advantage of block-compression over symbol by symbol compression.

In particular, we prove the following results.

Theorem 21. *For large n and any k,*

$$\overline{R}(\mathcal{M}_k^n) \le \sqrt{40 n \log k \log n}.$$

It follows that

Corollary 22. *For any $k = 2^{o(n/\log n)}$,*

$$\overline{R}(\mathcal{M}_k^n) = o(n).$$

We also provide a nearly matching lower bound.

Theorem 23. *For $k = 2^{\Omega(n)}$,*

$$\overline{R}(\mathcal{M}_k^n) = \Omega(n).$$

Theorem 24. *For $k = 2^{\omega(n)}$,*

$$n \log \frac{\log k}{n} - O(n) \leq \overline{R}(\mathcal{M}_k^n) \leq n \log\log k.$$

We extend these results to bound the redundancy of m-modal distributions in terms of monotone distributions.

Theorem 25. *For large n and any $k \geq m$*

$$\overline{R}(\mathcal{M}_{k,m}^n) \leq \log \binom{k}{m} + (m+1)\overline{R}(\mathcal{M}_k^n).$$

As a corollary we show that for a constant number of modes, $\mathcal{M}_k^n$ is universally compressible for $k \leq 2^{o(n/\log n)}$.

Corollary 26. *For $m = O(1)$ any $k = 2^{o(n/\log n)}$,*

$$\overline{R}(\mathcal{M}_k^n) = o(n).$$

4.3 Proofs

4.3.1 Proof of Theorem 21

The main ingredient of our proof is to show that even though the underlying monotone distribution has alphabet size k, it can be approximated in KL-divergence by a step distribution with $m \ll k$ steps. This can be thought of as reducing the *degres of freedom* from k to m. While such approximation results are known in ℓ_1 distance [48], proving the same for KL divergence is non-trivial as KL divergence is not a metric and does not satisfy triangle inequality. This is the most technical part of the paper and we defer it to the next section. Let $I_1, \ldots, I_b$ be a partition of $[k]$ into consecutive intervals. Let $\overline{\mathcal{M}}(I_1^b)$ be the set of monotone step distributions such that for any $i, j \in I_l$, then $p(i) = p(j)$. For an interval I, $|I|$ denotes the number of integers in that interval.

Theorem 27. *Let $b \geq 10 \log k$. There exists a set of intervals $I_1, I_2, \ldots I_b$ such that for ever $P \in \mathcal{M}$, there is a $\bar{P} \in \overline{\mathcal{M}}(I_1^b)$ such that*

$$D(P||\bar{P}) \leq 10 \frac{\log k}{b}.$$

Using the above result we first show that the redundancy is $o(n)$ for $k < 2^{o(n)}$.

For a distribution $P \in \mathcal{M}_k$ let $\bar{P} \in \overline{\mathcal{M}}(I_1^b)$ be the distribution whose mean in each interval is same as P, *i.e.,* for $x \in I_j$

$$\bar{P}(x) = \frac{P(I_j)}{|I_j|}. \tag{4.1}$$

As before $\bar{P}^n$ is the distribution obtained by sampling $\bar{P}$ *i.i.d.* n times. Then for any distribution Q_n over $[k]^n$,

$$\begin{aligned} D(P^n||Q_n) &= \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{P^n(x_1^n)}{Q_n(x_1^n)} \\ &= \sum_{x_1^n \in [k]^n} P^n(x_1^n) \left[\log \frac{P^n(x_1^n)}{\bar{P}^n(x_1^n)} + \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)} \right] \\ &= nD(P||\bar{P}) + \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)}, \end{aligned}$$

where the last step follows since the KL divergence for product distributions is the sum of KL divergence of distributions on each coordinate.

By Theorem 27 and the definition of redundancy

$$\overline{R}(\mathcal{M}_k^n) \le \frac{10n \log k}{b} + \inf_{Q_n} \sup_{P \in \mathcal{M}_k} \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)}.$$

We now use Equation (2.3) to obtain a distribution over $[k]^n$ that has a small KL divergence with respect to any distribution in $\mathcal{M}_k^n$. By Equation 2.3 there is a distribution $Q_{b,n}$ over $[b]^n$ such that for any distribution P over $[b]$,

$$D(P^n||Q_{b,n}) \le (b-1) \log n. \tag{4.2}$$

We use the intervals to map $[k]^n \to [b]^n$, and then use $Q_{b,n}$ to obtain a distribution over $[k]^n$.

For any $x \in [k]$, let j be the interval such that $x \in I_j$. Let $f(x) = j$, then f maps $[k]$ to $[b]$. Then $f(x_1^n) = f(x_1, \ldots, x_n) \stackrel{\text{def}}{=} f(x_1), \ldots, f(x_n)$ maps $[k]^n$ to $[b]^n$. Let $f(x_1^n) = j_1, \ldots, j_n$. The number of x_1^n that map to $j_1^n \stackrel{\text{def}}{=} j_1, \ldots, j_n$ is $|I_{j_1}| \ldots |I_{j_n}|$. Using this we define distribution $\bar{Q}_n$ over $[k]^n$ as

$$\bar{Q}_n(x_1^n) = \frac{Q_{b,n}(j_1^n)}{\prod_{i=1}^n |I_{j_i}|}.$$

Then for any distribution $P \in \mathcal{M}_k$

$$\begin{aligned}
&\sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)} \\
&= \sum_{j_1^n \in [b]^n} \sum_{x_1^n : f(x_1^n) = j_1^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)} \\
&\overset{(a)}{=} \sum_{j_1^n \in [b]^n} \sum_{x_1^n : f(x_1^n) = j_1^n} P^n(x_1^n) \log \frac{P(I_{j_1}) \ldots P(I_{j_n})}{Q_{b,n}(j_1^n)} \\
&= \sum_{j_1^n \in [b]^n} \log \frac{\prod_{i=1}^n P(I_{j_i})}{Q_{b,n}(j_1^n)} \left(\sum_{x_1^n : f(x_1^n) = j_1^n} P^n(x_1^n) \right),
\end{aligned}$$

where (a) uses Equation (4.1). Now, for $j_1^n \in [b]^n$,

$$\sum_{x_1^n : f(x_1^n) = j_1^n} P^n(x_1^n) = \sum_{x_1^n : x_i \in I_{j_i}} \prod_{i=1}^n P(x_i) = \prod_{i=1}^n P(I_{j_i}).$$

Therefore,

$$\sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)} = \sum_{j_1^n \in [b]^n} \prod_{i=1}^n P(I_{j_i}) \log \frac{\prod_{i=1}^n P(I_{j_i})}{Q_{b,n}(j_1^n)}$$

Now, a distribution P induces a distribution over the intervals $I_1, \ldots, I_b$ and this expression is the KL divergence of the product distribution over the intervals to $Q_{b,n}$. Therefore by Equation (4.1) is bounded by $(b-1)\log n$.

Plugging this we obtain,

$$\begin{aligned}
\overline{R}(\mathcal{M}_k^n) \leq & \frac{10n \log k}{b} + \inf_{Q_n} \sup_{P \in \mathcal{M}_k} \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)} \\
\leq & \frac{10n \log k}{b} + \sup_{P \in \mathcal{M}_k} \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{\bar{Q}_n(x_1^n)} \\
= & \frac{10n \log k}{b} + \sup_{P \in \mathcal{M}_k} \sum_{j_1^n \in [b]^n} \prod_{i=1}^n P(I_{j_i}) \log \frac{\prod_{i=1}^n P(I_{j_i})}{Q_{b,n}(j_1^n)} \\
\leq & \frac{10n \log k}{b} + (b-1) \log n.
\end{aligned}$$

Choosing $b = \sqrt{\frac{10n \log k}{\log n}}$ results in the Theorem. We now prove Theorem 27.

Proof of Theorem 27

We first state a simple result on the set of non-negative numbers.

Lemma 28. *For* $0 \le x_1 \le x_2 \ldots \le x_n$ *with mean* $\bar{x}$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \le n(x_n - x_1)\bar{x}.$$

Proof.

$$\begin{aligned} n(x_n - x_1)\bar{x} - \sum_{i=1}^{n}(x_i - \bar{x})^2 &= nx_n\bar{x} - \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - n\bar{x}x_1 \\ &\ge nx_n\bar{x} - \sum_{i=1}^{n} x_i^2 \ge 0. \end{aligned}$$

□

As a simple application of the above result we bound the KL divergence between P and $\bar{P}$. Let p_j^+ and p_j^- be the maximum and minimum value of probabilities in the interval I_j. Let k_j be the number of non-zero probabilities in interval I_j.

Lemma 29.

$$D(P||\bar{P}) \le \sum_{j=1}^{b} k_j (p_j^+ - p_j^-).$$

Proof. By Jensen's inequality,

$$\begin{aligned} \sum_{x \in I_j} p(x) \log \frac{p(x)}{\bar{p}_j} \le & \sum_{x \in I_j} p(x) \frac{p(x) - \bar{p}_j}{\bar{p}_j} \\ = & \sum_{x \in I_j} \frac{p^2(x) - \bar{p}_j^2}{\bar{p}_j} \\ = & \frac{1}{\bar{p}_j} \sum_{x \in I_j} (p(x) - \bar{p}_j)^2. \end{aligned}$$

The rest of the proof follows from Lemma 28 applied to each interval. □

Let $\gamma = \frac{2 \log k}{b}$. We now choose the intervals as follows:

$$|I_j| = \begin{cases} 1 & \text{if } j \le \frac{b}{2}, \\ \lfloor 2(1+\gamma)^{j-b/2} \rfloor & \text{else,} \end{cases}$$

Since

$$\begin{aligned}\sum_{i=1}^{b} |I_j| &\geq b/2 + \sum_{i=b/2+1}^{b} \lfloor 2(1+\gamma)^{j-b/2} \rfloor \\ &\geq \sum_{i=b/2+1}^{b} 2(1+\gamma)^{j-b/2} \\ &= 2\frac{1+\gamma}{\gamma}\left((1+\gamma)^{b/2}-1\right).\end{aligned}$$

For $\gamma \geq \frac{2\log k}{b}$, the above quantity is $\geq k$ and the intervals span all the alphabet. Since $|I_j|$ is 1 for $j \leq b/2$, we have $p_j^+ = p_j^-$ for $j \leq b/2$. By Lemma 29,

$$\begin{aligned}D(P||\bar{P}) &\leq \sum_{j=1}^{b} k_j(p_j^+ - p_j^-) \\ &= \sum_{j=\frac{b}{2}+1}^{b} k_j(p_j^+ - p_j^-) \\ &= k_{b/2+1}p_{b/2+1}^+ + \sum_{j=b/2+1}^{b} (p_{j+1}^+ k_{j+1} - p_j^- k_j) \\ &\leq k_{b/2+1}p_{b/2+1}^+ + \sum_{j=b/2+1}^{b} p_j^-(k_{j+1} - k_j).\end{aligned}$$

Observe that k_{j+1} is non-zero only if $k_j = |I_j|$. Hence if k_{j+1} is non-zero,

$$k_{j+1} - k_j \leq |I_{j+1}| - |I_j| \leq 2\gamma|I_j| \leq 2\gamma k_j.$$

The factor 2 appears because of the floor in defining I_j. Even if $k_j = 0$, we have $k_{j+1} - k_j \leq 2\gamma k_j$. Hence,

$$\begin{aligned}D(P||\bar{P}) &\leq k_{b/2+1}p_{b/2+1}^+ + 2\gamma \sum_{j=b/2+1}^{b} p_j^- k_j \\ &\leq |I_{b/2+1}|p_{b/2+1}^+ + 2\gamma.\end{aligned}$$

Substituting values of γ and $|I_{b/2+1}|$ and the fact that $p_{b/2+1}^+ \leq \frac{2}{b}$ proves the result.

4.3.2 Proof of Theorem 24

In this section we prove the upper bound of the theorem, and sketch the lower bound in the final section. We use the following simple lemma that states that block compression has smaller redundancy than using a single compression scheme.

Lemma 30. *For a class $\mathcal{P}$,*

$$\overline{R}(\mathcal{P}^n) \le n\overline{R}(\mathcal{P}).$$

Proof. By the definition of redundancy,

$$\begin{aligned}\overline{R}(\mathcal{P}^n) &= \inf_{Q_n} \sup_{P^n} D\left(P^n || Q_n\right), \\ &\le \inf_{Q_n:product} \sup_{P^n} D\left(P^n || Q_n\right) \\ &= \sum_{i=1}^{n} \inf_{Q} \sup_{P\in\mathcal{P}} D\left(P || Q\right) = n\overline{R}(\mathcal{P}),\end{aligned}$$

where the inequality follows by restricting Q_n to product distributions. □

Using this along with $\overline{R}(\mathcal{M}_k) \le \hat{R}(\mathcal{M}_k) \sim \log\log k$ gives,

$$\overline{R}(\mathcal{M}_k^n) \le n\overline{R}(\mathcal{M}_k) \sim n \log\log k.$$

4.3.3 m-modal distributions

In this section consider the redundancy of m-modal distributions and prove Theorem 25. We decompose the class $\mathcal{M}_{k,m}$ into $\binom{k}{m}$ classes and then invoke Lemma 17, which bounds the redundancy of union of classes.

For any m-modal distribution, if we are given one point from each of the modes, then the distribution is monotone within the intervals thus formed. The number of possibilities of the m points is at most $\binom{k}{m}$.

We therefore consider one class of distributions from $\mathcal{M}_{k,m}$, specified by m points. It is not hard to repeat the computations of monotone distributions on each of the $m+1$ monotone distributions formed over the intervals. Once again by monotonicity of redundancy with block-length, each and therefore the total extra number of bits can be bounded by $(m+1)\overline{R}(\mathcal{M}_k^n)$. Combining with Lemma 17 with $T = \binom{k}{m}$ proves the theorem.

4.3.4 Lower bounds

By Lemma 9, suppose there exist M distributions $P_1, \ldots, P_M$ in $\mathcal{M}_k$, and a partition of $[k]^n$ into $\mathcal{S}_1, \ldots, \mathcal{S}_M$ such that for some $0 < e < 1$,

$$P_i^n(\mathcal{S}_i) > 1 - e,$$

then

$$\overline{R}(\mathcal{M}_k^n) \geq (1 - e) \log M - h(e).$$

Note that since $\mathcal{M}_k \subset \mathcal{M}_{k+1}$, $\overline{R}(\mathcal{M}_k^n) \leq \overline{R}(\mathcal{M}_{k+1}^n)$, it will suffice to prove Theorem 23 for $k = 2^n$.

For $k = 2^n$, we now construct a class of $M = 2^{cn}$ distributions for a constant c, with the above property. This will give a lower bound of $\sim cn(1-e)$ on $\overline{R}(\mathcal{M}_k^n)$. We first restrict to the following subclass of $\mathcal{M}_k$. Consider distributions in $\mathcal{M}_k$ such that for each $j < \log n$, one of the following two conditions are satisfied.

1. For all $2^j < i \leq 2^{j+1}$,
$$p(i) = \frac{2}{3 \cdot 2^j} \frac{1}{n}.$$
2. For all $2^j < i \leq 2^{j+1}$,
$$p(i) = \frac{4}{3 \cdot 2^j} \frac{1}{n}.$$

Since we have partitioned the interval $[k]$ with interval size doubling at each step, we see that any distribution that satisfies the conditions above is monotone.

$$p(2^j) \geq \frac{2}{3 \cdot 2^{j-1}} = \frac{4}{3 \cdot 2^j} \geq p(2^j + 1).$$

Since a distribution changes values at precisely these points, it is monotone.

Any distribution satisfying the above condition assigns probability $1/2n$ or $3/2n$ to each interval $(2^j, 2^{j+1}]$. Assuming that n is even, such a distribution satisfies each of the above condition in exactly half of the intervals, (since probabilities sum to 1).

In fact this shows that the number of such distributions is precisely $\binom{n}{n/2}$. We now select a set of distributions from these to satisfy our requirements. For such a distribution P, let $S(P)$ be the $n/2$ intervals that satisfy condition 1.

By the Gilbert-Varshamov bound of Lemma 31, we can now show the following result.

Lemma 31. *There exist a class of* $M = 2^{n(1-h(\alpha)-o(1))}$ *monotone distributions* $P_1, \ldots, P_M$ *satisfying the condition described, such that any two such distributions* P_i *and* P_j *satisfy*

$$|S(P_i) \cap S(P_j)| < n(1-\alpha)/2.$$

In other words for any pair of distributions, their distributions are different in at least a fraction $(1-\alpha)/2$ of the intervals.

Consider n samples obtained from any such distribution, say P_1. The number of samples in $S(P_1)$ is $B(n, n/3)$, *i.e.,* about $n/3$ samples fall in these intervals and the remaining $2n/3$ from the other intervals. By Lemma 31, the number of samples from distribution P_j, $j \neq 1$, is $B(n, n(1+\alpha)/3)$.

By simple tail bounds on the Binomial distributions, it follows that we can choose an $\alpha < 1/2$ such that the distributions can be reconstructed from the samples with a constant probability. This proves Theorem 23.

By a similar construction over $\log k$ bins with doubling sizes also proves the lower bound of Theorem 24 and is omitted.

4.4 Summary

The following table summarizes the current knowledge of the redundancy of $\mathcal{M}_k^n$.

Table 4.2: Current bounds on $\overline{R}(\mathcal{M}_k^n)$

Range	Lower bound	Upper bound
$k = o(n^{1/3})$	$\Omega(k \log \frac{n}{k})$	$O(k \log \frac{n}{k})$
$k = n^{O(1)}$	$\Omega(n^{1/3})$	$O(n^{1/3} \log n)$
$k = \exp(o(n))$	$\Omega(n^{1/3})$	$O(\sqrt{n \log k})$
$k = \exp(\omega(n))$	$O(n \log((\log k)/n))$	$O(n \log \log k)$

Open problem 32. *Find $\overline{R}(\mathcal{M}_k^n)$, for the range $k = \exp(o(n))$ and $k = n^{\omega(1)}$. In particular, show that in this range,*

$$\overline{R}(\mathcal{M}_k^n) = \tilde{O}(n^{1/3}(\log k)^{2/3}).$$

Acknowledgement

Chapter 4 is partially adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, “Efficient compression of monotone and m-modal distributions”, submitted to *IEEE International Symposium on Information Theory (ISIT)*, 2014. The dissertation author was a primary researcher and author of this work.

Chapter 5

Compression of envelope classes

In the introduction we discussed that the class of *i.i.d.* distributions over an infinite underlying alphabet has infinite redundancy, and also saw in the previous chapter that the same holds even for the class of monotone distributions over $\mathbb{N}$. We considered monotone distributions over bounded support, and provided bounds on the redundancy in terms of support size k and block length n. In particular we showed that monotone distributions can be universally compressed in the regime $k = \exp(o(n))$.

In this chapter, we consider another natural and elegant approach proposed in [12]. They study universal compression of a fairly general class of distributions, called *envelope classes*, which as the name suggests are distributions that are bounded by an envelope. An envelope class is characterized by a function $f : \mathbb{N} \to \mathbb{R}^+$.

Definition 33. *The envelope class associated with a function f is the class*

$$\mathcal{P}_f \stackrel{\text{def}}{=} \{(p_1, p_2, \ldots) : 0 \leq p_j \leq f(j), \text{ and } p_1 + p_2 + \ldots = 1\}$$

of all distributions such that the symbol probability is bounded by the value of the function at that point.

Similar to the definition of $\mathcal{P}^n$, let

$$\mathcal{P}_f^n \stackrel{\text{def}}{=} \{P^n : P \in \mathcal{P}_f\}$$

be the class of length-n *i.i.d.* distributions of an envelope distribution.

[12] provide general bounds on the redundancy of envelope classes. The upper bounds on the worst-case redundancy are obtained by bounding the Shtarkov sum. They provide bounds on the more stringent (for lower bounds) average case redundancy. They provide a relatively complex bound employing the redundancy-capacity theorem.

Envelope classes capture a wide range of distributions. They also take into account scenarios in which there is a prior knowledge that the source distributions are close to a class of “nice” distributione. For example, the distributions could be from an *almost* power law class, in which we know that the probability of symbol i is close to a power law distribution.

There are a number of problems in which there is some form of side information about the data. In language modeling, it is common to assume that word distributions follow the Zipf distribution [39], which is a monotone distribution. In image compression, geometric distributions over integers arise [40, 54]. Power-law distribution is very common in modeling numerous types of random variables, for example the distribution of wealth, etc [55].

In this chapter we are interested in compression of restricted classes of *i.i.d.* distributions over countably infinite alphabets, *e.g.,* $\mathbb{N}$. In particular, we first discuss conditions on the class of distributions to be universally compressible. We then consider envelope classes of distributions. Based on the Poisson sampling framework we derive bounds on the redundancy of a general envelope class. These bounds are simple to represent and apply. We show the efficacy of our bounds by giving the exact growth rate of the power-law class, solving an open problem in [12]. We also provide alternate proof and slightly strengthen the bounds on the redundancy of exponential envelope class.

In the next section, we prove some relations between the redundancy of Poisson sampling to sampling exact n times. This shows that it suffices to consider Poisson-sampling.

5.1 Poisson redundancy

The next result captures some properties of redundancy of classes of the form $\mathcal{P}^n$ over any discrete alphabet (even countably infinite).

Lemma 34. *1.* Monotonicity: *For all n, $\hat{R}(\mathcal{P}^{n+1}) \geq \hat{R}(\mathcal{P}^n)$*

2. Linearity: *If $S(\mathcal{P}) < \infty$, then $\hat{R}(\mathcal{P}^n) < n\hat{R}(\mathcal{P})$*

3. Finiteness: $S(\mathcal{P}) < \infty \Leftrightarrow \hat{R}(\mathcal{P}^n) < \infty$

4. Sublinearity: *If $S(\mathcal{P}) < \infty$ then $\hat{R}(\mathcal{P}^n) = o(n)$.*

Proof. Monotonicity follows by marginalizing the $(n+1)$th coordinate.

$$\begin{aligned}
S(\mathcal{P}^{n+1}) &= \sum_{x_1^{n+1} \in \mathcal{X}^{n+1}} \sup_{P \in \mathcal{P}} P^n(x_1^{n+1}) \\
&\geq \sum_{x_1^n \in \mathcal{X}^n} \sup_{P \in \mathcal{P}} \left[P^n(x_1^n) \left(\sum_{x_{n+1} \in \mathcal{X}} P(x_{n+1}) \right) \right] \\
&\geq \sum_{x_1^n \in \mathcal{X}^n} \sup_{P \in \mathcal{P}} P^n(x_1^n) \\
&= S(\mathcal{P}^n).
\end{aligned}$$

The second item follows using $S(\mathcal{P}^n) < S(\mathcal{P})^n$ which is a consequence of $\hat{P}^n(\overline{x}) \leq \prod \hat{P}(x_j)$. Combining the first two items yield the third. The final part is a result of [12]. It states that if the worst case redundancy of a class is finite, then it must grow sub-linearly. □

Let

$$\mathcal{P}^{\text{poi}(n)} \stackrel{\text{def}}{=} \{P^{\text{poi}(n)} : P \in \mathcal{P}\},$$

where recall from Section 2.6 that $P^{\text{poi}(n)}$ is the distribution over $\mathcal{X}^*$ when P is sampled independently poi(n) times.

The next theorem bounds the redundancy of P^n in terms of the redundancy of $P^{\text{poi}(n)}$. This implies that finding bounds on one of them also provides bounds on the other. We would be primarily interested in the second bound, since will provide stronger upper bounds on the redundancy of envelope classes.

Theorem 35. *Suppose $\hat{R}(\mathcal{P}) < \infty$. For any $\epsilon > 0$ there is $n_0(\epsilon, \mathcal{P})$, such that for $n > n_0$,*

$$\hat{R}(\mathcal{P}^{\text{poi}(n(1-\epsilon)}) - 1 \le \hat{R}(\mathcal{P}^n) \le \hat{R}(\mathcal{P}^{\text{poi}(n)}) + 1$$

Proof. The second inequality is easier to show.

$$\begin{aligned} S(\mathcal{P}^{\text{poi}(n)}) &\overset{(a)}{=} \sum_{n' \ge 0} \text{poi}(n, n') S(\mathcal{P}^{n'}) \\ &\overset{(b)}{\ge} S(\mathcal{P}^n) \sum_{n' \ge n} \text{poi}(n, n') \\ &\overset{(c)}{\ge} \frac{1}{2} S(\mathcal{P}^n), \end{aligned}$$

where (a) follows from Lemma 7, (b) from monotonicity of $S(\mathcal{P}^n)$ and (c) from the fact that median of a Poisson distribution larger than its mean for large means. Taking logarithms gives the result.

For the first inequality

$$S(\mathcal{P}^{\text{poi}(n(1-\epsilon))}) = \sum_{n'} \text{poi}(n(1-\epsilon), n') S(\mathcal{P}^{n'})$$

By the Poisson tail bound of Lemma 4,for $n' \ge n$

$$\text{poi}(n, n') \le \Pr(\text{poi}(n(1-\epsilon)) \ge n') < e^{-n'/\epsilon^2}.$$

By item 4 of Lemma 34 since $\hat{R}(\mathcal{P}^n) = o(n)$, $S(\mathcal{P}^{n'}) = \exp(\hat{R}(\mathcal{P}^{n'})) = \exp(o(n'))$ the contribution of terms $\ge n$ are negligible. Using monotonicity yields the result. □

In the next section, we bound the redundancy of general envelope classes.

5.2 Redundancy bounds on envelope classes

We bound the redundancy of $\mathcal{P}^{\text{poi}(n)}$ in terms of the redundancy of the following *primitive* class

$$\text{POI}(\lambda^{\max}) \overset{\text{def}}{=} \{\text{poi}(\lambda) : \lambda < \lambda^{\max}\}.$$

This is the class of all Poisson distributions with mean bounded above. This class is simple enough, and we can bound its redundancy tightly.

We first investigate POI(λ) and then bound the redundancy of *i.i.d.* distributions in terms of $\hat{R}(\text{POI}(\lambda))$.

5.2.1 Redundancy of Poisson distributions

The maximum likelihood Poisson distribution of a non-negative integer j over all Poisson distributions is poi(j). The distributions in POI(λ) that maximizes the probability of an integer j is:

$$\arg\max_{\text{POI}(\lambda)} P(i) = \begin{cases} \text{poi}(i) & \text{if } i \le \lfloor \lambda \rfloor \\ \text{poi}(\lambda) & \text{otherwise.} \end{cases}$$

Using this, the Shtarkov sum of the class is

$$\begin{aligned} S\Big(\text{POI}(\lambda)\Big) &= \sum_{i=0}^{\lfloor\lambda\rfloor} e^{-i}\frac{i^i}{i!} + \sum_{\lfloor\lambda\rfloor+1}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} && (5.1) \\ &\overset{(a)}{=} 1 + \sum_{i=0}^{\lfloor\lambda\rfloor} \Big(e^{-i}\frac{i^i}{i!} - e^{-\lambda}\frac{\lambda^i}{i!}\Big), && (5.2) \end{aligned}$$

where (a) uses that $\sum_\mu \text{poi}(\lambda, \mu) = 1$.

From this Equation we obtain the following bound on redundancy of POI(λ).

Lemma 36. *For $\lambda \le 1$,*

$$\hat{R}\Big(\text{POI}(\lambda)\Big) = \log\left(2 - \exp(-\lambda)\right) \le \lambda,$$

and for $\lambda \ge 1$,

$$\sqrt{\frac{2(\lambda+1)}{\pi}} \le \hat{R}\Big(\text{POI}(\lambda)\Big) \le 2 + \sqrt{\frac{2\lambda}{\pi}}.$$

Proof. For the first part, we use Equation (5.2), for the equality, and the inequality follows from $e^{-x} + e^{x} \ge 2$. For the second part, using the Stirling's Approximation(Lemma 1) with Equation (5.1)

$$\begin{aligned}
S\Big(\text{POI}(\lambda)\Big) &< 2 + \sum_{j=1}^{\lfloor\lambda\rfloor} e^{-j}\frac{j^j}{j!} \\
&\overset{(a)}{\leq} 2 + \sum_{j=1}^{\lfloor\lambda\rfloor} e^{-j}\frac{j^j}{\sqrt{2\pi j}(\frac{j}{e})^j} \\
&= 2 + \sum_{j=1}^{\lfloor\lambda\rfloor} \frac{1}{\sqrt{2\pi j}} \\
&\overset{(b)}{\leq} 2 + \sqrt{\frac{2\lambda}{\pi}},
\end{aligned}$$

where (a) follows from $j! > \sqrt{2\pi j}(\frac{j}{e})^j$, and (b) from a simple integration. The lower bound follows from a similar computation. □

5.2.2 General envelope class

Recall that the class of distributions associated with an envelope function $f : \mathbb{N} \to \mathbb{R}^+$ is

$$\mathcal{P}_f \overset{\text{def}}{=} \{(p_1, p_2, \ldots) : 0 \leq p_j \leq f(j), \text{ and } p_1 + p_2 + \ldots = 1\}.$$

Suppose the envelope class is integrable, *i.e.*,

$$\sum_{j\geq 1} f(j) < \infty,$$

then Lemma 34 implies that $S(\mathcal{P}_f) < \infty$. If the sum is not finite then $S(\mathcal{P}_f) = \infty$. For an envelope characterized by f, let

$$\lambda_j^{\max} \overset{\text{def}}{=} nf(j)$$

be the largest possible value of np where p is the probability of j. Let l_f be the smallest integer such that

$$\sum_{j\geq l_f} f_j \leq 1 - \epsilon.$$

This also implies that $\sum_{j\geq l_f} \lambda_j^{\max} \leq n(1-\epsilon)$. We do not mention ℓ_f as a function of epsilon, since it will have little effect on the results, as seen later.

We now state our main result on envelope classes.

Theorem 37. *For the envelope class $\mathcal{P}_f$, for large n and fixed $\epsilon > 0$,*

$$\sum_{i=l_f}^{\infty} \hat{R}\left(\mathrm{POI}(\lambda_i^{\max})\right) \le \hat{R}\left(\mathcal{P}_f^{\mathrm{poi}(n)}\right) \le \sum_{i=1}^{\infty} \hat{R}\left(\mathrm{POI}(\lambda_i^{\max})\right),$$

where $\lambda_i^{\max} = nf_i$.

Proof. By Lemma 13, it suffices to consider the redunadncy of types of sequences. By item 2 of Lemma 6, for a distribution a distribution $P = (p_1, p_2, , \ldots)$ over $\mathcal{X} = \{1, 2, \ldots\}$ can be rewritten as

$$\tau(P^{\mathrm{poi}(n)}) = (\mathrm{poi}(np_1), \mathrm{poi}(np_2), \ldots),$$

where each coordinate is an independent Poisson distribution. Similar to this, we can show that the class of type distributions induced by $\mathcal{P}_f$ is

$$\tau\left(\mathcal{P}_f^{\mathrm{poi}(n)}\right) = \left\{(\mathrm{poi}(\lambda_1), \mathrm{poi}(\lambda_2) \ldots) : \lambda_i \le nf_i, \sum \lambda_i = n\right\}.$$

In other words, under Poisson sampling the distribution of types is a product of Poisson distributions, where the parameter of ith coordinate is at most $nf(i) = \lambda_i^{\max}$. It follows that

$$\mathcal{P}_f^{\mathrm{poi}(n)} \subset \mathrm{POI}(\lambda_1^{\max}) \times \mathrm{POI}(\lambda_2^{\max}) \times \ldots.$$

The product redunadncy lemma (Lemma 14) can be generalized to products of any countable number of distributions, and hence

$$\hat{R}\left(\mathcal{P}_f^{\mathrm{poi}(n)}\right) = \hat{R}\left(\tau(\mathcal{P}_f^{\mathrm{poi}(n)})\right) \le \hat{R}\left(\mathrm{POI}(\lambda_1^{\max})\right) + \hat{R}\left(\mathrm{POI}(\lambda_2^{\max})\right) + \ldots.$$

For the lower bound, note for any choice of $\lambda_i < \lambda_i^{\max}$ for $i \ge l_f$ corresponds to a distribution in $\mathcal{P}_f^{\mathrm{poi}(n)}$. In other words, all product distributions in

$$\mathrm{POI}(\lambda_{l_f}^{\max}) \times \mathrm{POI}(\lambda_{l_f+1}^{\max}) \times \ldots$$

are valid projections of a distribution in $\mathcal{P}_f^{\mathrm{poi}(n)}$ along the coordinates $i \ge l_f$. Therefore, along these coordinates Lemma 14 holds with equality. Dropping the other coordinates simply reduces the redundancy, thus proving the lower bound. □

We now consider power-law and exponential envelopes and apply Theorem 37 to obtain sharp bounds on redundancies of these classes improving the previous results.

5.3 Applications to specific envelope classes

We now find bounds on the redundancy of Power-law and exponential envelopes.

Definition 38. *The* power-law *envelope class $\Lambda_{c\cdot^{-\alpha}}$ with parameters $\alpha > 1$ and c is the collection of distributions over $\mathbb{N}$ bounded by a power-law envelope with exponent $\alpha > 1$ and coefficient c, i.e.,*

$$f(i) = \frac{c}{i^{\alpha}}.$$

The redundancy of $\Lambda^n_{c\cdot^{-\alpha}}$ was considered in [12] who prove that for large n,

$$C_0 n^{\frac{1}{\alpha}} \le \hat{R}(\Lambda^n_{c\cdot^{-\alpha}}) \le \left(\frac{2cn}{\alpha-1}\right)^{\frac{1}{\alpha}} (\log n)^{1-\frac{1}{\alpha}} + O(1), \tag{5.3}$$

where C_0 is a constant (function of α and c).

We show that simply applying Theorem 37 to these classes and bounding the resulting expressions gives tight redundancy bounds, removing the logarithmic factor. We prove that the lower bound of [12] on the average redundancy is within a constant factor of the actual worst case redundancy by proving the following theorem.

Theorem 39. *For large n*

$$(cn)^{1/\alpha}\left[\frac{\alpha}{2} + \frac{1}{2(\alpha-1)} - \frac{\log 3}{2}\right] - 1 \le \hat{R}(\Lambda^n_{c\cdot^{-\alpha}}) \le (cn)^{1/\alpha}\left[\frac{\alpha}{2} + \frac{1}{\alpha-1} + \log 3\right] + 1.$$

Proof. By Definition 38, for power-law class $\Lambda_{c\cdot^{-\alpha}}$,

$$\lambda_i^{\max} = \frac{cn}{i^{\alpha}}$$

is the largest expected multiplicity of symbol i. Let $b \overset{\text{def}}{=} (cn)^{1/\alpha}$, then $\lambda_i^{\max} \ge 1$ for $i \le b$ and $\lambda_i^{\max} < 1$ otherwise.

Then,

$$\begin{aligned}
\hat{R}(\Lambda^{\text{poi}(n)}_{c\cdot^{-\alpha}}) &\overset{(a)}{\le} \sum_{i\le b} \hat{R}(\text{POI}(\lambda_i^{\max})) + \sum_{i>b} \hat{R}(\text{POI}(\lambda_i^{\max})) \\
&\overset{(b)}{\le} \sum_{i\le b} \log\left(2 + \sqrt{\frac{2\lambda_i^{\max}}{\pi}}\right) + \sum_{i>b}^{\infty} \lambda_i^{\max},
\end{aligned}$$

where (a) follows from Theorem 37 and (b) from Lemma 36.

We consider the two summations separately. For the first term, we note that for $\lambda \geq 1$, $2 + \sqrt{2\lambda/\pi} < 3\sqrt{\lambda}$ and use it with the following simplification.

$$\sum_{i=1}^{B} \log \frac{B}{i} = \log \frac{B^B}{B!} \leq B,$$

which follows from $B! > (B/e)^B$ using Stirling's approximation. Therefore,

$$\begin{aligned}\sum_{i=1}^{b} \log\left(2 + \sqrt{\frac{2\lambda_i^{\max}}{\pi}}\right) &< \sum_{i=1}^{b} \log\left(3\sqrt{\frac{cn}{i^\alpha}}\right)\\ &= b\log(3) + \frac{\alpha}{2}\sum_{i=1}^{b}\log\left(\frac{(cn)^{\frac{1}{\alpha}}}{i}\right)\\ &\overset{(a)}{<} (cn)^{1/\alpha}\Big(\log(3) + \frac{\alpha}{2}\Big),\end{aligned}$$

where (a) follows since $b = (cn)^{1/\alpha}$.

Taking the second term,

$$\begin{aligned}\sum_{i=b+1}^{\infty} \lambda_i^{\max} &= cn \sum_{i=b+1}^{\infty} \frac{1}{i^\alpha}\\ &= \frac{c^{1/\alpha}}{\alpha - 1} n^{1/\alpha},\end{aligned}$$

where (b) follows by using $s = b + 1 = (cn)^{1/\alpha} + 1$ in

$$\sum_{i=s}^{\infty} \frac{1}{i^r} \leq \int_s^\infty \frac{1}{(x-1)^r} \leq \frac{(s-1)^{1-r}}{(r-1)}. \tag{5.4}$$

Finally applying Theorem 35,

$$\hat{R}(\Lambda^n_{c\cdot^{-\alpha}}) \leq \hat{R}(\Lambda^{\mathrm{poi}(n)}_{c\cdot^{-\alpha}}) + 1 \leq (cn)^{1/\alpha}\Big[\log 3 + \frac{\alpha}{2} + \frac{1}{\alpha - 1}\Big] + 1.$$

We now prove the lower bound. Let $\epsilon > 0$ be any constant. Then by Equation (5.4)

$$\sum_{j=\ell+1}^{\infty} \lambda_j^{\max} \leq cn\frac{\ell^{1-\alpha}}{(\alpha-1)},$$

and therefore for

$$\ell \overset{\mathrm{def}}{=} \left(\frac{c}{(\alpha-1)(1-\epsilon)}\right)^{\frac{1}{\alpha-1}}$$

the sum above is at most $n(1-\epsilon)$. Applying the lower bound from Theorem 37,

$$\begin{aligned}\hat{R}(\Lambda_{c\cdot^{-\alpha}}^{\text{poi}(n)}) &\le \sum_{i>\ell} \hat{R}(\text{POI}(\lambda_i^{\max})) \\ &= \sum_{\ell<i\le b} \hat{R}(\text{POI}(\lambda_i^{\max})) + \sum_{i>b} \hat{R}(\text{POI}(\lambda_i^{\max})).\end{aligned}$$

Considering the second term and using $2-e^{-\lambda} < 1+\lambda/2$,

$$\sum_{i>b} \hat{R}(\text{POI}(\lambda_i^{\max})) > \frac{cn}{2}\sum_{i>b}\frac{1}{i^\alpha} > \frac{cn}{2}\frac{(b+2)^{1-\alpha}}{\alpha-1}.$$

Substituting $b=(cn)^{1/\alpha}$ and since $b \gg 2$, we can bound the lower bound the expression above with $\frac{1}{2}(cn)^{1/\alpha}/(\alpha-1)-1$.

We use that $\hat{R}(\text{POI}(\lambda)) \ge \sqrt{\lambda/3}$, which follows from a Shtarkov sum argument. Therefore,

$$\hat{R}(\Lambda_{c\cdot^{-\alpha}}^{n}) \ge \frac{1}{2}\sum_{i=\ell}^{b}\log\frac{cn}{i^\alpha} \ge \frac{1}{2}\sum_{i=1}^{b}\log\frac{cn}{3i^\alpha} - \frac{\ell}{2}\log n \ge \frac{b}{2}[\alpha-\log 3] - \mathcal{O}(\log n)$$

Combining the two bounds and using Theorem 35, we obtain

$$\hat{R}(\Lambda_{c\cdot^{-\alpha}}^{n}) \ge \frac{(cn)^{1/\alpha}}{2}\left[\alpha+\frac{1}{\alpha-1}-\log 3\right]. \qquad \square$$

Remark By obtaining tighter bounds on $\hat{R}(\text{POI}(\lambda))$, it should be possible to obtain upper and lower bounds within an *additive* $\ell\log n$.

5.3.1 Exponential envelope

Definition 40. *The* exponential-law *envelope class with parameters α and c is the class of distributions $\Lambda_{ce^{-\alpha\cdot}}$ over $\mathbb{N}$ such that $\forall i \in \mathbb{N}$,*

$$p_i \le ce^{-\alpha i}.$$

The redundancy of $\Lambda_{ce^{-\alpha\cdot}}^{n}$ was considered in [12] who proved

$$\frac{\log^2 n}{8\alpha}(1+o(1)) \le \hat{R}(\Lambda_{ce^{-\alpha\cdot}}^{n}) \le \frac{\log^2 n}{2\alpha}+O(1).$$

[13] showed the precise growth rate of exponential envelopes and showed that

$$\hat{R}(\Lambda_{ce^{-\alpha\cdot}}^{n}) = \frac{\log^2 n}{4\alpha}(1+o(1)).$$

An analysis of their algorithm shows that the $o(1)$ term is of the form $\frac{\log\log n}{\log n}$, namely their upper bound shows that

$$\hat{R}(\Lambda^n_{ce^{-\alpha\cdot}}) = \frac{\log^2 n}{4\alpha} + O(\log n \log\log n),$$

where c and α are hidden in the order terms. We provide a simple proof of a slightly stronger version of this result using the Poisson sampling technique and prove that

Theorem 41.

$$\hat{R}(\Lambda^n_{ce^{-\alpha\cdot}}) = \frac{\log^2 n}{4\alpha} + O(\log c \log n).$$

Proof. Note that

$$j \le \frac{\log(cn)}{\alpha} \quad \Leftrightarrow \quad \lambda_j^{\max} \ge 1.$$

As with the power-law, let $b \overset{\text{def}}{=} \frac{\log(cn)}{\alpha}$ be the location of this transition. Then

$$\begin{aligned}\hat{R}(\Lambda^{\text{poi}(n)}_{ce^{-\alpha\cdot}}) &\overset{(a)}{\le} \sum_{i\le b} \hat{R}(\text{POI}(\lambda_i^{\max})) + \sum_{i>b} \hat{R}(\text{POI}(\lambda_i^{\max})) \\ &\le \sum_{j=1}^{b} \log\left(2 + \sqrt{\frac{2\lambda_j^{\max}}{\pi}}\right) + \sum_{j=b}^{\infty} \lambda_j^{\max}\end{aligned}$$

Using $e^{b\alpha} = cn$,

$$\sum_{i=b}^{\infty} \lambda_i^{\max} = \sum_{i=b}^{\infty} cne^{-\alpha i} = cne^{-\alpha b}\frac{1}{1-e^{-\alpha}} = \frac{1}{1-e^{-\alpha}}.$$

For any $\lambda > 1$, $2 + \sqrt{2\lambda/\pi} < 3\sqrt{\lambda}$.

$$\begin{aligned}\sum_{i=1}^{b} \log(2 + \sqrt{\frac{2\lambda_i^{\max}}{\pi}}) &\le \sum_{i=1}^{b} \log(3\sqrt{\lambda_i^{\max}}) \\ &= b\log 3 + \frac{1}{2}\sum_{i=1}^{b} \log[cne^{-\alpha i}] \\ &= b\log 3 + \frac{1}{2}\log[(cn)^b e^{-\alpha b(b+1)/2)}] \\ &= b\log 3 + \frac{1}{2}\log[(cn)^b \cdot (cn)^{-(b+1)/2}] \\ &= b\log 3 + \frac{(b-1)}{4}\log(cn) \\ &< b\log(3c) + \frac{b}{4}\log n.\end{aligned}$$

Substituting $b = \log(cn)/\alpha$, and noting that $\log^2 n$ is the dominant term,

$$\hat{R}(\Lambda_{ce^{-\alpha\cdot}}) \leq \frac{\log^2 n}{4\alpha} + \mathcal{O}\Big(\log n\Big).$$

We now prove the lower bound by a similar argument. Clearly, for $a \geq \frac{1}{\alpha}\log(\frac{c}{1-e^{-\alpha}})$,

$$\sum_{j=a}^{\infty} \lambda_j^{\max} < n.$$

Let $b' = \frac{1}{\alpha}\log(\frac{c}{1-e^{-\alpha}})$.

$$\begin{aligned}
\hat{R}(\Lambda_{ce^{-\alpha\cdot}}) &\geq \sum_{j=b'}^{b} \log(\sqrt{\lambda_j^{\max}/4}) \\
&\geq \frac{1}{2}\sum_{j=b'}^{b} \log(cne^{-\alpha j}) - b\log 4 \\
&\geq \frac{1}{2}\log((cn)^b e^{-\alpha b(b+1)/2)}) - \mathcal{O}(\log n) \\
&\geq \frac{1}{2}\frac{(b-1)}{4}\log(cn) - \mathcal{O}(\log n) \\
&\geq \frac{\log^2 n}{4\alpha} - \mathcal{O}(\log n).
\end{aligned}$$

□

Acknowledgement

Chapter 5 is partially adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Poissonization and universal compression of envelope classes", submitted to *IEEE International Symposium on Information Theory*, 2014. The dissertation author was a primary researcher and author of this paper.

Chapter 6

Pattern redundancy - Tight bounds

6.1 Introduction

We recap some of the definitions and motivations for pattern based compression and other applications of patterns.

Definition 42. *The* pattern *of a sequence* $x_1^n \overset{\text{def}}{=} x_1 \ldots x_n$*, denoted* $\overline{\psi}(x_1^n)$ *is the integer sequence obtained by replacing each symbol in* x_1^n *by the number of distinct symbols up to (and including) its first appearance.*

The pattern of the length-11 sequence *abracadabra* over the english letters, *i.e.,* $\mathcal{X} = \{a, b, \ldots, z\}$, is 12314151231. The pattern of the 5 word phrase *to be or not to be*, where all possible english words is the underlying alphabet, is $\overline{\psi}$(*to be or not to be*) = 123412. A sequence can be described by encoding its pattern and the *dictionary* separately. Such a coding scheme was proposed by [14, 15]. For example, one can encode the sequence *abracadabra* by first compressing 12314151231 and then conveying the dictionary as $1 \to a, 2 \to b, 3 \to r, 4 \to c, 5 \to d$.

Patterns capture all the structural information present in a sequence, disregarding the meaning of individual symbols. However, since many sequences map to the same pattern, by the function redundancy lemma, patterns of sequences generated by a source can be compressed with smaller redundancy than sequences themselves.

6.2 Pattern probability and redundancy

Let Ψ^n be the set of all possible patterns of length n. [15] showed a bijection from Ψ^n to all partitions of a set with n elements. This shows that $|\Psi^n| = B_n$, the nth Bell number. We consider the induced distributions on Ψ^n when length$-n$ sequences are generated *i.i.d.*. Let P be a distribution over an underlying alphabet $\mathcal{X}$. The probability of a sequence $x_1^n \in \mathcal{X}^n$ under P is $P^n(x_1^n) \stackrel{\text{def}}{=} \prod_{i=1}^n P(x_i)$. The probability of a pattern is

$$P^n(\overline{\psi}) \stackrel{\text{def}}{=} \sum_{x_1^n : \overline{\psi}(x_1^n) = \overline{\psi}} P^n(x_1^n),$$

the probability of observing a sequence with pattern $\overline{\psi}$. For example, the probability of the pattern 1232 under distribution P over $\mathcal{X} = \{A, B, \ldots, Z\}$ is

$$P^3(1232) = P^3(ABCB) + P^3(ABDB) + \ldots + P^3(ZYXY).$$

Let $\mathcal{I}^n$ be the class of all length$-n$ *i.i.d.* distributions over any discrete alphabet of any size. Let $\mathcal{I}_\Psi^n$ denote the class of all distributions induced on Ψ^n by distributions in $\mathcal{I}^n$ *i.e.,*

$$\mathcal{I}_\Psi^n = \{P' : P'(\overline{\psi}) = P^n(\overline{\psi}) \text{ where } P^n \in \mathcal{I}^n\}.$$

The redundancy of $\mathcal{I}_\Psi^n$ is

$$\begin{aligned}\overline{R}(\mathcal{I}_\Psi^n) =& \inf_Q \sup_{P \in \mathcal{I}_\Psi^n} D(P||Q),\\ \hat{R}(\mathcal{I}_\Psi^n) =& \inf_Q \sup_{P \in \mathcal{I}_\Psi^n} \sup_{\overline{\psi} \in \Psi^n} \log \frac{P(\overline{\psi})}{Q(\overline{\psi})}.\end{aligned}$$

Note that here the infimum Q is over all possible distributions on Ψ^n.

6.3 Related work and known results

The redundancy of patterns was considered in [15] who show that the worst case redundancy of patterns is $\mathcal{O}(\sqrt{n})$ and at least $\Theta(n^{1/3})$ and therefore patterns of *i.i.d.* distributions have diminishing per-symbol redundancy and they are universally compressible. A conclusion of this is that the structure of sequences can be compressed efficiently, and for large alphabets, and almost all the redundancy is in compressing the dictionary. In [56] the authors provide a different analysis to improve the constant in the lower bound. [57] consider the problem of finding the average redundancy of patterns of *i.i.d.* sequences. Restricting to the class of *i.i.d.* distributions with at most k symbols, they show that for $k \le n^{1/3}$, $\overline{R}(\mathcal{I}_\Psi^n)$ grows at least linearly with k, and extending to all k, they prove that for any fixed $\epsilon \ge 0$, $\overline{R}(\mathcal{I}_\Psi^n) \ge n^{1/3-\epsilon}$ for large n. Similarly, they prove that for $k \le \sqrt{n}$, $\overline{R}(\mathcal{I}_\Psi^n)$ grows at most linearly with k. [24] improved the upper bound on average redundancy to $n^{0.4}$, still leaving a large gap between the known lower and upper bounds. The lower bound on average redundancy was improved from $n^{1/3-\epsilon}$ by Garivier [58], who showed that $\overline{R}(\mathcal{I}_\Psi^n) \ge 1.84 \left(\frac{n}{\log n}\right)^{1/3}$.

Combining these results, the best known bounds on pattern redundancy before this work can be surmised in the following equation.

$$1.84\left(\frac{n}{\log n}\right)^{1/3} \le \overline{R}(\mathcal{I}_\Psi^n) \le \mathcal{O}(n^{0.4}) \tag{6.1}$$

$$\frac{3}{2}n^{1/3} < \hat{R}(\mathcal{I}_\Psi^n) < \left(\pi\sqrt{\frac{2}{3}}\right)n^{1/2}. \tag{6.2}$$

We determine the exact growth exponent of both average and worst case pattern redundancy and prove that

$$0.5n^{1/3} \le \overline{R}(\mathcal{I}_\Psi^n) \le 3n^{1/3}\log^{4/3} n \tag{6.3}$$

$$\hat{R}(\mathcal{I}_\Psi^n) < 110n^{1/3}\log^{5/3} n. \tag{6.4}$$

We now define profiles of sequences and show that they are a sufficient statistic of the pattern, much the same way as *type* is a sufficient statistic of a sequence.

6.4 Profiles

Definition 43. *The* profile *of a sequence x_1^n, denoted $\bar{\varphi}(x_1^n)$ is the multiset of the multiplicities of all symbols appearing in it.*

The profile of the length-4 strings *room*, *abac*, *isit* are all $\{1, 1, 2\}$, meaning that there is a symbol that appears twice and two symbols that appear once in each of them.

Let Φ^n be the set of all possible patterns of length n. [15] showed a bijection from Ψ^n to all partitions of an integer n. This shows that $|\Psi^n| = p_n$, the nth partition number. The upper bound in Equation (6.2) was derived using the bound of $\exp(O(\sqrt{n}))$ on p_n.

The probability of a profile $\bar{\varphi} \in \Phi^n$ under P is the probability that an *i.i.d.* sequence generated according to P has profile $\bar{\varphi}$, *i.e.,*

$$P^n(\bar{\varphi}) \overset{\text{def}}{=} \sum_{\overline{x}:\bar{\varphi}(x_1^n)=\bar{\varphi}} P^n(x_1^n).$$

As with patterns, let $\mathcal{I}_\Phi^n = \{P' : P'(\bar{\varphi}) = P^n(\bar{\varphi})$ where P is any discrete distribution$\}$, denote all induced distributions on profiles via *i.i.d.* sequences, and

$$\begin{aligned}\overline{R}(\mathcal{I}_\Phi^n) =& \inf_Q \sup_{P\in\mathcal{I}_\Phi^n} D(P||Q),\\ \hat{R}(\mathcal{I}_\Psi^n) =& \inf_Q \sup_{P\in\mathcal{I}_\Phi^n} \sup_{\bar{\varphi}\in\Phi^n} \log \frac{P(\bar{\varphi})}{Q(\bar{\varphi})}.\end{aligned}$$

It is easy to see that patterns are functions of sequences and profiles are functions of patterns. For *i.i.d.* distributions, profiles are particularly interesting due to the following observation. Its states that under *i.i.d.* sampling, profiles are a sufficient statistic for the patterns.

Lemma 44 ([15])**.** *For a distribution P and two patterns $\overline{\psi}_1$ and $\overline{\psi}_2$ with the same profile*

$$P^n(\overline{\psi}_2) = P^n(\overline{\psi}_1).$$

Proof. For *i.i.d.* sampling, the probability of any sequence is unchanged under any permutation of its symbols. For example, when $n = 4$, $P(isit) = P(itis) =$

$P(siit) = \cdots$. Using this, for any two patterns $\overline{\psi}_1$ and $\overline{\psi}_2$ with the same profile, there is a bijection between sequences with pattern $\overline{\psi}_1$ and those with pattern $\overline{\psi}_2$ such that each sequence is mapped to a sequence that can be obtained by permuting its symbols, and hence has the same probability. Summing over all sequences proves the result. □

Therefore, if f is a function that maps a pattern to a profile, then it obeys Equation (3.1), and by applying Lemma 12 proves

Lemma 45. $\overline{R}(\mathcal{I}^n_\Psi) = \overline{R}(\mathcal{I}^n_\Phi)$ *and* $\hat{R}(\mathcal{I}^n_\Psi) = \hat{R}(\mathcal{I}^n_\Phi)$.

We therefore consider only the redundancy of profiles.

In a number of problems in machine learning, profiles are a sufficient statistic. For example, learning the support size, estimating the entropy, etc the individual symbols do not matter. Such properties are *symmetric properties* of distributions [16]. For the problems of classification and closeness testing the notion of *joint profile* can be defined [59, 60, 21, 22]. Using joint profiles, [21, 22] prove results on classification and closeness testing that are independent of the underlying alphabet size $|\mathcal{X}|$. These results are along the lines of the results of pattern redundancy being sublinear.

For a discrete distribution P let $\mathcal{M}(P)$ denote the multiset of probability values in P. For example, a distribution on a loaded die given by $P(1) = 3/21, P(2) = 5/21, P(3) = 1/21, P(4) = 5/21, P(5) = 2/21, P(6) = 5/21$ has $\mathcal{M} = \{1/21, 2/21, 3/21, 5/21, 5/21, 5/21\}$. Any permutation of these values is a distribution on a die with the same $\mathcal{M}$. Since patterns are obtained by striping off the symbol labels and keeping only the ordering, we obtain

Lemma 46. *If* $\mathcal{M}(P) = \mathcal{M}(Q)$, *then for any pattern* $\overline{\psi}$, $P^n(\overline{\psi}) = Q^n(\overline{\psi})$ *and for any profile* $\bar{\varphi}$, $P^n(\bar{\varphi}) = Q^n(\bar{\varphi})$.

Proof. Since P and Q have the same multiset of probabilities there is a bijection between their supports. Then an argument similar to that of Lemma 44 shows the result. □

Therefore, any distribution in $\mathcal{I}^n_\Phi$ can be described by its multiset.

6.5 Alternate profile probability via Poissonization

Similar to the bounds on redundancy of sequences in Chapter 5, we consider the redundancy of profiles generated via Poisson sampling. $\mathcal{I}_\Phi^n$ is the class of induced distributions on $\Phi^n$by all *i.i.d.* distributions over all support sizes, when sampled n times. Let $\mathcal{I}_\Phi^{\text{poi}(n)}$ be the class of induced distributions on Φ^* by all *i.i.d.* distributions, when sampled poi(n) times.

Consider the following method to generate a profile in Φ^* from a distribution P with $\mathcal{M}(P) = \{p_1, \ldots\}$. For an integer n, let $\lambda_i = np_i$, then $\sum \lambda_i = n$. Generate $\mu_i \sim \text{poi}(\lambda_i)$ independently. Output the multiset of non-negative μ_i's.

Using Lemma 6 it follows that the distribution above is the same as the distribution induced by $P^{\text{poi}(n)}$ over Φ^*.

Therefore any distribution in $\mathcal{I}_\Phi^{\text{poi}(n)}$ can be described as a multiset $\overline{\Lambda} = \{\lambda_1 \overset{\text{def}}{=} np_1, \lambda_2 \overset{\text{def}}{=} np_2, \ldots\}$. The profile generated by Poisson sampling is a multiset $\bar{\varphi} = \{\mu_1, \mu_2, \ldots\}$, where each μ_i generated independently according to poi(λ_i). The probability that Λ generates $\bar{\varphi}$ is [21, 15],

$$\Lambda(\bar{\varphi}) = \frac{1}{\prod_{\mu=0}^{\infty} \varphi_\mu!} \sum_\sigma \prod_i \text{poi}(\lambda_{\sigma(i)}, \mu_i), \tag{6.5}$$

where φ_μ is the number of appearances of μ in $\bar{\varphi}$.

6.5.1 Profile redundancy bounds

We now relate the redundancy of profiles under Poisson sampling to the redundancy under sampling exactly n times.

To prove our lower bound on $\overline{R}(\mathcal{I}_\Phi^n)$, we use Lemma 9. Therefore, it suffices to provide a good lower bound on $M(\mathcal{I}_\Phi^n, \delta)$. We do so by proving the following lemma, and then bounding $M(\mathcal{I}_\Phi^{\text{poi}(n)}, \delta)$, whose calculation is easier owing to independence.

Lemma 47.

$$M(\mathcal{I}_\Phi^{\text{poi}(n-\sqrt{n\log n})}, \delta) \le M(\mathcal{I}_\Phi^n, 2\delta).$$

Proof. Consider the set of distributions that give rise to $M \stackrel{\text{def}}{=} M(\mathcal{I}_\Phi^{\text{poi}(n-\sqrt{n\log n})}, \delta), \delta)$ distributions. By Lemma 4 for large n,

$$\Pr\left(\text{poi}\left(n - \sqrt{n\log n}\right) \geq n\right) \leq \delta.$$

Therefore, these M distributions generate a profile of length at most n with probability $\geq 1-\delta$. Given profiles of length n generated by a distribution we can generate profiles of any length $n' \leq n$ from the same distribution. Therefore these set of distributions are also distinguishable over length-n profiles with probability of error at most 2δ. □

For the upper bound, we can show a result analogous to Theorem 35 for profiles. Using monotonicity of $\mathcal{I}_\Phi^n$ with n, and the fact that the median of a Poisson random variable is close to its mean, we obtain

Lemma 48. *For $R \in \{\overline{R}, \hat{R}\}$,*

$$R(\mathcal{I}_\Psi^n) \leq R(\mathcal{I}_\Psi^{\text{poi}(n)}) + 1.$$

These results imply that we can only consider Poisson-sampled profiles.

6.6 Lower bound on average pattern redundancy

We now lower bound $M(\mathcal{I}_\Phi^{\text{poi}(n)}, \delta)$, for some $\delta > 0$. We will construct a class of *distinguishable distributions* $\overline{\Lambda}_1, \ldots, \overline{\Lambda}_M$, each in $\mathcal{I}_\Phi^{\text{poi}(n)}$, such that there exist disjoint $\Phi_j \subset \Phi^*$, with $\overline{\Lambda}_i(\Phi_i) > 1-\delta$. Maximizing the size of M yields a lower bound on $\overline{R}(\mathcal{I}_\Phi^{\text{poi}(n)})$.

We now describe our class of distributions. For notational simplification, we use n instead of $n - \sqrt{n}\log n$. Recall that a distributionin $\mathcal{I}_\Phi^{\text{poi}(n)}$ is a collection of Poisson parameters that sum to n, namely each $\overline{\Lambda}_i$ is a collection of positive reals that sum to n. Let $C > 0$ and $\lambda_i \stackrel{\text{def}}{=} Ci^2$, where the value of C will be chosen later.

Let

$$\mathcal{S} = \{\lambda_i : 1 \leq i \leq K\},$$

where $K = \lfloor (3n/C)^{1/3} \rfloor$ ensures that

$$\sum_{\lambda \in \mathcal{S}} \lambda \leq n.$$

For a binary string $\overline{x} = x_1 x_2 \ldots x_K$, let

$$\overline{\Lambda}_{\overline{x}} = \{\lambda_i^* : x_i = 1\} \cup \left\{ n - \sum \lambda_i^* x_i \right\}.$$

This maps every length-K binary string to a distribution in $\mathcal{I}_\Phi^{\text{poi}(n)}$. The distribution adds a λ_i whenever $x_i = 1$, and the final element is added to ensure that the sum of elements in the distribution is n. We consider a subset of these 2^K distributions that correspond to a binary code.

Let $\mathcal{C}(K, \alpha K)$ be a code satisfying Lemma 5. For this code, let

$$\mathcal{L} = \{\overline{\Lambda}_{\overline{c}} : \overline{c} \in \mathcal{C}\},$$

be the collection of distributions generated by the codewords in $\mathcal{C}(K, \alpha K)$. Then,

$$|\mathcal{L}| = |\mathcal{C}(K, \alpha K)| \geq 2^{(\frac{3}{C})^{1/3}(1-h(\alpha))n^{1/3}}. \tag{6.6}$$

We now show that for suitable choices of C and α, and large K, the distributions in $\mathcal{L}$ are $\delta-$distinguishable (see Section 3.1), for some δ that approaches 0 as K grows by proving the following theorem.

Theorem 49. *Let $C > 0$, and $\frac{1}{2} > \alpha > 2.01e^{-C/2}$. Then, the set of distributions in $\mathcal{L}$ are $\delta-$distinguishable, and K large.*

Proof. The set $\mathcal{L}$ consists of δ-distinguishable distributions if there is a map $f : \Phi^* \to \mathcal{L}$, such that for any $\overline{\Lambda} \in \mathcal{L}$, if $\bar{\varphi} \sim \overline{\Lambda}$, then

$$P(f(\bar{\varphi}) \neq \overline{\Lambda}) < \delta. \tag{6.7}$$

The map f is constructed as follows. We first describe a function from Φ^* to $\{0,1\}^K$, and then map the string to the codeword with the least Hamming distance from it.

Formally, let $\bar{\varphi} = \{\mu_1, \mu_2, \ldots\}$ be a profile. For each $j = 1, 2, \ldots, K$, let

$$x_i = \begin{cases} 1 & \text{if } \exists j \text{ such that } i = \arg\min_r |\mu_j - \lambda_r| \\ 0 & \text{otherwise.} \end{cases}$$

In other words, for each multiplicity μ_j if λ_i is closest to μ_j we set x_i to 1. Let $\overline{x}(\bar{\varphi}) = x_1 \ldots x_K$ be the sequence generated by this process. Let $\hat{c}(\bar{\varphi}) \in \mathcal{C}$ be the code with minimum Hamming distance from $\overline{x}(\bar{\varphi})$. Let

$$f(\bar{\varphi}) = \overline{\Lambda}_{\hat{c}}.$$

We now pick any distribution $\Lambda_{\overline{c}}$ and and find conditions unver which Equation (6.7) holds.

Two adjacent λ's are separated by

$$\Delta_i \overset{\text{def}}{=} \lambda_{i+1} - \lambda_i = C(i+1)^2 - Ci^2 = (2i+1)C > 2\sqrt{C\lambda_i}. \tag{6.8}$$

Let $\lambda_i \in \overline{\Lambda}_{\overline{c}}$ be any element. Let Y_i be a random variable that is 1 if the multiplicity μ_i generated by λ_i is closest to a λ_j, $j \neq i$ and 0 otherwise. Since the minimum distance of the code is αK, the probability of error is at most the probability that $2\sum Y_i \geq \frac{\alpha K}{2}$. Thus, for $\bar{\varphi} \sim \overline{\Lambda}_{\overline{c}}$

$$P(f(\bar{\varphi}) \neq \overline{\Lambda}_{\overline{c}}) \leq P\left(\sum Y_i \geq \frac{\alpha K}{4}\right).$$

Applying this Lemma to Equation (6.8)

$$\begin{aligned} P(Y_i = 1) \leq & P\left(\text{poi}(\lambda_i) \leq \lambda_i - \frac{(\lambda_i - \lambda_{i-1})}{2}\right) + P\left(\text{poi}(\lambda_i) \leq \lambda_i + \frac{(\lambda_{i+1} - \lambda_i)}{2}\right) \\ \leq & \exp\left(-\frac{(C(2i-1))^2}{2Ci^2}\right) + \exp\left(-\frac{((2i+1)C)^2}{2(Ci^2+Ci)}\right) < 2\exp(-C/2). \end{aligned}$$

Without loss of generality, assume that at least half of the codewords in $\mathcal{C}(K, \alpha K)$ have weight at most $K/2$. If not we can take the complement of each codeword, and still maintain the same minimum distance, but ensuring that the new code has at least half the codewords of weight at most $K/2$. So, we throw away the distributions that have more than $K/2$ elements (at most half the distributions satisfy this) So,

$$\mathbb{E}\left[\sum Y_i\right] = \sum P(Y_i = 1) \leq 2e^{-C/2}\frac{K}{2} = e^{-C/2}K.$$

By the independence of Y_i's,

$$\sigma^2 \overset{\text{def}}{=} V\left[\sum Y_i\right] = \sum_i V[Y_i] \leq e^{-C/2}K.$$

By the Chebychev's Inequality,

$$\begin{aligned} P(f(\bar{\varphi}) \neq \overline{\Lambda}_{\bar{c}}) &\leq P\left(\sum Y_i \geq \frac{\alpha K}{4}\right) \\ &\leq P\left(\left|\sum Y_i - e^{-C/2}K\right| \geq \frac{(\alpha - 2e^{-C/2})K}{2}\right) \\ &\leq \frac{e^{-C/4}}{K(\alpha - 2e^{-C/2})^2}. \end{aligned}$$

When the conditions of the theorem are satisfied, and K grows, error probability goes to 0. □

The value of $\log M(\mathcal{I}_\Phi^{\text{poi}(n)}, \delta)$ is at most the largest value attained by $\left(\frac{3}{C}\right)^{\frac{1}{3}}(1-h(\alpha)n^{1/3}$, subject to $\exp(-C/4) < \alpha < \frac{1}{2}$. When $C = 19$, a simple calculation shows that $\log M(\mathcal{I}_\Phi^{\text{poi}(n)}, \delta) > 0.5n^{1/3}$. Applying this result to Lemma 47 and Lemma 9 yields the lower bound.

6.7 Upper bound on average pattern redundancy

In this section we upper bound the average redundancy of patterns under *i.i.d.* sampling. The bound on the worst-case is obtained via a more refined analysis and is presented later. Even though the basic principle behind proving both the results are similar, the arguments that hold for an expected profile do not hold over all profiles. We circumvent this problem by showing that it suffices to consider a smaller class of profiles, for which a result similar to the average case holds.

Let $\overline{\mu} = (\mu_1, \dots, \mu_m)$, where $\mu_i \sim \text{poi}(\lambda_i)$ are independent Poisson random variables. Any collection of m λ_i's induces a distribution on m-tuples of non-negative integers. Let

$$\mathcal{I}(m, (\lambda_0, \lambda_0 + \Delta]) \stackrel{\text{def}}{=} \{(\text{poi}(\lambda_1), \dots, \text{poi}(\lambda_m)) : \lambda_0 < \lambda_i \leq \lambda_0 + \Delta\}$$

be the set of all products of m Poisson distributions, with each parameter in the interval $(\lambda_0, \lambda_0 + \Delta]$. In order to generate a profile from a distribution, we simply consider the multiset $\{\mu_1, \dots, \mu_m\}$. We first prove an upper bound on the redundancy of this class, and since profiles are functions of the multiplicity tuple, we obtain an upper bound on the profile redundancy.

Let $\overline{\Lambda} = (\lambda_1, \ldots, \lambda_m)$ and $\overline{\Lambda}' = (\lambda'_1, \ldots, \lambda'_m)$ be any two distributions in $\mathcal{I}(m, (\lambda_0, \lambda_0 + \Delta])$. For a profile $\bar{\varphi} \in \Phi^*$, let $\overline{\Lambda}(\bar{\varphi})$ be probability of observing profile $\bar{\varphi}$, a function of $\overline{\mu}$. Let $\mathcal{I}_\Phi(m, (\lambda_0, \lambda_0 + \Delta])$ denote this induced class of distributions over profiles. Since KL-divergence adds up on product spaces, for any two distributions

$$\begin{aligned}
D\Big(\overline{\Lambda}(\bar{\varphi})||\overline{\Lambda}'(\bar{\varphi})\Big) &\overset{(a)}{\le} D\Big(\overline{\Lambda}(\overline{\mu})||\overline{\Lambda}'(\overline{\mu})\Big) && (6.9)\\
&= \sum_{j=1}^{m} D(\mathrm{poi}(\lambda_j)||\mathrm{poi}(\lambda'_j)) && (6.10)\\
&\overset{(b)}{=} \sum_{j=1}^{m} \lambda'_j - \lambda_j + \lambda_j \log\left(\frac{\lambda_j}{\lambda'_j}\right) && (6.11)\\
&\overset{(c)}{\le} \sum_{j=1}^{m} \lambda'_j - \lambda_j + \lambda_j \Big(\frac{\lambda_j}{\lambda'_j} - 1\Big) && (6.12)\\
&= \sum_{j=1}^{m} \frac{(\lambda'_j - \lambda_j)^2}{\lambda'_j} && (6.13)\\
&\overset{(d)}{\le} m\frac{\Delta^2}{\lambda_0}, && (6.14)
\end{aligned}$$

where (a) uses the following form of Data-processing inequality. For any function f, and random variables X and Y,

$$D(f(X)||f(Y)) \le D(X||Y),$$

which follows from the convexity of logarithms. The function f simply maps $(\mu_1, \ldots, \mu_m)$ to $\{\mu_1, \ldots, \mu_m\}$. (b) is the expression of KL divergence between Poisson distributions, and (c) uses $\log x \le x - 1$, the (d) from the fact that all λ's are in an interval of width Δ, starting at λ_0. Combining this result with the definition of redundancy bounds the redundancy of $\mathcal{I}(m, (\lambda_0, \lambda_0 + \Delta])$ yields

Lemma 50.

$$\overline{R}\Big(\mathcal{I}_\Phi(m, (\lambda_0, \lambda_0 + \Delta])\Big) \le \overline{R}\Big(\mathcal{I}(m, (\lambda_0, \lambda_0 + \Delta])\Big) \le \frac{m\Delta^2}{\lambda_0}.$$

Note that the right hand side of this bound reduces with Δ, namely distributions with *close* parameters have small redundancy. We will therefore divide the interval $(0, n]$ into intervals. We define a partition of $\mathcal{I}_{\Phi}^{\text{poi}(n)}$ based on the choice of intervals. Using Lemma 17, we show that considering one of the classes suffices. We then apply Lemma 50 over all the intervals. Optimizing over the choice of intervals we obtain the upper bound.

6.7.1 Construction of distribution classes

Any collection of positive reals $\overline{\Lambda} \stackrel{\text{def}}{=} \{\lambda_1, \lambda_2, \ldots\}$ is defines a distribution over Φ^*, where the profile generated is the collection of poi(λ_i) random variables. By Lemma 48, we bound the redundancy of

$$\mathcal{I}_{\Phi}^{\text{poi}(n)} = \Big\{\overline{\Lambda} : \sum_{\lambda \in \overline{\Lambda}} \lambda = n\Big\},$$

the collection of all distributions whose elements add up to n.

Let b be a positive integer (specified later). Consider any partition of $(0, n]$ into $b+1$ consecutive intervals $I_0, I_1, I_2, \ldots, I_b$ of lengths $\Delta_0 = 1, \Delta_1, \Delta_2, \ldots, \Delta_b$. In other words, the first interval is $(0, 1]$. We will be mostly interested in the b intervals $I_1, \ldots, I_b$.

For any $\overline{\Lambda} \in \mathcal{I}_{\Phi}^{\text{poi}(n)}$,

- For $j = 0, 1, \ldots, b$, $\overline{\Lambda}_j \stackrel{\text{def}}{=} \overline{\Lambda} \cap I_j$ be the multiset of elements of $\overline{\Lambda}$ in I_j,
- For $j = 1, 2, \ldots, b$, let $m_j \stackrel{\text{def}}{=} m_j(\overline{\Lambda}) \stackrel{\text{def}}{=} |\overline{\Lambda}_j|$ be the number of elements of $\overline{\Lambda}$ in I_j, and $\overline{m} \stackrel{\text{def}}{=} \overline{m}(\overline{\Lambda})$ is the b-tuple of m_j's.

Two distributions have same $\overline{m}$ if they have the same number of λ's in $I_1, \ldots, I_b$. A partition of $(0, n]$ induces a partition of $\mathcal{I}_{\Phi}^{\text{poi}(n)}$ via $\overline{m}$, *i.e.,* distributions having same $\overline{m}$ are in the same class. Let $\mathcal{I}_{\Phi}^{\text{poi}(n)}(i)$, $i = 1, \ldots, T_n$ be these classes of distributions. Since $\Delta_0 = 1$, and sum of elements in a distribution in $\mathcal{I}_{\Phi}^{\text{poi}(n)}$ is n, this shows that $m_j \le n$, for all $j = 1, \ldots, m$. Thus, $\overline{m}$ is a $b-$tuple of non-negative integers each at most n. Therefore,

$$T_n \le n^b.$$

Combining with Lemma 17,

$$\overline{R}(\mathcal{I}_{\Phi}^{\text{poi}(n)}) \leq \max_{1\leq i\leq T_n} \overline{R}(\mathcal{I}_{\Phi}^{\text{poi}(n)}(i)) + b\log n. \tag{6.15}$$

6.7.2 Bounding redundancy of each class

Let λ_j^* denote the start point of interval I_j, *i.e.,* $I_j = (\lambda_j^*, \lambda_{j+1}^*]$, and $\Delta_j = \lambda_{j+1}^* - \lambda_j^*$. Let $\mathcal{I}$ be one of the T_n classes, with $\overline{m} = (m_1, \ldots, m_b)$. We bound the redundancy of $\mathcal{I}$ by proving the following theorem.

Theorem 51. *For any class $\mathcal{I}$,*

$$\overline{R}(\mathcal{I}) \leq \sum_{j=1}^{b} m_j \frac{\Delta_j^2}{\lambda_j^*} + 2\log^2 n.$$

Proof. For any distribution $\overline{\Lambda} \in \mathcal{I}_{\Phi}^{\text{poi}(n)}$, let $\bar{\varphi}_j$ denote the profile generated by $\overline{\Lambda}_j$, where recall that $\overline{\Lambda}_j$ is $\overline{\Lambda} \cap I_j$. By the independence of Poisson sampling, $\bar{\varphi}_j$'s are all independent. Moreover, the profile $\bar{\varphi} = \bar{\varphi}_0 \cup \ldots \cup \bar{\varphi}_b$, a function of the tuple $(\bar{\varphi}_0, \ldots, \bar{\varphi}_b)$. For $j = 1, \ldots, b$

$$\mathcal{I}_j \overset{\text{def}}{=} \{\overline{\Lambda}_j : \overline{\Lambda} \in \mathcal{I}\}$$

be the class of marginal distributions in interval j.

By definition, $\mathcal{I}_{\Phi}(m_j, (\lambda_j^*, \lambda_{j+1}^*])$ consists of all possible distributions with m_j elements in $(\lambda_j^*, \lambda_{j+1}^*]$. Therefore, $\mathcal{I}_j \subseteq \mathcal{I}_{\Phi}(m_j, (\lambda_j^*, \lambda_{j+1}^*])$. Using Lemma 17 and the independence of $\bar{\varphi}_j$'s, and Lemma 50,

$$\overline{R}(\mathcal{I}_{\Phi}^{\text{poi}(n)}) \leq \sum_{j=0}^{b} \overline{R}(\mathcal{I}_j) \leq \overline{R}(\mathcal{I}_0) + \sum_{j=1}^{b} \overline{R}(\mathcal{I}_j) \leq \overline{R}(\mathcal{I}_0) + \sum_{j=1}^{b} m_j \frac{\Delta_j^2}{\lambda_j^*}.$$

Any distribution in $\mathcal{I}_0$ is a collection of λ's that are all ≤ 1 and with sum most n. To bound $\overline{R}(\mathcal{I}_0)$, we provide a single explicit encoding of profiles with expected length $\leq 5\log^2 n$ for all distributions in $\mathcal{I}_0$. The largest expected codelength of any code over a class of distributions is clearly an upper bound on average redundancy. One way to describe a collection of multiplicities is to describe the set of distinct multiplicities, with their respective number appearances. We will describe all positive integers using Elias coding. Elias Codes [61] are prefix free coding scheme over positive integers that uses $2\lfloor\log\mu\rfloor + 1$ bits to represent μ. Using concavity of logarithms, a random variable X can be encoded with at most $2\log\mathbb{E}[X] + 1$ expected bits. Let $\mu_{\max}$ be the largest multiplicity generated by a distribution $\{\lambda_1, \lambda_2, \ldots\}$, where $\lambda_j \le 1$, and $\sum\lambda \le n$. For any $j \ge 1$,

$$P(\mu_{\max} \ge j) \overset{(a)}{\le} \sum_{\lambda_i} e^{-\lambda_i}\left(\frac{e\lambda_i}{j}\right)^j \overset{(b)}{\le} \left(\frac{e}{j}\right)^j \sum_{\lambda_i} \lambda_i^j \overset{(c)}{\le} \left(\frac{e}{j}\right)^j \cdot n,$$

where (a) follows by the union bound, (b) uses $\exp(-\lambda_i) \le 1$, and (c) uses $\lambda_j \le 1$ and $j \ge 1$. Now, for $j > \log n$ the probability falls exponentially with j. This combined with the fact that the number of times a multiplicity appears is at most the number of total samples, which is $\text{poi}(n)$ shows that Elias code has expected codelength at most $\log^2 n$ for any distribution in $\mathcal{I}_0$.

□

6.7.3 Final bound

We now use Theorem 51 with Equation 6.15. The average redundancy is bounded by,

$$\sum_{j=1}^{b} \frac{m_j\Delta_j^2}{\lambda_j^*} + \log^2 n + b\log n,$$

for any choice of intervals.

We choose the intervals in a geometric progression. The start of the first interval is 1, and the ending point of the bth interval is n. The starting point of the jth interval is $(1+c)^{j-1}$, for $j = 1, \ldots, b$, where $1+c$ is the parameter of the geometric series. Then,

$$(1+c)^b = n.$$

Thus, the jth interval has length,

$$\Delta_j = (1+c)^j - (1+c)^{j-1} = c(1+c)^{j-1} = c\lambda_j^*.$$

Plugging these $\overline{R}(\mathcal{I}_\Phi^{\text{poi}(n)})$ can be upper bounded by,

$$\sum_{j=1}^{b} \frac{m_j\Delta_j^2}{\lambda_j^*} + b\log n + \log^2 n = c^2 \sum_{j=1}^{b} m_j\lambda_j^* + b\log n + \log^2 n \le c^2 n + b\log n + \log^2 n,$$

where we used the fact that the sum of all elements of a distribution is at most n, thus $\sum_j m_j \lambda_j^* \le n$. Now, $(1+c)^b = n$, therefore, $b \log(1+c) = \log n$. We are interested in the case where b is a polynomial of n, and hence c is small. Thus for any $\epsilon > 0$, as n grows,

$$c \le (1+\epsilon)\frac{\log n}{b}.$$

Plugging this,

$$\begin{aligned}\overline{R}(\mathcal{I}_\Phi^n) &\le \overline{R}(\mathcal{I}_\Phi^{\mathrm{poi}(n)}) + \log n \\ &\le (1+\epsilon)^2 \frac{n \log^2 n}{b^2} + b \log n + \log^2 n + \log n \\ &\le 3n^{1/3} \log^{4/3} n,\end{aligned}$$

by choosing $b = n^{1/3} \log^{1/3} n$.

6.8 Upper bound on worst-case pattern redundancy

6.8.1 Overview

Any collection of non-negative reals $\{\lambda_1, \ldots\}$ implies a distribution over Φ^* by Poisson sampling. From each distribution in $\mathcal{I}_\Phi^{\mathrm{poi}(n)}$, we generate three distributions, the collection of λ's that are at most $n^{1/3}$, those between $n^{1/3}$ and H, and those larger than H. The value of H will be close to $n^{2/3}$ and specified later. Using independence of the multiplicities generated via Poisson sampling, we show that it suffices to consider distributions that have all the λ's in the middle range, namely in $(n^{2/3}, H]$.

We then partition all such distributions into T_n classes, where $\log(T_n) = b \log n$. Using Lemma 17, we need to consider only one of these classes. Using Lemmas 14 and 11 and some other manipulations over Poisson distributions, we prove that each such class has redundancy $\tilde{\mathcal{O}}(n^{1/3})$.

6.8.2 Details

For any distribution $\overline{\Lambda} \in \mathcal{I}_{\Phi}^{\text{poi}(n)}$, let

$$\begin{aligned}\overline{\Lambda}_{\text{low}} &\stackrel{\text{def}}{=} \{\lambda \in \overline{\Lambda} : \lambda \le n^{1/3}\},\\ \overline{\Lambda}_{\text{med}} &\stackrel{\text{def}}{=} \{\lambda \in \overline{\Lambda} : n^{1/3} < \lambda \le H\},\\ \overline{\Lambda}_{\text{high}} &\stackrel{\text{def}}{=} \{\lambda \in \overline{\Lambda} : \lambda > H\}.\end{aligned}$$

Let $\bar{\varphi}_{\text{low}}, \bar{\varphi}_{\text{med}}, \bar{\varphi}_{\text{high}}$ denote the profiles generated by $\overline{\Lambda}_{\text{low}}, \overline{\Lambda}_{\text{med}}, \overline{\Lambda}_{\text{high}}$ respectively. Thus, $\overline{\Lambda} = \overline{\Lambda}_{\text{low}} \cup \overline{\Lambda}_{\text{med}} \cup \overline{\Lambda}_{\text{high}}$, and any $\bar{\varphi}$ generated by a distribution in $\mathcal{I}_{\Phi}^{\text{poi}(n)}$ is of the form $\bar{\varphi}_{\text{low}} \cup \bar{\varphi}_{\text{med}} \cup \bar{\varphi}_{\text{high}}$, and hence a function of the triple $(\bar{\varphi}_{\text{low}}, \bar{\varphi}_{\text{med}}, \bar{\varphi}_{\text{high}})$. Let,

$$\begin{aligned}\mathcal{I}_{\Phi_{\text{low}}}^{\text{poi}(n)} &\stackrel{\text{def}}{=} \Big\{\overline{\Lambda} : \lambda < n^{1/3}\ \forall\lambda,\ \text{and} \sum_{\lambda\in\overline{\Lambda}} \lambda \le n\Big\},\\ \mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)} &\stackrel{\text{def}}{=} \Big\{\overline{\Lambda} : n^{1/3} < \lambda \le H\ \forall\lambda,\ \text{and} \sum_{\lambda\in\overline{\Lambda}} \lambda \le n\Big\},\\ \mathcal{I}_{\Phi_{\text{high}}}^{\text{poi}(n)} &\stackrel{\text{def}}{=} \Big\{\overline{\Lambda} : \lambda > H\ \forall\lambda \in \overline{\Lambda},\ \text{and} \sum_{\lambda\in\overline{\Lambda}} \lambda \le n\Big\}.\end{aligned}$$

It is each to see that for any distribution in $\mathcal{I}_{\Phi}^{\text{poi}(n)}$, $\overline{\Lambda}_{\text{low}} \in \mathcal{I}_{\Phi_{\text{low}}}^{\text{poi}(n)}$, $\overline{\Lambda}_{\text{med}} \in \mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}$, and $\overline{\Lambda}_{\text{high}} \in \mathcal{I}_{\Phi_{\text{high}}}^{\text{poi}(n)}$. We can now state the following lemma that enables us to concentrate only these classes separately.

Lemma 52.

$$\hat{R}(\mathcal{I}_{\Phi}^{\text{poi}(n)}) \le \hat{R}(\mathcal{I}_{\Phi_{\text{low}}}^{\text{poi}(n)}) + \hat{R}(\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}) + \hat{R}(\mathcal{I}_{\Phi_{\text{high}}}^{\text{poi}(n)})$$

Proof. $\bar{\varphi}$ is a function of the triple $(\bar{\varphi}_{\text{low}}, \bar{\varphi}_{\text{med}}, \bar{\varphi}_{\text{high}})$, which itself is a product random variable by Poisson sampling. Using Lemma 11, and an extension of Lemma 14 to three products gives the result. □

The redundancies of $\mathcal{I}_{\Phi_{\text{low}}}^{\text{poi}(n)}$, and $\mathcal{I}_{\Phi_{\text{high}}}^{\text{poi}(n)}$ are easier to bound compared to $\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}$ and are done first.

Lemma 53.

$$\hat{R}(\mathcal{I}_{\Phi_{\text{low}}}^{\text{poi}(n)}) < 4n^{1/3}\log n,\ \textit{and}\ \hat{R}(\mathcal{I}_{\Phi_{\text{high}}}^{\text{poi}(n)}) < \frac{n}{H}\log n.$$

Proof. The proof is similar to bounding the redundancy of $\mathcal{I}_0$ in the proof of Theorem 51. We encode $\bar{\varphi}_{\text{low}}$ and $\bar{\varphi}_{\text{high}}$ using Elias codes described before. We show the proof for $\mathcal{I}^{\text{poi}(n)}_{\Phi_{\text{high}}}$ and sketch it for $\mathcal{I}^{\text{poi}(n)}_{\Phi_{\text{low}}}$. The number of distinct multiplicities in $\bar{\varphi}_{\text{high}}$ is at most n/H. Now the probability of the largest multiplicity exceeding n falls down exponentially with μ. Therefore as before, the redundancy can be bounded by $n/H \log n$. The proof of $\mathcal{I}^{\text{poi}(n)}_{\Phi_{\text{low}}}$ is similar and is omitted.

□

6.8.3 Bounding $\hat{R}(\mathcal{I}^{\text{poi}(n)}_{\Phi_{\text{med}}})$

We first consider $\mathcal{I}(m, s, (\lambda_0, \lambda_0 + \Delta])$, which is the subset of $\mathcal{I}(m, (\lambda_0, \lambda_0 + \Delta])$ that have $\lfloor (\sum \lambda_i) \times n^{94} \rfloor = s$. Thus, distributions in the such classes have same number of elements and their sums are extremely close (differ by at most $1/n^{94}$).

As with the average case we are interested in finding an upper bound on worst-case redundancy of this class. For average redundancy the KL divergence between distributions had a nice expression that we bound. We are now interested in finding bounds on the sum of maximum likelihood probabilities over all profiles, by distributions in this class. Using Poisson tail bounds we first show that it suffices to consider only a class of all profiles, namely the one's with all multiplicities *around* the interval of interest (*i.e.,* $(\lambda_0, \lambda_0 + \Delta]$). We then bound the Shtarkov sum of this class of profiles.

Consider any profile $\bar{\varphi}$ with all multiplicities in $[\lambda + \Delta/2 - \Theta/2, \lambda + \Delta/2 - \Theta/2]$. Consider any uniform distribution $\overline{\Lambda}' = \{m \times \lambda'\}$ in $\mathcal{I}(m, s, (\lambda_0, \lambda_0 + \Delta])$. Then by definition, $|\sum \lambda_j - m\lambda'| < 1/n^{94}$. We bound the Shtarkov sum of these profiles using the following theorem.

Theorem 54. *For any distribution* $\overline{\Lambda} = \{\lambda_1, \ldots, \lambda_m\}$ *in* $\mathcal{I}(m, s, (\lambda_0, \lambda_0 + \Delta])$*, then for any profile* $\bar{\varphi}$ *with multiplicities in the interval specified,*

$$\frac{\overline{\Lambda}(\bar{\varphi})}{\overline{\Lambda}'(\bar{\varphi})} \leq \sqrt{2} \exp\left[m \Big(\frac{\Delta\Theta}{\lambda_0} \Big)^2 \right].$$

Proof. Let $\bar{\varphi} = \{\mu_1, \dots, \mu_m\}$. By Equation (6.5),

$$\overline{\Lambda}(\bar{\varphi}_j) = N(\bar{\varphi})\frac{\exp(-\sum_{j=1}^{m}\lambda_j)}{\prod \mu_j!}\left(\sum_{\sigma\in S_m}\prod_{l=1}^{m}\lambda_{\sigma(l)}^{\mu_l}\right).$$

Taking the ratio with $\overline{\Lambda}'$,

$$\begin{aligned}\frac{\overline{\Lambda}(\bar{\varphi})}{\overline{\Lambda}'(\bar{\varphi})} &= \exp\Big(-\sum \lambda_j + m\lambda\Big)\cdot\frac{1}{m!}\left(\sum_{\sigma\in S_m}\prod_{l=1}^{m}\left(\frac{\lambda_{\sigma(l)}}{\lambda'}\right)^{\mu_l}\right)\\ &\le \exp\left(\frac{1}{n^{94}}\right)\cdot\frac{1}{m!}\left(\sum_{\sigma\in S_m}\prod_{l=1}^{m}\left(\frac{\lambda_{\sigma(l)}}{\lambda'}\right)^{\mu_l}\right)\end{aligned}$$

Let $\delta_j \overset{\text{def}}{=} \lambda_j - \lambda'$. Let $\mu_{\text{ave}} = \frac{\sum\mu}{m}$ be the average of multiplicities in $\bar{\varphi}$, and $\theta_j \overset{\text{def}}{=} \mu_j - \mu_{\text{ave}}$. Then, $\sum_{l=1}^{m}\theta_l = 0$, and $|\sum_{l=1}^{m_j}\delta_j| < \frac{1}{n^{94}}$. Therefore,

$$\begin{aligned}\frac{1}{m!}\left(\sum_{\sigma\in S_m}\prod_{l=1}^{m}\left(\frac{\lambda_{\sigma(l)}}{\lambda'}\right)^{\mu_l}\right) &= \frac{1}{m!}\left(\sum_{\sigma\in S_m}\prod_{l=1}^{m}\left(1+\frac{\delta_{\sigma(l)}}{\lambda'}\right)^{\mu_l}\right)\\ &\overset{(a)}{\le} \frac{1}{m!}\sum_{\sigma\in S_m}\exp\left(\sum_{l=1}^{m}\frac{\delta_{\sigma(l)}(\mu_{\text{ave}}+\theta_l)}{\lambda'}\right)\\ &\overset{(b)}{\le} \exp\left(\frac{\mu_{\text{ave}}}{n^{94}\lambda'}\right)\cdot\frac{1}{m_j!}\sum_{\sigma\in S_{m_j}}\exp\left(\sum_{l=1}^{m_j}\frac{\delta_{\sigma(l)}\theta_l}{\lambda'}\right),\end{aligned}$$

where (a) uses $1+x\le\exp(x)$ and (b) uses $|\sum\delta_l| \le \frac{1}{n^{94}}$. We now want to bound a function of the form,

$$f(\{\delta_l\},\{\theta_l\}) = \frac{1}{m!}\sum_{\sigma}\left[\exp\left(\sum_{l=1}^{m}\frac{\delta_l\theta_{\sigma_l}}{\lambda'}\right)\right].$$

Now $|\theta_j| \le \Theta$, $|\delta_j| < \Delta$, θ_j's sum to 0 and δ_j's sum to a very small quantity, (less than $1/n^{94}$ in absolute value). Therefore, the convexity of exponential functions we can assume that for m even, the function is maximized when $m/2$ of the θ_j's are Θ and the other $m/2$ are $-\Theta$. Similarly, half the δ_j's are Δ and the other half $-\Delta$. Note that the problem is independent of permutations of the random variables.

We are therefore interested in the following problem (after suitable normalization). Let $x_1, \dots, x_m$ and $y_1, \dots y_m$ are such that half of the x_j's and y_j's are 1 and the remaining half -1. We are interested in $F(\sigma) = \sum_j x_{\sigma(j}y_j$ over all permutations. Since m is even, the value of F is of the form $m-4k$ for $k = 0, 1, \ldots, m/2$. By a simple counting, the number of permutations that lead to the value $m - 4k$ can be shown to be

$$\Big((\frac{m}{2})!\Big)^2 \binom{\frac{m}{2}}{k}^2.$$

Plugging in these values

$$\begin{aligned}
f(\{\delta_l\}, \{\theta_l\}) &\leq \frac{\left(\frac{m}{2}!\right)^2}{m!} \left(\sum_{k=0}^{m/2} \binom{\frac{m}{2}}{k}^2 \exp\left((m-4k)\frac{\Delta\Theta}{\lambda'} \right) \right) \\
&\overset{(a)}{\leq} \frac{\left(\frac{m}{2}!\right)^2 \binom{\frac{m}{2}}{\frac{m}{4}}}{m!} \exp\Big(\frac{m\Delta\Theta}{\lambda'}\Big) \sum_{k=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{k} \exp\left(\frac{-4k\Delta\Theta}{\lambda'} \right) \\
&\overset{(b)}{\leq} \frac{\sqrt{2}}{2^{m/2}} \left(1 + \exp\left(-\frac{4\Delta\Theta}{\lambda'} \right) \right)^{\frac{m}{2}} \exp\left(\frac{2\Delta\Theta}{\lambda'} \right)^{\frac{m}{2}} \\
&= \sqrt{2} \left(\frac{\exp\left(-\frac{2\Delta\Theta}{\lambda_0} \right) + \exp\left(\frac{2\Delta\Theta}{\lambda'} \right)}{2} \right)^{\frac{m}{2}} \\
&\overset{(c)}{\leq} \sqrt{2} \exp\left[m\Big(\frac{\Delta\Theta}{\lambda_0}\Big)^2 \right],
\end{aligned}$$

where (a) follows since $\binom{n}{k}$ is maximized at $k = n/2$, (b) uses Stirling's approximation, and (c) uses the fact that for any x,

$$\frac{e^x + e^{-x}}{2} \leq e^{x^2/2},$$

which can be proved by Taylor series expansion of $\exp(x)$. The theorem follows since the other terms present are small compared to n^{94} by the definition of profiles.

□

Remark

1. It is possible to prove a stronger form of the theorem that does not have the $\sqrt{2}$ term. It makes the proof slightly longer and does not yield any improvements in the leading terms of the worst-case redundancy and therefore not presented.

2. For any fixed profile, as we let $\Delta \to 0$ all distributions look alike and hence the ratio on the left hand side of the theorem should converge to 1, as can be shown . We bound it by $\sqrt{2}$.

We now prove a result equivalent to Lemma 50 for average redudancy. Recall that we are interested in $\mathcal{I}^{\text{poi}(n)}_{\Phi_{\text{med}}}$, where all λ's are between $n^{1/3}$ and n/H, and hence there are at most $n^{2/3}$ elements in any such distribution. Consider any interval $(\lambda_0, \lambda_0 + \Delta]$. Let Φ^{near} be the set of all profiles with all multiplicities in $I_{\text{near}} \overset{\text{def}}{=} [\lambda_0 - 2\sqrt{\lambda_0 \log n}, (\lambda_0 + \Delta) + 2\sqrt{(\lambda_0 + \Delta) \log n}]$. Applying Poisson tail bounds from Lemma 4 a poi(λ) random variable lies in $\lambda \pm 2\sqrt{\lambda \log n}$ with probability at least $1 - 1/n^4$, and the probability falls exponentially beyond the range. By the union bound, a profile generated by $\mathcal{I}(m, s, (\lambda_0, \lambda_0 + \Delta])$ for $m \leq n$ is in Φ^{near} with probability $> 1 - 1/n^3$. We use this to first show that it suffices to consider Φ^{near}.

Lemma 55. $\sum_{\bar{\varphi} \in \Phi} \sup_{\overline{\Lambda} \in \mathcal{I}(m,s,(\lambda_0,\lambda_0+\Delta])} \overline{\Lambda}(\bar{\varphi}) < 2 \sum_{\bar{\varphi} \in \Phi^{\text{near}}} \sup_{\overline{\Lambda} \in \mathcal{I}(m,s,(\lambda_0,\lambda_0+\Delta])} \overline{\Lambda}(\bar{\varphi}).$

Proof. Recall from Equation (6.5) that the probability of a profile $\bar{\varphi} = \{\mu_1, \ldots, \mu_k\}$ is

$$\begin{aligned}\overline{\Lambda}(\bar{\varphi}) =& \frac{1}{\prod_{i=0}^{\infty} \varphi_i!} \sum_{\sigma \in S_m} \prod_{j=1}^{k} \text{poi}(\lambda_{\sigma(j)}, \mu_j) \\ =& \frac{1}{\prod_{i=0}^{\infty} \varphi_i!} \sum_{\sigma \in S_m} \Big[\prod_{\mu_j \in I_{\text{near}}} \text{poi}(\lambda_{\sigma(j)}, \mu_j) \prod_{\mu_j \notin I_{\text{near}}} \text{poi}(\lambda_{\sigma(j)}, \mu_j) \Big].\end{aligned}$$

A profile $\bar{\varphi}$ can be written as the union

$$\bar{\varphi} = \bar{\varphi}_{\text{near}} \cup \bar{\varphi}_{\text{far}},$$

where $\bar{\varphi}_{\text{near}}$ is the collection of all multiplicities in I_{near}, and $\bar{\varphi}_{\text{far}}$ has all elements outside this interval. Now when we are interested in distributions in $\overline{\Lambda} \in \mathcal{I}(m, s, (\lambda_0, \lambda_0 + \Delta])$ a multiplicity outside I_{near} is maximized by either λ or $\lambda + \Delta$. Therefore, for any distribution in $\overline{\Lambda} \in \mathcal{I}(m, s, (\lambda_0, \lambda_0 + \Delta])$

$$\overline{\Lambda}(\bar{\varphi}) \leq m^{i_1+i_2} \cdot \prod_{\mu_i \notin I_{\text{near}}} \max\{\text{poi}(\lambda_0 - \Delta, \mu_i), \text{poi}(\lambda_0 + \Delta, \mu_i)\} \Big(\max \overline{\Lambda}(\bar{\varphi}_{\text{near}}) \Big)$$

Now consider all profiles with the same $\bar{\varphi}_{\text{near}}$,

$$\begin{aligned}
&\sum_{\bar{\varphi}|\bar{\varphi}\sim\bar{\varphi}_{\text{near}}} \max \overline{\Lambda}(\bar{\varphi}) \\
&\le \max \overline{\Lambda}(\bar{\varphi}_{\text{near}}) \cdot \sum_j m^j \prod_{\mu_i \notin I_{\text{near}}} \max\{\text{poi}(\lambda_0 - \Delta, \mu_i), \text{poi}(\lambda_0 + \Delta, \mu_i)\} \\
&\overset{(a)}{\le} \max \overline{\Lambda}(\bar{\varphi}_{\text{near}}) \cdot \sum_{i_1, i_2} \left(\frac{m}{n^2}\right)^j,
\end{aligned}$$

where (a) uses Lemma 4. Summing over all j and using the fact that $m \le n$ proves the result. □

Using these two lemmas, we bound $\hat{R}(\mathcal{I}_\Phi(m, s, (\lambda_0, \lambda_0 + \Delta]))$.

Lemma 56.

$$\hat{R}(\mathcal{I}_\Phi(m, s, (\lambda_0, \lambda_0 + \Delta])) \le \frac{3}{2} + m\frac{\Delta^2\left(\Delta + 4\sqrt{(\Delta + \lambda_0)\log n}\right)^2}{\lambda_0^2}.$$

We divide $\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}$ into T_n classes of distributions, and bound the redundancy of each class. Invoking Lemma 17 gives the result.

Partition of $\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}$

We form partition of the set of distributions similar to the construction for average redundancy. Partition $(n^{1/3}, H]$ into b consecutive intervals $I_1, I_2, \ldots, I_b$ of lengths $\Delta_1, \Delta_2, \ldots, \Delta_b$. We will optimize for b later. A distribution in $\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}$ is a collection of positive reals in $(n^{1/3}, H]$ that sum to at most n.

Let b be a positive integer (specified later). Consider any partition of $(0, n]$ into $b + 1$ consecutive intervals $I_0, I_1, I_2, \ldots, I_b$ of lengths $\Delta_0 = 1, \Delta_1, \Delta_2, \ldots, \Delta_b$. In other words, the first interval is $(0, 1]$. We will be mostly interested in the b intervals $I_1, \ldots, I_b$.

A distribution in $\mathcal{I}_{\Phi}^{\text{poi}(n)}$ is a collection of positive reals in $(n^{1/3}, H]$. For any $\overline{\Lambda} \in \mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}$,

- For $j = 1, 2, \ldots, b$, $\overline{\Lambda}_j \overset{\text{def}}{=} \overline{\Lambda} \cap I_j$ be the multiset of elements of $\overline{\Lambda}$ in I_j,

- For $j = 1, 2, \ldots, b$, let $m_j \overset{\text{def}}{=} m_j(\overline{\Lambda}) \overset{\text{def}}{=} |\overline{\Lambda}_j|$ be the number of elements of $\overline{\Lambda}$ in I_j, and $\overline{m} \overset{\text{def}}{=} \overline{m}(\overline{\Lambda})$ is the b-tuple of m_j's.

- $s_j \overset{\text{def}}{=} s_j(\overline{\Lambda}_{\text{med}}) \overset{\text{def}}{=} \lfloor n^{94} \cdot \sum_{\lambda \in \overline{\Lambda}_j} \lambda \rfloor$, and $\overline{s}(\overline{\Lambda}_{\text{med}}) \overset{\text{def}}{=} (s_1, \ldots, s_b)$.

A partition of $(n^{1/3}, H]$ induces a partition of $\mathcal{I}_{\Phi}^{\text{poi}(n)}$ via $\overline{m}$ and $\overline{s}$, *i.e.*, distributions having same $\overline{m}$ and $\overline{s}$ are in the same class.

Let $\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}(i)$, $i = 1, \ldots, T_n$ be these classes of distributions. Note that each m_j is at most $n^{2/3}$ and any s_j an integer that is at most n^{95}. The number of possible tuples $\overline{m}$ and $\overline{s}$ is therefore at most $n^b \cdot (n^{95})^b \le n^{96b}$. Therefore, $T_n \le n^{96b}$.

Using Lemma 17 with this yields the following result.

Lemma 57.

$$\hat{R}(\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}) \le \max_{1 \le i \le T_n} \hat{R}(\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}(i)) + 100 b \log n.$$

Bounding redundancy of each class

Let $\mathcal{I}$ be one of these T_n classes. Let the class $\mathcal{I}$ correspond to $\overline{m} = (m_1, \ldots, m_b)$ and $\overline{s} = (s_1, \ldots, s_b)$ and λ_j^* be the start of interval I_j.

We bound $\hat{R}(\mathcal{I})$ by proving a result analogous to Theorem 51.

Theorem 58.

$$\hat{R}(\mathcal{I}_{\Phi_{\text{med}}}^{\text{poi}(n)}) \le \frac{3b}{2} + \sum_{j=1}^{b} m_j \frac{\Delta_j^2 \left(\Delta_j + 4\sqrt{\lambda_{j+1}^* \log n} \right)^2}{(\lambda_j^*)^2}.$$

Proof. Analogous to the average case, let $\bar{\varphi}_j$ be the profile generated by $\overline{\Lambda}_j$. Then $\bar{\varphi}_{\text{med}} = \bar{\varphi}_1 \cup \ldots \cup \bar{\varphi}_b = f(\bar{\varphi}_1, \ldots, \bar{\varphi}_b)$, where $\bar{\varphi}_j$'s are independent due to Poisson sampling. For $j = 1, \ldots, b$,

$$\mathcal{I}_j \overset{\text{def}}{=} \{\overline{\Lambda}_j : \overline{\Lambda} \in \mathcal{I}\}$$

be the class of marginal distributions in interval j.

By definition, $\mathcal{I}_\Phi(m_j, s_j, (\lambda_j^*, \lambda_{j+1}^*])$ consists of all possible distributions with m_j elements in $(\lambda_j^*, \lambda_{j+1}^*]$. Therefore, $\mathcal{I}_j \subseteq \mathcal{I}_\Phi(m_j, , s_j, (\lambda_j^*, \lambda_{j+1}^*])$.

Since sampling is Poisson, $\bar{\varphi}_j$'s are independent. By

$$\hat{R}(\mathcal{I}) \overset{(a)}{\leq} \sum_{j=1}^{b} \hat{R}(\mathcal{I}_j) \overset{(b)}{\leq} \sum_{j=1}^{b} \Big[\frac{3}{2} + m_j \frac{\Delta_j^2\Big(\Delta_j + 4\sqrt{\lambda_{j+1}^* \log n}\Big)^2}{(\lambda_j^*)^2}\Big] \tag{6.16}$$

where (a) uses Lemmas 11 and 14 and (b) follows from Lemma 56. □

Interval lengths and redundancy

We will bound worst case redundancy by choosing intervals in a geometric progression similar to the average case. Recall that $\lambda_1^* = n^{1/3}$. If we take the interval ends in a geometric progression with ratio c, then $n^{1/3}(1+c)^b = H$. Now $\lambda_{j+1}^* = \lambda_j^*(1+c)$, and $\Delta_j = c\lambda_j^*$. As before $c \leq \frac{\log n}{3b}$. Also, $\Theta_j \leq \Delta_j + 2\sqrt{\lambda_j^* \log n}$ yields, $\Theta_j^2 \leq 2(\Delta_j^2 + 4\lambda_j^* \log n)$. Plugging these reduces the bound in Equation (6.16), and using $\lambda_j^* \leq H$,

$$\begin{aligned}\frac{3b}{2} + \sum_{j=1}^{b} 2m_j c^2(\Delta_j^2 + 4\lambda_j^* \log n) \leq & \frac{3b}{2} + 8nc^2 \log n + \sum_{j=1}^{b} 2m_j c^4 (\lambda_j^*)^2 \\ \leq & \frac{3b}{2} + 8nc^2 \log n + \sum_{j=1}^{b} 2m_j c^4 \lambda_j^* H \\ \leq & \frac{3b}{2} + 8nc^2 \log n + 2n \cdot Hc^4.\end{aligned}$$

Combining the results of Lemmas 52, 53, and 57,

$$\begin{aligned}\hat{R}(\mathcal{I}_\Phi^{\text{poi}(n)}) \leq & 4n^{1/3} \log n + \frac{n}{H} \log n + \frac{3b}{2} + 8nc^2 \log n + 2n \cdot Hc^4 + 100b \log n \\ \leq & 4n^{1/3} \log n + 1.5b + \frac{n}{H} \log n + 8n\frac{\log^3 n}{b^2} + 2n \cdot H\Big(\frac{\log n}{b}\Big)^4 + 100b \log n\end{aligned}$$

To optimize the value of this expression, let $H = n^{2/3}/\log^{1/6} n$, and $b = n^{1/3} \log^{2/3} n$, then for large n,

$$\begin{aligned}\hat{R}(\mathcal{I}_\Phi^{\text{poi}(n)}) \leq & 4n^{1/3} \log n + 1.5n^{1/3} \log^{2/3} n + 6n^{1/3} \log^{7/6} n + 108n^{1/3} \log^{5/3} n \\ < & 110n^{1/3} \log^{5/3} n.\end{aligned}$$

Acknowledgement

Chapter 6 is partially adapted from Jayadev Acharya, Hirakendu Das, Alon Orlitsky, "Tight bounds on profile redundancy and distinguishability", *Neural Information Processing Systems (NIPS)*, 2012, and Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Ananda Theertha Suresh, "Tight Bounds for Universal Compression of Large Alphabets", *IEEE International Symposium on Information Theory (ISIT)*, 2013. The dissertation author was a primary researcher and author of this paper.

Bibliography

[1] C. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] L. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 783–795, Nov. 1973.

[3] B. M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Probl. Inform. Transm.*, vol. 2, no. 2, pp. 1–7, 1966.

[4] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Transactions on Information Theory*, vol. 24, no. 6, pp. 674–682, Nov. 1978.

[5] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information Theory*, vol. 30, no. 4, pp. 629–636, July 1984.

[6] R. Krichevsky and V. Trofimov, "The preformance of universal coding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.

[7] T. Cover and J. Thomas, *Elements of Information Theory, 2nd Ed.* Wiley Interscience, 2006.

[8] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games.* New York, NY, USA: Cambridge University Press, 2006.

[9] P. D. Grünwald, *The Minimum Description Length Principle.* The MIT Press, 2007.

[10] D. Foster, R. Stine, and A. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1713–1720, June 2002.

[11] G. I. Shamir, "Universal source coding for monotonic and fast decaying monotonic distributions," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7194–7211, 2013.

[12] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 358–373, 2009.

[13] D. Bontemps, "Universal coding on infinite alphabets: Exponentially decreasing envelopes," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1466–1478, 2011.

[14] J. Åberg, Y. M. Shtarkov, and B. J. M. Smeets, "Multialphabet coding with separate alphabet description," in *Proceedings of Compression and Complexity of Sequences*, 1997.

[15] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469– 1481, July 2004.

[16] P. Valiant, "Testing symmetric properties of distributions," Ph.D. dissertation, Cambridge, MA, USA, 2008, aAI0821026.

[17] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 2200–2203, 2004.

[18] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.

[19] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Annual Symposium on Foundations of Computer Science*, 2000, p. 259.

[20] ——, "Testing closeness of discrete distributions," *J. ACM*, vol. 60, no. 1, p. 4, 2013.

[21] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," in *COLT*, vol. 19, 2011.

[22] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. T. Suresh, "Competitive classification and closeness testing," in *COLT*, 2012, pp. 22.1–22.18.

[23] A. Orlitsky, N. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, October 17 2003.

[24] G. Shamir, "A new upper bound on the redundancy of unknown alphabets," in *CISS, Princeton*, 2004.

[25] Y. M. Shtarkov, “Universal sequential coding of single messages,” *Problems of Information Transmission*, vol. 23, no. 3, pp. 3–17, 1987.

[26] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, “Efficient universal noiseless source codes,” *IEEE Transactions on Information Theory*, vol. 27, no. 3, pp. 269–279, 1981.

[27] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, “The context-tree weighting method: basic properties.” *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[28] T. Cover, “Universal portfolios,” *Mathematical Finance*, vol. 1, no. 1, pp. 1–29, January 1991.

[29] J. Rissanen, “Fisher information and stochastic complexity,” *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[30] W. Szpankowski, “On asymptotics of certain recurrences arising in universal coding,” *Problems of Information Transmission*, vol. 34, no. 2, pp. 142–146, 1998.

[31] Q. 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.

[32] W. Szpankowski and M. J. Weinberger, “Minimax redundancy for large alphabets,” in *ISIT*, 2010, pp. 1488–1492.

[33] A. Orlitsky and N. Santhanam, “Speaking of infinity,” *IEEE Transactions on Information Theory*, vol. To appear, 2004.

[34] G. Valiant and P. Valiant, “Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts,” in *STOC*, 2011, pp. 685–694.

[35] M. Mitzenmacher and E. Upfal, *Probability and computing - randomized algorithms and probabilistic analysis.* Cambridge Univ. Press, 2005.

[36] R. M. Roth, *Introduction to coding theory.* Cambridge University Press, 2006.

[37] K. Pearson, “Contributions to the Mathematical Theory of Evolution,” *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[38] T. Batu, R. Kumar, and R. Rubinfeld, “Sublinear algorithms for testing monotone and unimodal distributions,” in *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, ser. STOC ’04. New York, NY, USA: ACM, 2004, pp. 381–390.

[39] G. K. Zipf, *The Psychobiology of Language.* New York, NY, USA: Houghton-Mifflin, 1935.

[40] N. Merhav, G. Seroussi, and M. J. Weinberger, "Optimal prefix codes for sources with two-sided geometric distributions," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 121–135, 2000.

[41] P. Elias, "Universal codeword sets and representations of integers," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 194–203, Mar. 1975.

[42] J. Rissanen, "Minimax codes for finite alphabets," *Information Theory, IEEE Transactions on*, vol. 24, no. 3, pp. 389–392, 1978.

[43] B. Y. Ryabko, "Coding of a source with unknown but ordered probabilities," *Problemy Peredachi Informatsii*, vol. 15, no. 2, pp. 71–77, 1979.

[44] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes." *IEEE Transactions on Information Theory*, vol. 26, no. 2, pp. 166–174, 1980.

[45] M. Khosravifard, H. Saidi, M. Esmaeili, and T. A. Gulliver, "The minimum average code for finite memoryless monotone sources." *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 955–975, 2007.

[46] J. Acharya, H. Das, and A. Orlitsky, "Tight bounds on profile redundancy and distinguishability," in *NIPS*, 2012.

[47] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. Suresh, "Tight bounds for universal compression of large alphabets," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2013, pp. 2875–2879.

[48] L. Birgé, "On the risk of histograms for estimating decreasing densities," *Annals of Statistics*, vol. 15, no. 3, pp. 1013–1022, 1987.

[49] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning k-modal distributions via testing," in *SODA*, 2012, pp. 1371–1385.

[50] M. Woodroofe and J. Sun, "Testing uniformity versus a monotone density," *The Annals of Statistics*, vol. 27, no. 1, pp. pp. 338–360, 1999.

[51] O. Goldreich, S. Goldwasser, E. Lehman, and D. Ron, "Testing monotonicity," *Foundations of Computer Science, IEEE Annual Symposium on*, vol. 0, p. 426, 1998.

[52] R. Rubinfeld and R. A. Servedio, "Testing monotone high-dimensional distributions," *Random Struct. Algorithms*, vol. 34, no. 1, pp. 24–44, 2009.

[53] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, “A competitive test for uniformity of monotone distributions,” in *AISTATS*, 2013, pp. 57–65.

[54] N. Merhav, G. Seroussi, and M. J. Weinberger, “Coding of sources with two-sided geometric distributions and unknown parameters,” *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 229–236, 2000.

[55] Z. Burda, D. Johnston, J. Jurkiewicz, M. Kaminski, M. A. Nowak, G. Papp, and I. Zahed, “Wealth condensation in pareto macro-economies,” 2001.

[56] N. Jevtić, A. Orlitsky, and N. Santhanam, “A lower bound on compression of unknown alphabets,” 2005.

[57] G. Shamir, “Universal lossless compression with unknown alphabets—the average case,” *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 4915–4944, Nov. 2006.

[58] A. Garivier, “A lower-bound for the maximin redundancy in pattern coding,” *Entropy*, vol. 11, no. 4, pp. 634–642, 2009.

[59] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, “Testing closeness of discrete distributions,” *CoRR*, vol. abs/1009.5397, 2010.

[60] H. Das, “Competitive tests and estimators for properties of distributions,” Ph.D. dissertation, UCSD, 2012.

[61] P. Elias, “Universal codeword sets and representations of the integers,” *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 194–203, Mar 1975.