UNIVERSITY OF CALIFORNIA, SAN DIEGO

Statistical Inference over Large Domains

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

in

Electrical Engineering (Communication Theory and Systems)

by

Ananda Theertha Suresh

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Sanjoy Dasgupta
Professor Young-Han Kim
Professor Ramamohan Paturi
Professor Alexander Vardy

2016

The dissertation of Ananda Theertha Suresh is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2016

DEDICATION

In memory of my dear father S. Suresh.

EPIGRAPH

What is unseen, is not necessarily unknown

— Wendelessen

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to my advisor, colleagues, friends and family for their support and contribution in the making of this dissertation. My sincerest thanks go to my advisor Alon Orlitsky for the deep and fruitful interaction we had over the years. I am inspired by his style of thinking, dedication, approachability, and pursuit for elegant solutions and I hope to incorporate them in my future research. His support, encouragement, and guidance were instrumental for this dissertation.

I am extremely thankful to my committee members Sanjoy Dasgupta, Young-Han Kim, Ramamohan Paturi, and Alexander Vardy for helpful comments on my research and teaching amazing courses. I benefited a lot from the research-oriented graduate courses that I took with them.

Research at UCSD was fun and encouraging, thanks to my current and past lab mates Jayadev Acharya, Hirakendu Das, Moein Falahatgar, Ashkan Jafarpour, Shengjun Pan, and Venkatadheeraj Pichapathi. Their collaborations on my various problems broadened my horizon in various directions. I am especially thankful to Jayadev for giving me the often unnecessary wake-up calls early in the morning, Hirakendu for patiently explaining all of his research during my first year of PhD, and Shengjun for teaching me everything I know about Latex hacks.

I am also thankful to Alon's former students, in particular Narayana Prasad Santhanam and Krishnamurthy Vishwanathan for comments and discussions. I was also fortunate to have interacted and worked with various postdocs at UCSD, including Sudeep Kamath, Yonathan Kaspi, Mesrob Ohannessian, and Himanshu Tyagi.

I want to thank Mehryar Mohri and Michael Riley for mentoring me during a fruitful summer at Google Research, New York. During that summer and the following spring, I had the chance to work with Aditya Bhaskara, Morteza Zadimoghaddam, and Yihong Wu, and I am thankful to them for many technical and non-technical discussions.

There is a long list of friends at UCSD and outside, which cannot fit into this page. I am happy and thankful to have come across them and to have enjoyed their company during the time of my PhD.

I am indebted to my parents and to my sister Rohini, for their tireless love and support. They have done everything in their capacity, and beyond, to help me achieve whatever I have. Simply put, I am nothing without them.

Chapter 2 is adapted from Alon Orlitsky and Ananda Theertha Suresh, "Competitive distribution estimation: Why is Good-Turing good", *Neural Information Processing Systems (NIPS),* 2015.

Chapters 3 and 4 are adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Optimal probability estimation with applications to prediction and classification", *Conference on Learning Theory (COLT)*, 2013; and Alon Orlitsky and Ananda Theertha Suresh, "Competitive distribution estimation: Why is Good-Turing good", *Neural Information Processing Systems (NIPS),* 2015.

Chapters 5 and 6 are adapted from Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu, "Estimating the number of unseen species: A bird in the hand is worth $\log n$ in the bush", Manuscript, 2015.

Chapter 7 is adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Near-optimal-sample estimators for spherical Gaussian mixtures", *Neural Information Processing Systems (NIPS),* 2014.

VITA

2010	B. Tech. in Engineering Physics, Indian Institute of Technology Madras
2012	M. S. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego
2016	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

Estimating the number of unseen species: A bird in the hand is worth $\log n$ in the bush, A. Orlitsky, A. T. Suresh, Yihong Wu, Manuscript, 2015.

Competitive distribution estimation: Why is Good-Turing good, A. Orlitsky, A. T. Suresh, *Neural Information Processing Systems*, 2015.

Faster algorithms for testing under conditional sampling, M. Falahatgar, A. Jafarpour, A. Orlitsky, A. T. Suresh, V. Pichapati, *Conference on Learning Theory*, 2015.

On learning distributions from their samples, S. Kamath, A. Orlitsky, A. T. Suresh, V. Pichapati, *Conference on Learning Theory*, 2015.

Universal compression of power-law distributions, M. Falahatgar, A. Jafarpour, A. Orlitsky, A. T. Suresh, V. Pichapati, *IEEE International Symposium on Information Theory*, 2015.

Automata and graph compression, M. Mohri, M. Riley, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2015.

Sparse solutions to nonnegative linear systems and applications, A. Bhaskara, A. T. Suresh, M. Zaghimoghaddam, *International Conference on Artificial Intelligence and Statistics*, 2015.

The complexity of estimating Rényi entropy, J. Acharya, A. Orlitsky, A. T. Suresh, H. Tyagi, *Symposium on Discrete Algorithms*, 2015.

Near-optimal-sample estimators for spherical Gaussian mixtures, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *Neural Information Processing Systems*, 2014.

Sorting with adversarial comparators and application to density estimation, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2014.

Efficient compression of monotone and m-modal distributions, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2014.

Poissonization and universal compression of envelope classes, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2014.

Sublinear algorithms for outlier detection and generalized closeness testing, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2014.

Optimal probability estimation with applications to prediction and classification, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *Conference on Learning Theory*, 2013.

A competitive test for uniformity of monotone distributions, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, *International Conference on Artificial Intelligence and Statistics*, 2013.

Tight bounds on worst-case pattern redundancy, J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2013.

Competitive classification and closeness testing, J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, A. T. Suresh, *Conference on Learning Theory*, 2012.

On the query computation and verification of functions, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, A. T. Suresh, *IEEE International Symposium on Information Theory*, 2012.

Strong and weak secrecy in wiretap channels, A. Subramanian, A. T. Suresh, S. Raj, A. Thangaraj, M. Bloch and S. W. McLaughlin , *International Symposium on Turbo Codes and Iterative Information Processing*, 2010.

Strong secrecy for erasure wiretap channels, A. T. Suresh, A. Subramanian, A. Thangaraj, M. Bloch and S. W. McLaughlin, *IEEE Information Theory Workshop*, 2010.

Interplay between optimal selection scheme design, selection criterion, and discrete rate adaptation in opportunistic wireless systems, N. B. Mehta, R. Talak, A. T. Suresh, *IEEE Transactions on Communications,* 2013.

On optimal timer-based distributed selection for rate-adaptive multi-user diversity systems, A. T. Suresh, N. B. Mehta, V. Shah, *National Conference on Communications*, India, 2010.

Universal compression of envelope classes: Tight characterization via Poisson sampling, J. Acharya, A. Jafarpour, A. Orlitsky, A. T. Suresh, Submitted to *IEEE Transactions on Infornmation Theory.*

Estimating Rényi entropy of discrete distributions, J. Acharya, A. Orlitsky, A. T. Suresh, H. Tyagi, Submitted to *IEEE Transactions on Infornmation Theory.*

On the computation and verification query complexity of symmetric functions, J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, A. T. Suresh, Submitted to *IEEE Transactions on Infornmation Theory.*

ABSTRACT OF THE DISSERTATION

Statistical Inference over Large Domains

by

Ananda Theertha Suresh

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

University of California, San Diego, 2016

Professor Alon Orlitsky, Chair

Motivated by diverse applications in ecology, genetics, and language modeling, researchers in learning, computer science, and information theory have recently studied several fundamental statistical questions in the large domain regime, where the domain size is large relative to the number of samples.

We study three such basic problems with rich history and wide applications. In the course of analyzing these problems, we also provide provable guarantees for several existing practical estimators and propose estimators with better guarantees.

Competitive distribution estimation and classification: Existing theory does not explain why absolute-discounting, Good-Turing, and related estimators outperform the asymptotically min-max optimal estimators in practice.

We explain their performance by showing that a variant of Good-Turing estimators performs near optimally for all distributions. Specifically, for distributions over k symbols and n samples, we show that a simple variant of Good-Turing estimator is always within KL divergence of $(3+o_n(1))/n^{1/3}$ from a genie-aided estimator that knows the underlying distribution up to a permutation, and that a more involved estimator is within $\tilde{\mathcal{O}}_n(\min(k/n, 1/\sqrt{n}))$. We extend these results to classification, where the goal is to classify a test sample based on two training samples of length n each.

Estimating the number of unseen species: We study species estimation, where given n independent samples from an unknown species distribution, we would like to estimate the number of new species that will be observed among the next $t \cdot n$ samples. Existing algorithms provide guarantees only for the prediction range $t \leq 1$. We significantly extend the range of predictability and prove that a class of estimators including Efron-Thisted accurately predicts the number of unseen species for $t \propto \log n$. Conversely, we show that no estimator is accurate for $t = \Omega(\log n)$.

Learning Gaussian mixtures: We derive the first sample-efficient polynomial-time estimator for high-dimensional spherical Gaussian mixtures. It estimates mixtures of k spherical Gaussians in d-dimensions to within ℓ_1 distance ϵ using $\mathcal{O}(dk^9(\log^2 d)/\epsilon^4)$ samples and $\mathcal{O}_{k,\epsilon}(d^3 \log^5 d)$ computation time. Conversely, we show that any estimator requires $\Omega\big(dk/\epsilon^2\big)$ samples, hence the algorithm's sample complexity is nearly optimal in the number of dimensions. We also construct a simple estimator for one-dimensional Gaussian mixtures that uses $\tilde{\mathcal{O}}(k/\epsilon^2)$ samples and $\tilde{\mathcal{O}}((k/\epsilon)^{3k+1})$ computation time.

Chapter 1

Introduction

Most processes in modern day engineering and science are inferred via probabilistic models. For example, in natural language processing the words are often modeled according to a Markov distribution, in speech processing Gaussian mixtures are employed, and in genetics Bernoulli-product models are used. Hence, inferring properties of these probabilistic processes, often refered to as *statistical inference*, is one of the most fundamental problems of modern day science and engineering.

Given the applicability of statistical inference, it has been studied extensively over the last century. Traditionally, most of the statistical inference has been studied in the asymptotic regime where the the number of samples far exceeds the domain size. For example, in natural language processing, this translates to having observed a lot of words compared to the vocabulary size.

While such assumptions are true for many classical examples, modern applications often require us to draw inference in the opposite regime, where the underlying domain size is comparable to or far exceeds the number of samples. Even if the number of samples is comparable to the domain size, we do not even observe all the underlying symbols even once! For example, in genetics every sample is a human DNA, hence the number of all possible samples is far less than the number of all possible human DNAs.

Even if data is plentiful, the number of possible models can be very large, thus rendering traditional methods computationally inefficient. For example, in

speech processing Gaussian mixtures are often used, and the number of possible Gaussian mixtures scales exponentially with the number of dimensions, thus one cannot use traditional approaches to provably infer the underlying mixture.

Thus there is an inherent need for inference algorithms that are both *data-efficient* and *computation-efficient*. Focusing on these goals, in this dissertation, we study three fundamental problems in the large domain regime. Our objective is to (i) formulate relevant and basic questions for each problem, (ii) determine the fundamental limit: minimum amount of data required, (iii) provide theoretical justification for existing efficient algorithms and if possible improve upon them, (iv) derive such algorithms when they do not exist.

We start with the most basic question in the large domain regime: how well can one estimate a discrete distribution? Surprisingly, even after decades of research, it is not well understood why practical estimators such as Good-Turing work better than min-max optimal estimators.

1.1 How well can one estimate?

Estimating distributions over large alphabets is a fundamental machine-learning tenet. In its simplest form, given n independent samples from an unknown discrete distribution over k symbols, it asks for an estimate of the underlying distribution. An obvious and intuitive estimator is the empirical frequency estimator that assigns to each symbol a probability proportional to the number of times it appears. For example, if we toss a coin 5 times and observe 3 heads and 2 tails, the empirical estimate assigns probability $3/5 = 0.6$ to heads and $2/5 = 0.4$ to tails.

While this empirical estimator is intuitive, it performs poorly in practice for a variety of reasons. For example, the empirical estimate always assigns probability 0 to unseen symbols. This might be accurate for large number of samples, but it is often inaccurate for small number of samples. .

To overcome these shortcomings with empirical estimators, distribution estimation has been studied extensively starting with Laplace. Yet no method is

known to estimate all distributions well. For example, add-constant estimators are nearly min-max optimal but often perform poorly in practice, and practical estimators such as absolute discounting, Jelinek-Mercer, and Good-Turing are not known to be near optimal for essentially any distribution.

Instead of the well studied min-max approach, we propose to study distribution estimation in a *competitive setting*. Specifically, for every discrete distribution, we construct estimators that are provably nearly the best in the following two competitive ways. First they estimate every distribution nearly as well as the best estimator designed with prior knowledge of the distribution up to a permutation. Second, they estimate every distribution nearly as well as the best estimator designed with prior knowledge of the exact distribution, but as all natural estimators, restricted to assign the same probability to all symbols appearing the same number of times.

Specifically, for distributions over k symbols and n samples, we show that for both comparisons, a simple variant of Good-Turing estimator is always within KL divergence of

$$\frac{3+o_n(1)}{n^{1/3}}$$

from the best estimator, and that a more involved estimator is within

$$\tilde{\mathcal{O}}_n\left(\min\left(\frac{k}{n},\frac{1}{\sqrt{n}}\right)\right).$$

Notice that the above results are independent of the domain size k and hence is particularly useful for large domain settings. Conversely, we show that any estimator must have a KL divergence at least $\tilde{\Omega}_n(\min(k/n,1/n^{2/3}))$ over the best estimator for the first comparison, and at least $\tilde{\Omega}_n(\min(k/n,1/\sqrt{n}))$ for the second.

We modify the estimator to derive a linear-complexity classifier that takes two length-n training sequences, one distributed *i.i.d.* according to a distribution p and one according to q, and classifies a single test sample generated by p or q, with error at most $\widetilde{\mathcal{O}}_n(n^{-1/5})$ higher than that achievable by the best classifier that knows p and q up to a permutation. We also show an $\widetilde{\Omega}_n(n^{-1/3})$ lower bound on this additional error for any classifier.

Our main proof technique is to relate the problem of competitive distribution estimation to that of combined-probability mass estimation and then provide

uniform bounds on the estimation of combined probability masses.

1.2 How far can one predict?

Population estimation is an important problem in many scientific endeavors. Its most popular formulation, introduced by Fisher, uses n samples to predict U, the number of hitherto unseen elements that will be observed among $t \cdot n$ new samples. In 1956, Good and Toulmin [1] approximated U by a fascinating estimator that has since intrigued statisticians and mathematicians [2]. For example, in Stanford University's Statistics Department's brochure published in the early 90's [3], and slightly abbreviated in Figure 1.1, Bradley Efron credited the problem and its elegant solution with kindling his interest in statistics.

The Good-Toulmin estimator fails to predict further due to its high variance. Later Efron and Thisted showed empirically that a variation of this estimator approximates U even for some $t > 1$, but no theoretical guarantees are known.

We derive a class of estimators that *provably* predict U not just for constant $t > 1$, but all the way up to

$$t \propto \log n,$$

with a *normalized mean squared error* of

$$\mathcal{O}\left(\frac{1}{n^{1/t}}\right).$$

This shows that the number of species can be estimated for a population $\log n$ times larger than that observed, a factor that grows arbitrarily large as n increases. We also show that this range is the best possible and that the estimators' mean-square error is optimal up to constants for any t. Our approach yields the first provable guarantee for the Efron-Thisted estimator and, in addition, a variant which achieves stronger theoretical and experimental performance than existing methodologies on a variety of synthetic and real datasets.

From the time I was a little boy until my senior year in college I wanted to be a mathematician. Then I learned that I really wanted to be a 19th century mathematician, the kind who does a little theory, a lot of computation, and some consulting with real scientists. The field of statistics has allowed me to do all three things, in whatever proportions I desired. Here is an example of the three faces of statistics, done in the early 1940's.

In the early 1940's, naturalist Corbet had spent two years trapping butterflies in Malaya. At the end of that time he constructed a table to show how many times he had trapped the various butterfly species. For example, 118 species were so rare that Corbet had trapped only one specimen of each, 74 species had been trapped twice each, etc.

Frequency	1	2	3	4	5	6	7	8	9	10	11	...
Species	118	74	44	24	29	22	20	19	20	15	12	...

Corbet returned to England with his table, and asked R.A. Fisher, the greatest of all statisticians, how many new species he would see if he returned to Malaya for another two years of trapping. This question seems impossible to answer, since it refers to a column of Corbet's table that doesn't exist, the "0" column. Fisher provided an interesting answer to the question, which was later improved on, the number of new species you can expect to see in two years of additional trapping is

$$118 - 74 + 44 - 24 + \ldots - 12 + 6 = 75.$$

Figure 1.1: Excerpt from Efron's comments in the Stanford statistics brochure from early 2000's.

The estimators we derive are simple linear estimators that are computable in time proportional to n. The performance guarantees hold uniformly for all distributions, and apply to all four standard sampling models commonly used across various scientific disciplines: multinomial, Poisson, hypergeometric, and Bernoulli product.

1.3 How efficiently can one learn in high dimensions?

Meaningful information often resides in high-dimensional spaces: voice signals are expressed in many frequency bands, credit ratings are influenced by multiple parameters, and document topics are manifested in the prevalence of numerous words. Some applications, such as topic modeling and genomic analysis consider data in over 1000 dimensions, [4, 5].

Typically, information can be generated by different types of sources: voice is spoken by men or women, credit parameters correspond to wealthy or poor individuals, and documents address topics such as sports or politics. In such cases the overall data follow a mixture distribution [6, 7, 8].

Mixtures of high-dimensional distributions are therefore central to the understanding and processing of many natural phenomena. Methods for recovering the mixture components from the data have consequently been extensively studied by statisticians, engineers, and computer scientists.

We learn Gaussian mixtures in the *PAC learning* framework, where the goal is to output a mixture that is at a ℓ_1 distance at most ϵ to the underlying one. We provide the first sample-efficient polynomial-time estimator for high-dimensional spherical Gaussian mixtures.

For mixtures of any k d-dimensional spherical Gaussians, we derive an intuitive spectral-estimator that uses

$$\mathcal{O}_k\left(\frac{d\log^2 d}{\epsilon^4}\right)$$

samples and runs in time

$$\mathcal{O}_{k,\epsilon}(d^3\log^5 d),$$

to output a mixture that is ϵ close to the underlying mixture in ℓ_1 distance. Our sample and time complexities are significantly lower than previously known. The constant factor $\mathcal{O}_k$ is polynomial for sample complexity and is exponential for the time complexity, again much smaller than what was previously known. Furthermore, the results are independent of any parameters of the Gaussian mixture.

We also show that $\Omega_k\left(\frac{d}{\epsilon^2}\right)$ samples are needed for any algorithm. Hence the sample complexity is near-optimal in the number of dimensions. We also derive a simple estimator for k-component one-dimensional mixtures that uses

$$\mathcal{O}\left(\frac{k\log\frac{k}{\epsilon}}{\epsilon^2}\right)$$

samples and runs in time

$$\widetilde{\mathcal{O}}\left(\left(\frac{k}{\epsilon}\right)^{3k+1}\right).$$

1.4 Thesis organization

The rest of the thesis is organized as follows.

Part I: How well can one estimate?

Chapter 2: We describe competitive distribution estimation and relate it to min-max combined-probability estimation.

Chapter 3: We study the problem of combined-probability estimation and provide guarantees for Good-Turing type estimators.

Chapter 4: We extend the estimation results to competitive classification and provide classifiers that are uniformly close to the genie-aided classifier that knows the distributions up to a permutation.

Part II: How far can one predict?

Chapter 5: We study the unseen species estimation problem and provide linear estimators that are extend the predictability range to $\mathcal{O}(n\log n)$ under the Poisson sampling model.

Chapter 6: We extend the results to other three popular models: multinomial, hypergeometric, and Bernoulli-product. We then prove that the performance

of proposed learning estimators are near-optimal for multinomial and Poisson sampling models.

Part III: How efficiently can one learn in high dimensions?

Chapter 7: We propose an algorithm to learn spherical Gaussian mixtures whose sample complexity is near-optimal in the number of dimensions.

Part I

How well can one estimate?

Chapter 2

Competitive distribution estimation

2.1 Introduction

2.1.1 Background

Many learning applications, ranging from language-processing staples such as speech recognition and machine translation to biological studies in virology and bioinformatics, call for estimating large discrete distributions from their samples. Probability estimation over large alphabets has therefore long been the subject of extensive research, both by practitioners deriving practical estimators [9, 10], and by theorists searching for optimal estimators [11].

Yet even after all this work, provably-optimal estimators remain elusive. The add-constant estimators frequently analyzed by theoreticians are nearly min-max optimal, yet perform poorly for many practical distributions, while common practical estimators, such as absolute discounting [12], Jelinek-Mercer [13], and Good-Turing [14], are not well understood and lack provable performance guarantees.

To understand the terminology and approach a solution we need a few definitions. A *probability distribution* over a discrete set $\mathcal{X}$ is a mapping $p : \mathcal{X} \to [0, 1]$ such that $\sum_{x \in \mathcal{X}} p_x = 1$. A distribution *estimator* over a support set $\mathcal{X}$

associates with any observed sample sequence $x^* \in \mathcal{X}^*$ a distribution $q(x^*)$ over $\mathcal{X}$. The performance of an estimator q for an underlying distribution p is typically evaluated in terms of the *Kullback-Leibler (KL) divergence* [15],

$$D(p||q) \stackrel{\text{def}}{=} \sum_x p_x \log \frac{p_x}{q_x},$$

reflecting the expected increase in the ambiguity about the outcome of p when it is approximated by q. KL divergence is also the increase in the number of bits over the entropy that q uses to compress the output of p, and is also the *log-loss* of estimating p by q. It is therefore of interest to construct estimators that approximate a large class of distributions to within small KL divergence. We now describe one of the problem's simplest formulations.

2.1.2 Min-max loss

Given n samples $X^n \stackrel{\text{def}}{=} X_1, X_2, \ldots, X_n$, generated independently according to a distribution p over $\mathcal{X}$, the expected KL loss of the estimator q is

$$r_n(q, p) = \mathop{\mathbb{E}}_{X^n \sim p^n} [D(p||q(X^n))].$$

Let $\mathcal{P}$ be a known collection of distributions over a discrete set $\mathcal{X}$. The worst-case loss of an estimator q over all distributions in $\mathcal{P}$ is

$$r_n(q, \mathcal{P}) \stackrel{\text{def}}{=} \max_{p \in \mathcal{P}} r_n(q, p), \tag{2.1}$$

and the lowest worst-case loss for $\mathcal{P}$, achieved by the best estimator, is the min-max loss

$$r_n(\mathcal{P}) \stackrel{\text{def}}{=} \min_q r_n(q, \mathcal{P}) = \min_q \max_{p \in \mathcal{P}} r_n(q, p). \tag{2.2}$$

Min-max performance can be viewed as regret relative to an oracle that knows the underlying distribution. Hence from here on we refer to it as *regret*.

The most natural and important collection of distributions, and the one we study here, is the set of all discrete distributions over an alphabet of some size k,

which without loss of generality we assume to be $[k] = \{1, 2, \ldots k\}$. Hence the set of all distributions is the *simplex* in k dimensions,

$$\Delta_k \stackrel{\text{def}}{=} \Big\{(p_1, \ldots, p_k) : p_i \geq 0 \text{ and } \sum p_i = 1\Big\}.$$

Following [16], researchers have studied $r_n(\Delta_k)$ and related quantities [17]. We outline some of the results derived.

2.1.3 Add-constant estimators

The *add-β* estimator assigns to a symbol that appeared t times a probability proportional to $t + \beta$. For example, if three coin tosses yield one heads and two tails, the add-1/2 estimator assigns probability $1.5/(1.5 + 2.5) = 3/8$ to heads, and $2.5/(1.5 + 2.5) = 5/8$ to tails. [18] showed that as for every k, as $n \to \infty$, an estimator related to add-3/4 is near optimal and achieves

$$r_n(\Delta_k) = \frac{k-1}{2n} \cdot (1 + o(1)). \tag{2.3}$$

The more challenging, and practical, regime is where the sample size n is not overwhelmingly larger than the alphabet size k. For example in English text processing, we need to estimate the distribution of words following a context. But the number of times a context appears in a corpus may not be much larger than the vocabulary size. Several results are known for other regimes as well. When the sample size n is linear in the alphabet size k, $r_n(\Delta_k)$ can be shown to be a constant, and [11] showed that as $k/n \to \infty$, add-constant estimators achieve the optimal

$$r_n(\Delta_k) = \log \frac{k}{n} \cdot (1 + o(1)), \tag{2.4}$$

While add-constant estimators are nearly min-max optimal, the distributions attaining the min-max regret are near uniform. In practice, large-alphabet distributions are rarely uniform, and instead, tend to follow a power-law. For these distributions, add-constant estimators under-perform the estimators described in the next subsection.

2.1.4 Practical estimators

For real applications, practitioners tend to use more sophisticated estimators, with better empirical performance. These include the Jelinek-Mercer estimator that cross-validates the sample to find the best fit for the observed data. Or the absolute-discounting estimators that rather than add a positive constant to each count, do the *opposite*, and subtract a positive constant.

Perhaps the most popular and enduring have been the *Good-Turing* estimator [14] and some of its variations. Let $n_x \stackrel{\text{def}}{=} n_x(x^n)$ be the number of times a symbol x appears in x^n and let $\varphi_t \stackrel{\text{def}}{=} \varphi_t(x^n)$ be the number of symbols appearing t times in x^n. The basic Good-Turing estimator posits that if $n_x = t$,

$$q_x(x^n) = \frac{\varphi_{t+1}}{\varphi_t} \cdot \frac{t+1}{n},$$

surprisingly relating the probability of an element not just to the number of times it was observed, but also to the number other elements appearing as many, and one more, times. It is easy to see that this basic version of the estimator may not work well, as for example it assigns any element appearing $\geq n/2$ times 0 probability. Hence in practice the estimator is modified, for example, using empirical frequency to elements appearing many times.

The Good-Turing Estimator was published in 1953, and quickly adapted for language-modeling use, but for half a century no proofs of its performance were known. Following [19], several papers, *e.g.,* [20, 21], showed that Good-Turing variants estimate the combined probability of symbols appearing any given number of times with accuracy that does not depend on the alphabet size, and [22] showed that a different variation of Good-Turing similarly estimates the probabilities of each previously-observed symbol, and all unseen symbols combined.

However, these results do not explain why Good-Turing estimators work well for the actual probability estimation problem, that of estimating the probability of each element, not of the combination of elements appearing a certain number of times. To define and derive uniformly-optimal estimators, we take a different, competitive, approach.

2.2 Competitive optimality

2.2.1 Overview

To evaluate an estimator, we compare its performance to the best possible performance of two estimators designed with some prior knowledge of the underlying distribution. The first estimator is designed with knowledge of the underlying distribution up to a permutation of the probabilities, namely knowledge of the probability multiset, *e.g.,* $\{.5, .3, .2\}$, but not of the association between probabilities and symbols. The second estimator is designed with exact knowledge of the distribution, but like all *natural estimators*, forced to assign the same probabilities to symbols appearing the same number of times. For example, upon observing the sample a, b, c, a, b, d, e, the estimator must assign the same probability to a and b, and the same probability to c, d, and e.

These estimators cannot be implemented in practice as in reality we do not have prior knowledge of the estimated distribution. But the prior information is chosen to allow us to determine the best performance of any estimator designed with that information, which in turn is better than the performance of any *data-driven* estimator designed without prior information. We then show that certain variations of the Good-Turing estimators, designed without any prior knowledge, approach the performance of both prior-knowledge estimators for every underlying distribution.

2.2.2 Competing with near full information

We first define the performance of an *oracle-aided* estimator, designed with some knowledge of the underlying distribution. Suppose that the estimator is designed with the aid of an oracle that knows the value of $f(p)$ for some given function f over the class Δ_k of distributions.

The function f partitions Δ_k into subsets, each corresponding to one possible value of f. We denote the subsets by P, and the partition by $\mathbb{P}$, and as before, denote the individual distributions by p. Then the oracle knows the unique partition part P such that $p \in P \in \mathbb{P}$. For example, if $f(p)$ is the multiset of p,

then each subset P corresponds to set of distributions with the same probability multiset, and the oracle knows the multiset of probabilities.

For every partition part $P \in \mathbb{P}$, an estimator q incurs the worst-case regret in (2.1),

$$r_n(q, P) = \max_{p \in P} r_n(q, p).$$

The oracle, knowing the unique partition part P, incurs the least worst-case regret (2.2),

$$r_n(P) = \min_q r_n(q, P).$$

The *competitive regret* of q over the oracle, for all distributions in P is

$$r_n(q, P) - r_n(P),$$

the competitive regret over all partition parts and all distributions in each is

$$r_n^{\mathbb{P}}(q, \Delta_k) \stackrel{\text{def}}{=} \max_{P \in \mathbb{P}} \left(r_n(q, P) - r_n(P)\right),$$

and the best possible competitive regret is

$$r_n^{\mathbb{P}}(\Delta_k) \stackrel{\text{def}}{=} \min_q r_n^{\mathbb{P}}(q, \Delta_k).$$

Consolidating the intermediate definitions,

$$r_n^{\mathbb{P}}(\Delta_k) = \min_q \max_{P \in \mathbb{P}} \left(\max_{p \in P} r_n(q, p) - r_n(P) \right).$$

Namely, an oracle-aided estimator who knows the partition part incurs a worst-case regret $r_n(P)$ over each part P, and the competitive regret $r_n^{\mathbb{P}}(\Delta_k)$ of data-driven estimators is the least overall increase in the part-wise regret due to not knowing P. The following examples evaluate $r_n^{\mathbb{P}}(\Delta_k)$ for the two simplest partitions.

Example 2.1. The *singleton partition* consists of $|\Delta_k|$ parts, each a single distribution in Δ_k,

$$\mathbb{P}_{|\Delta_k|} \stackrel{\text{def}}{=} \{\{p\} : p \in \Delta_k\}.$$

An oracle-aided estimator that knows the part containing p knows p. The competitive regret of data-driven estimators is therefore the min-max regret,

$$\begin{aligned} r_n^{\mathbb{P}_{|\Delta_k|}}(\Delta_k) &= \min_q \max_{p\in\Delta_k} \left(r_n(q,\{p\}) - r_n(\{p\})\right) \\ &= \min_q \max_{p\in\Delta_k} r_n(q,p) \\ &= r_n(\Delta_k), \end{aligned}$$

where the middle equality follows as $r_n(q,\{p\}) = r_n(q,p)$, and $r_n(\{p\}) = 0$.

Example 2.2. The *whole-collection* partition has only one part, the whole collection Δ_k,

$$\mathbb{P}_1 \stackrel{\text{def}}{=} \{\Delta_k\}.$$

An estimator aided by an oracle that knows the part containing p has no additional information, hence no advantage over a data-driven estimator, and the competitive regret is 0,

$$\begin{aligned} r_n^{\mathbb{P}_1}(\Delta_k) &= \min_q \max_{P\in\{\Delta_k\}} \left(\max_{p\in P} r_n(q,p) - r_n(P)\right) \\ &= \min_q \left(\max_{p\in\Delta_k} r_n(q,p) - r_n(\Delta_k)\right) \\ &= \min_q \max_{p\in\Delta_k} \left(r_n(q,p)\right) - r_n(\Delta_k) \\ &= r_n(\Delta_k) - r_n(\Delta_k) \\ &= 0. \end{aligned}$$

The examples show that for the coarsest partition of Δ_k, into a single part, the competitive regret is the lowest possible, 0, while for the finest partition, into singletons, the competitive regret is the highest possible, $r_n(\Delta_k)$.

A partition $\mathbb{P}'$ *refines* a partition $\mathbb{P}$ if every part in $\mathbb{P}$ is partitioned by some parts in $\mathbb{P}'$. For example $\{\{a,b\},\{c\},\{d,e\}\}$ refines $\{\{a,b,c\},\{d,e\}\}$. We show that if $\mathbb{P}'$ refines $\mathbb{P}$ then for every q, $r_n^{\mathbb{P}'}(q,\Delta_k) \geq r_n^{\mathbb{P}}(q,\Delta_k)$.

Lemma 2.3. If $\mathbb{P}'$ refines $\mathbb{P}$ then for any q,

$$r_n^{\mathbb{P}'}(q,\Delta_k) \geq r_n^{\mathbb{P}}(q,\Delta_k). \tag{2.5}$$

Proof. The definition implies that if $P' \subseteq P$ then $r_n(P') \le r_n(P)$, for every distribution class P and P'. Hence for every q,

$$\begin{aligned}
r_n^{\mathbb{P}'}(q, \Delta_k) &= \max_{P' \in \mathbb{P}'} (r_n(q, P') - r_n(P')) \\
&= \max_{P \in \mathbb{P}} \max_{P \supseteq P' \in \mathbb{P}'} (r_n(q, P') - r_n(P')) \\
&\ge \max_{P \in \mathbb{P}} \max_{P \supseteq P' \in \mathbb{P}'} (r_n(q, P') - r_n(P)) \\
&= \max_{P \in \mathbb{P}} \big(\max_{P \supseteq P' \in \mathbb{P}'} r_n(q, P') - r_n(P) \big) \\
&= \max_{P \in \mathbb{P}} (r_n(q, P) - r_n(P)) \\
&= r_n^{\mathbb{P}}(q, \Delta_k).
\end{aligned}$$

□

Considering the collection Δ_k of all distributions over $[k]$, it follows that as we start with single-part partition $\{\Delta_k\}$ and keep refining it till the oracle knows p, the competitive regret of estimators will increase from 0 to $r_n(q, \Delta_k)$. A natural question is therefore how much information can the oracle have and still keep the competitive regret low? We show that the oracle can know the distribution exactly up to permutation, and still the regret will be very small.

Two distributions p and p' *permutation equivalent* if for some permutation σ of $[k]$,

$$p'_{\sigma(i)} = p_i,$$

for all $1 \le i \le k$. For example, $(0.5, 0.3, 0.2)$ and $(0.3, 0.5, 0.2)$ are permutation equivalent. Permutation equivalence is clearly an equivalence relation, and hence partitions the collection of distributions over $[k]$ into equivalence classes. Let $\mathbb{P}_\sigma$ be the corresponding partition. We construct estimators q that uniformly bound $r_n^{\mathbb{P}_\sigma}(q, \Delta_k)$, thus the same estimator uniformly bounds $r_n^{\mathbb{P}}(q, \Delta_k)$ for any coarser partition of Δ_k, such as partitions into classes of distributions with the same support size, or entropy. Note that the partition $\mathbb{P}_\sigma$ corresponds to knowing the underlying distribution up to permutation, hence $r_n^{\mathbb{P}_\sigma}(\Delta_k)$ is the additional KL loss compared to an estimator designed with knowledge of the underlying distribution up to permutation.

This notion of competitiveness has appeared in several contexts. In data compression it is called *twice-redundancy* [23, 24, 25, 26], while in statistics it is often called *adaptive* or *local min-max* [27, 28, 29, 30, 31], and recently in property testing it is referred as competitive [32, 33, 34] or *instance-by-instance* [35]. Subsequent to this work, [36] studied competitive estimation in ℓ_1 distance, however their regret is $\mathrm{poly}(1/\log n)$, compared to our $\widetilde{\mathcal{O}}(1/\sqrt{n})$.

2.2.3 Competing with natural estimators

Our second comparison is with an estimator designed with exact knowledge of p, but forced to be *natural*, namely, to assign the same probability to all symbols appearing the same number of times in the sample. For example, for the observed sample a, b, c, a, b, d, e, the same probability must be assigned to a and b, and the same probability to c, d, and e. Since data-driven estimators derive all their knowledge of the distribution from the data, we expect them to be natural.

We compare the regret of data-driven estimators to that of *natural oracle-aided* estimators. Let $\mathcal{Q}^{\mathrm{nat}}$ be the set of all natural estimators. For a distribution p, the lowest regret of a natural estimator, designed with prior knowledge of p is

$$r_n^{\mathrm{nat}}(p) \stackrel{\mathrm{def}}{=} \min_{q \in \mathcal{Q}^{\mathrm{nat}}} r_n(q, p).$$

The regret of an estimator q relative to the least-regret natural-estimator is

$$r_n^{\mathrm{nat}}(q, p) = r_n(q, p) - r_n^{\mathrm{nat}}(p).$$

The regret of data-driven estimators relative to natural estimators over Δ_k is therefore,

$$r_n^{\mathrm{nat}}(\Delta_k) = \min_q \max_{p \in \Delta_k} r_n^{\mathrm{nat}}(q, p).$$

In the next section we state the results, showing in particular that $r_n^{\mathrm{nat}}(\Delta_k)$ is uniformly bounded. In Section 2.4 we describe experiments comparing the performance of competitive estimators to that of min-max motivated estimators and in Section 2.5 we provide the proofs. Finally in Section 2.6 we prove the lower bound.

2.3 Results

Recall that φ_t denotes the number of symbols appearing t times. For a sequence x^n, let the *combined prbability mass* $S_t \stackrel{\text{def}}{=} S_t(x^n)$ denote the total probability of symbols appearing t times. For notational convenience, we use S_t to denote both $S_t(x^n)$ and $S_t(X^n)$ and the usage becomes clear in the context. Similar to KL divergence between distributions, we define KL divergence between S and their estimates $\hat{S}$ as

$$D(S||\hat{S}) = \sum_{t=0}^{n} S_t \log \frac{S_t}{\hat{S}_t},$$

and the ℓ_1 distance between S and $\hat{S}$ as

$$\left|\left|S - \hat{S}\right|\right|_1 = \sum_{t=0}^{n} |S_t - \hat{S}_t|.$$

Our main result relates the two competitive formulations and further relate them to the min-max estimation of the combined probability mass S_t.

Theorem 2.4. For a natural estimator q, let $\hat{S}_t = \sum_{x:N_x=t} q_x$, then

$$r_n^{\text{nat}}(q, \Delta_k) = \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})].$$

Furthermore,

$$r_n^{\mathbb{P}_\sigma}(\Delta_k) \le r_n^{\text{nat}}(\Delta_k) = \min_{\hat{S}} \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})].$$

The above result relates the competitive regret of distribution estimation to the min-max regret for the combined-probability mass estimation. Good-Turing estimators are often used in conjunction with empirical frequency, where Good-Turing estimates low probabilities and empirical frequency estimates large probabilities. If $n_x = t$,

$$q'_x = \frac{C_t}{\varphi_t},$$

where C_t is a variation of the God-Turing estimate

$$C_t = \begin{cases} \frac{t \cdot \varphi_t}{N} & \text{if } t \ge t_0, \\ (\varphi_{t+1} + 1) \cdot \frac{t+1}{N} & \text{else,} \end{cases}$$

In the next chapter, we first show that even this simple Good-Turing version C estimates combined-probability mass well and hence is uniformly optimal for all distributions. For simplicity we prove the result when the number of samples is $n' \sim \text{poi}(n)$, a Poisson random variable with mean n. A similar result holds with exactly n samples, but the proof is more involved as the multiplicities are dependent. Specifically in Theorem 3.1 we show that for any k and n, upon observing $n' \sim \text{poi}(n)$ samples,

$$\max_{p \in \Delta_k} \mathbb{E}[D(S||C)] \leq \frac{3 + o_n(1)}{n^{1/3}},^*$$

Furthermore, we show that this bound is tight for any simple combination of Good-Turing and empirical estimators is optimal up to logarithmic factors in Lemma 3.2. We then show that a more complex a more complex variant of the Good-Turing estimator for the combined probability mass, denoted F', achieves a faster convergence rate in Theorem 3.3. Namely, for every distribution p and every n, F' satisfies, with probability at least $1 - 1/n$,

$$D(S||F') = \widetilde{\mathcal{O}}_n\left(\min\left(\frac{1}{n^{1/2}}, \frac{k}{n}\right)\right),$$

and hence by Pinsker's inequality, with probability at least $1 - 1/n$,

$$||F' - S||_1 = \widetilde{\mathcal{O}}_n\left(\min\left(\frac{1}{n^{1/4}}, \frac{\sqrt{k}}{\sqrt{n}}\right)\right).$$

Where $\widetilde{\mathcal{O}}_n$, and below also $\tilde{\Omega}_n$ and $\tilde{\Theta}_n$ hide multiplicative logarithmic factors in n. The above result holds with probability at least $1 - 1/n$. Using the fact that $F'_t \geq 1/n^2$, one can easily convert it to a result on expectation and show that

$$\max_{p \in \Delta_k} \mathbb{E}[D(S||F')] = \widetilde{\mathcal{O}}_n\left(\min\left(\frac{1}{n^{1/2}}, \frac{k}{n}\right)\right),$$

and

$$\max_{p \in \Delta_k} \mathbb{E}[||F' - S||_1] = \widetilde{\mathcal{O}}_n\left(\min\left(\frac{1}{n^{1/4}}, \frac{\sqrt{k}}{\sqrt{n}}\right)\right).$$

Furthermore, we show that matching information theoretic lower bound for estimating the combined probability mass S in Theorem 3.4.

$^*a_n = o_n(1)$, means $\limsup_{n\to\infty} a_n = 0$.

Fano's inequality usually yields lower bounds on KL loss, not regret. By carefully constructing distribution classes, we lower bound the regret.

Theorem 2.5. For any k and n,

$$r_n^{\mathbb{P}_\sigma}(\Delta_k) \geq \tilde{\Omega}_n\left(\min\left(\frac{1}{n^{2/3}}, \frac{k}{n}\right)\right).$$

To summarize, we have proved that

$$\tilde{\Omega}_n\left(\min\left(\frac{1}{n^{2/3}}, \frac{k}{n}\right)\right) \leq r_n^{\mathbb{P}_\sigma}(\Delta_k) \leq r_n^{\text{nat}}(\Delta_k) = \tilde{\Theta}_n\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right).$$

2.3.1 Illustration and implications

Figure 2.1 demonstrates some of the results. The horizontal axis reflects the set Δ_k of distributions illustrated on one dimension. The vertical axis indicates the KL loss, or absolute regret, for clarity, shown for $k \gg n$. The blue line is the previously-known min-max upper bound on the regret, which by (2.4) is very high for this regime, $\log(k/n)$. The red line is the regret of the estimator designed with prior knowledge of the probability multiset. Observe that while for some probability multisets the regret approaches the $\log(k/n)$ min-max upper bound, for other probability multisets it is much lower, and for some, such as uniform over 1 or over k symbols, where the probability multiset determines the distribution it is even 0. For many practically relevant distributions, such as power-law distributions and sparse distributions, the regret is small compared to $\log(k/n)$. The green line is an upper bound on the absolute regret of the data-driven proposed estimator q''. By Theorem 3.3, it is always at most $1/\sqrt{n}$ larger than the red line. It follows that for many distributions, possibly for distributions with more structure, such as those occurring in nature, the regret of q'' is significantly smaller than the pessimistic min-max bound implies.

We observe a few consequences of these results.

- Theorems 3.1 and 3.3 establish two uniformly-optimal estimators q' and q''. Their relative regrets diminish to zero at least as fast as $1/n^{1/3}$, and $1/\sqrt{n}$ respectively, independent of how large the alphabet size k is.

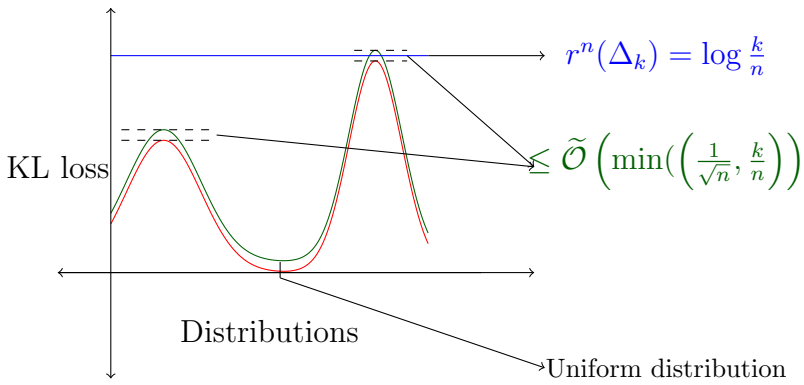


Figure 2.1: Qualitative behavior of the KL loss as a function of distributions in different formulations.

- Although the results are for relative regret, as shown in Figure 2.1, they lead to estimator with smaller absolute regret, namely, the expected KL divergence.
- The same regret upper bounds hold for all coarser partitions of Δ_k *i.e.,* where instead of knowing the multiset, the oracle knows some property of multiset such as entropy.

2.4 Experiments

Recall that for a sequence x^n, n_x denotes the number of times a symbol x appears and φ_t denotes the number of symbols appearing t times. For small values of n and k, the near-optimal estimator q'' proposed in the next chapter simplifies to a combination of Good-Turing and empirical estimators. By Lemmas 3.17 and 3.18 (stated in the next chapter), for symbols appearing t times, if $\varphi_{t+1} \geq \tilde{\Omega}(t)$, then the Good-Turing estimate is close to the underlying combined-probability mass, otherwise the empirical estimate is closer. Hence, for a symbol appearing t times, if $\varphi_{t+1} \geq t$ we use the Good-Turing estimator, otherwise we use the empirical

estimator. If $n_x = t$,

$$q_x = \begin{cases} \frac{t}{N} & \text{if } t > \varphi_{t+1}, \\ \frac{\varphi_{t+1}+1}{\varphi_t} \cdot \frac{t+1}{N} & \text{else,} \end{cases}$$

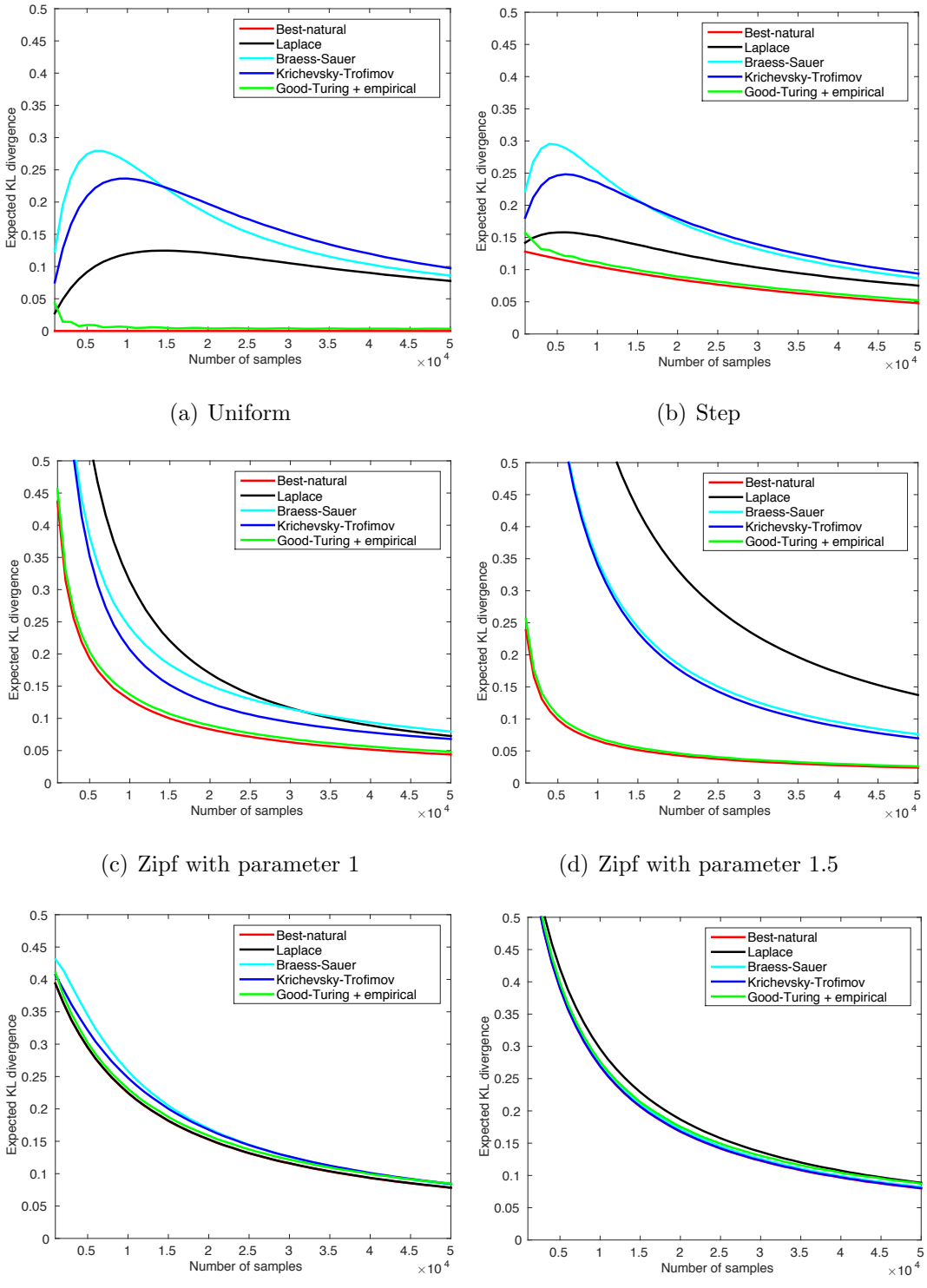
where N is a normalization factor. Note that we have replaced φ_{t+1} in the Good-Turing estimator by $\varphi_{t+1} + 1$ to ensure that every symbol is assigned a non-zero probability.

We compare the performance of this estimator to four estimators: three popular add-β estimators and the optimal natural estimator. An add-beta estimator $\hat{S}$ has the form

$$q_x^{\hat{S}} = \frac{n_x + \beta_{n_x}^{\hat{S}}}{N(\hat{S})},$$

where $N(\hat{S})$ is a normalization factor to ensure that the probabilities add up to 1. The Laplace estimator, $\beta_t^L = 1 \,\forall\, t$, minimizes the expected loss when the underlying distribution is generated by a uniform prior over Δ_k. The Krichevsky-Trofimov estimator, $\beta_t^{KT} = 1/2 \,\forall\, t$, is asymptotically min-max optimal for the cumulative regret, and minimizes the expected loss when the underlying distribution is generated according to a Dirichlet-1/2 prior. The Braess-Sauer estimator, $\beta_0^{BS} = 1/2, \beta_1^{BS} = 1, \beta_t^{BS} = 3/4 \;\forall\, t > 1$, is asymptotically min-max optimal for $r_n(\Delta_k)$. Finally, as shown in Lemma 2.9, the optimal estimator $q_x = \frac{S_{n_x}}{\varphi_{n_x}}$ achieves the lowest loss of any natural estimator designed with knowledge of the underlying distribution.

We compare the performance of the proposed estimator to that of the four estimators above. We consider six distributions: uniform distribution, step distribution with half the symbols having probability $1/2k$ and the other half have probability $3/2k$, Zipf distribution with parameter 1 ($p_i \propto i^{-1}$), Zipf distribution with parameter 1.5 ($p_i \propto i^{-1.5}$), a distribution generated by the uniform prior on Δ_k, and a distribution generated from Dirichlet-1/2 prior. All distributions have support size $k = 10000$ and $n \leq 50000$ samples. Results are averaged over 200 trials. Figure 2.2 shows the results. Observe that the proposed estimator performs similarly to the best natural estimator for all six distributions. It also significantly outperforms the other estimators for Zipf, uniform, and step distributions.



(a) Uniform

(b) Step

(c) Zipf with parameter 1

(d) Zipf with parameter 1.5

(e) Uniform prior (Dirichlet 1)

(f) Dirichlet 1/2 prior

Figure 2.2: Simulation results for support 10000, number of samples ranging from 1000 to 50000, averaged over 200 trials.

The performance of other estimators depends on the underlying distribution. For example, since Laplace is the optimal estimator when the underlying distribution is generated from the uniform prior, it performs well in Figure 2.2(e), however performs poorly on other distributions.

Furthermore, even though for distributions generated by Dirichlet priors, all the estimators have similar looking regrets (Figures 2.2(e), 2.2(f)), the proposed estimator performs better than estimators which are not designed specifically for that prior.

2.5 Relating the two competitive formulations

Our goal is to show that every estimator q, $r_n^{\mathbb{P}_\sigma}(q, \Delta_k) \leq r_n^{\text{nat}}(q, \Delta_k)$ and then upper bound $r_n^{\text{nat}}(q, \Delta_k)$ by the min-max regrets in estimating the combined probability mass. To that end, we first prove the following result.

Lemma 2.6. For every class $P \in \mathbb{P}_\sigma$, $r_n(P) \geq \max_{p \in P} r_n^{\text{nat}}(p)$.

Proof. We first show that there is an optimal estimator q that is natural. In particular, let

$$q''_y(x^n) = \frac{\sum_{p \in P} p(x^n y)}{\sum_{p' \in P} p'(x^n)}.$$

We show that $q''_y(x^n)$ is an optimal estimator for P. Since $q''_y(x^n) = q''_{\sigma(y)}(\sigma(x^n))$ for any permutation σ, the estimator achieves the same loss for every $p \in P$,

$$\max_{p \in P} r_n(q'', p) = \frac{1}{k!} \sum_{p \in P} r_n(q'', p'). \tag{2.6}$$

For any estimator q,

$$\begin{aligned}
\max_{p\in P}\mathbb{E}[D(p||q)] &\overset{(a)}{\geq} \frac{1}{k!}\sum_{p\in P}\mathbb{E}_p[D(p||q)] \\
&\overset{(b)}{=} \frac{1}{k!}\sum_{p\in P}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p) \\
&= \frac{1}{k!}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}}\sum_{p\in P} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p) \\
&\overset{(c)}{\geq} \frac{1}{k!}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}}\sum_{p\in P} p(x^n y)\log\frac{\sum_{p'\in P}p'(x^n)}{\sum_{p''\in P}p''(x^n y)} - H(p) \\
&= \frac{1}{k!}\sum_{p\in P}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}} p(x^n y)\log\frac{1}{{q''}_y(x^n)} - H(p) \\
&\overset{(d)}{=} \frac{1}{k!}\sum_{p\in P} r_n(q'',p).
\end{aligned}$$

(a) follows from the fact that maximum is larger than the average. (b) follows from the fact that every distribution in P has the same entropy. Non-negativity of KL divergence implies (c). All distributions in P has the same entropy and hence (d). Hence together with Equation (2.6)

$$\begin{aligned}
r_n(P) &= \min_q \max_{p\in P}\mathbb{E}[D(p||q)] \\
&\geq \frac{1}{k!}\sum_{p\in P} r_n(q'',p') \\
&= \max_{p\in P} r_n(q'',p).
\end{aligned}$$

Hence q'' is an optimal estimator. Recall that n_y denote the number of times symbol y appears in the sequence. q'' is natural as if $n_y = n_{y'}$, then ${q''}_y(x^n) = {q''}_{y'}(x^n)$. Since there is a natural estimator that achieves minimum in $r_n(P)$,

$$\begin{aligned}
r_n(P) &= \min_q \max_{p\in P}\mathbb{E}[D(p||q)] \\
&= \min_{q\in\mathcal{Q}^{\text{nat}}} \max_{p\in P}\mathbb{E}[D(p||q)] \\
&\geq \max_{p\in P} \min_{q\in\mathcal{Q}^{\text{nat}}}\mathbb{E}[D(p||q)] \\
&= \max_{p\in P} r_n^{\text{nat}}(p),
\end{aligned}$$

where the last inequality follows from the fact that min-max is bigger than max-min. □

Lemma 2.7. For every estimator q,

$$r_n^{\mathbb{P}_\sigma}(q, \Delta_k) \leq r_n^{\text{nat}}(q, \Delta_k).$$

Proof.

$$\begin{aligned} r_n^{\mathbb{P}_\sigma}(q, \Delta_k) &= \max_{P \in \mathbb{P}_\sigma} \left(\max_{p \in P} \mathbb{E}[D(p||q)] - r_n(P) \right) \\ &\overset{(a)}{\leq} \max_{P \in \mathbb{P}_\sigma} \left(\max_{p \in P} \mathbb{E}[D(p||q)] - \max_{p \in P} r_n^{\text{nat}}(p) \right) \\ &\overset{(b)}{\leq} \max_{P \in \mathbb{P}_\sigma} \max_{p \in P} \left(\mathbb{E}[D(p||q)] - r_n^{\text{nat}}(p) \right) \\ &= \max_{p \in \Delta_k} \left(\mathbb{E}[D(p||q)] - r_n^{\text{nat}}(p) \right) \\ &= r_n^{\text{nat}}(q, \Delta_k). \end{aligned}$$

Lemma 2.6 implies (a). Difference of maximums is smaller than maximum of differences, hence (b). □

2.5.1 Relation between $r_n^{\text{nat}}(q, \Delta_k)$ and combined-probability estimation

We now relate the regret in estimating distribution to that of estimating the combined probability mass. Since the natural estimator assigns same probability to symbols that appear the same number of times, estimating probabilities is same as estimating the total probability of symbols appearing a given number of times. We formalize it in the next lemma.

Lemma 2.8. For a natural estimator q let $\hat{S}_t = \sum_{x:N_x=t} q_x$, then

$$r_n^{\text{nat}}(q, p) = \mathbb{E}[D(S||\hat{S})].$$

The proof uses the following lemma which computes the best natural estimator. For a random sequence X^n, let $\Phi_t \overset{\text{def}}{=} \varphi_t(X^n)$.

Lemma 2.9. Let $q_x^*(x^n)=\frac{S_{n_x}}{\varphi_{n_x}}$, then

$$q^* = \arg\min_{q\in\mathcal{Q}^{\text{nat}}} r_n(q,p)$$

and

$$r_n^{\text{nat}}(p) = \mathbb{E}\left[\sum_{t=0}^{n} S_t \log\frac{\Phi_t}{S_t}\right] - H(p). \tag{2.7}$$

Proof. For a natural estimator q, if $n_y = n_{y'}$, then $q_y(x^n) = q_{y'}(x^n)$. Hence, with a slight abuse of notation let $q_{n_y}(x^n) = q_y(x^n)$. For a sequence x^n and estimator q,

$$\begin{aligned}
\sum_{y\in\mathcal{X}} p_y \log\frac{1}{q_y(x^n)} - \sum_{t=0}^{n} S_t \log\frac{\varphi_t}{S_t} &= \sum_{t=0}^{n}\sum_{y:n_y=t} p_y \log\frac{1}{q_y(x^n)} - \sum_{t=0}^{n} S_t \log\frac{\varphi_t}{S_t} \\
&= \sum_{t=0}^{n} S_t \log\frac{1}{q_t(x^n)} - \sum_{t=0}^{n} S_t \log\frac{\varphi_t}{S_t} \\
&= \sum_{t=0}^{n} S_t \log\frac{S_t}{\varphi_t q_t(x^n)} \\
&\geq 0,
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{t=0}^{n} S_t = \sum_{t=0}^{n} \varphi_t q_t(x^n) = 1$ and KL divergence is non-negative. Furthermore, equality is achieved only by the estimator that assigns $q_x^* = \frac{S_{n_x}}{\varphi_{n_x}}$. Hence,

$$r_n^{\text{nat}}(p) = \min_{q\in\mathcal{Q}^{\text{nat}}} \mathbb{E}\left[\sum_{y\in\mathcal{X}} p_y \log\frac{p_y}{q_y(X^n)}\right] = -H(p) + \mathbb{E}\left[\sum_{t=0}^{n} S_t \log\frac{\Phi_t}{S_t}\right].$$

□

Proof of Lemma 2.8. As before, with a slight abuse of notation let $q_{n_y}(x^n) = q_y(x^n)$ for natural estimators q. For any natural estimator q and sequence x^n,

$$\begin{aligned}
\sum_{y\in\mathcal{X}} p_y \log\frac{1}{q_y(x^n)} &= \sum_{t=0}^{n}\sum_{y:n_y=t} p_y \log\frac{1}{q_y(x^n)} \\
&= \sum_{t=0}^{n} S_t \log\frac{S_t}{\varphi_t q_t(x^n)} + \sum_{t=0}^{n} S_t \log\frac{\varphi_t}{S_t} \\
&= \sum_{t=0}^{n} S_t \log\frac{S_t}{\hat{S}_t} + \sum_{t=0}^{n} S_t \log\frac{\varphi_t}{S_t}.
\end{aligned}$$

Thus by Lemma 2.9,

$$\begin{aligned} r_n^{\text{nat}}(q,p) &= -H(p) + \mathbb{E}\left[\sum_{t=0}^{n} S_t \log \frac{S_t}{\hat{S}_t} + \sum_{t=0}^{n} S_t \log \frac{\Phi_t}{S_t}\right] + H(p) - \mathbb{E}\left[\sum_{t=0}^{n} S_t \log \frac{\Phi_t}{S_t}\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{n} S_t \log \frac{S_t}{\hat{S}_t}\right] \\ &= \mathbb{E}[D(S||\hat{S})]. \end{aligned}$$

□

We now show that exist natural estimators that achieve $r_n^{\text{nat}}(\Delta_k)$ and $r_n^{\mathbb{P}_\sigma}(\Delta_k)$.

Lemma 2.10. The exists a natural estimator q'' such that

$$r_n^{\text{nat}}(q'', \Delta_k) = r_n^{\text{nat}}(\Delta_k).$$

Similar there exists a natural estimator q' such that

$$r_n^{\mathbb{P}_\sigma}(q', \Delta_k) = r_n^{\mathbb{P}_\sigma}(\Delta_k).$$

Proof. We prove the result for $r_n^{\text{nat}}(\Delta_k)$. The result for $r_n^{\mathbb{P}_\sigma}(\Delta_k)$ is similar and omitted. Let profile $\bar{\varphi}$ of a sequence x^n be the vector of its prevalences i.e., $\bar{\varphi}(x^n) \stackrel{\text{def}}{=} (\varphi_0(x^n), \varphi_1(x^n), \varphi_2(x^n), \ldots \varphi_n(x^n))$. For any optimal estimator q and sequence $x^n y$ such that $\bar{\varphi}(x^n) = \bar{\varphi}_n$ and $n_y(x^n) = t$, let

$$q''_y(x^n) = \frac{\sum_{w^n z : \bar{\varphi}(w^n) = \bar{\varphi}_n, n_z = t} q_z(w^n)}{\sum_{u^n v : \bar{\varphi}(u^n) = \bar{\varphi}_n, n_v = t} 1}.$$

q'' is a natural estimator as if for any sequence x^n, $n_y(x^n) = n_{y'}(x^n)$, then $q''_y(x^n) = q''_{y'}(x^n)$. We show that q'' is an optimal estimator. Observe that for any $P \in \mathbb{P}_\sigma$

$$r_n(q,P) \stackrel{(a)}{\geq} \frac{1}{k!} \sum_{p \in P} r_n(q,p) \stackrel{(b)}{\geq} \frac{1}{k!} \sum_{p \in P} r_n(q'',p) \stackrel{(c)}{=} r_n(q'',P). \tag{2.8}$$

Maximum is larger than average and hence (a). Every distribution in P has the

same KL loss for q'' and hence (c). To prove (b), observe that

$$\begin{aligned}
\sum_{p\in P} r_n(q,p) &= \sum_{p\in P}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p)\\
&= \sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}}\sum_{p\in P} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p)\\
&= \sum_{\bar\varphi_n,t}\sum_{x^n:\bar\varphi(x^n)=\bar\varphi_n}\sum_{y:n_y=t}\sum_{p\in P} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p)\\
&\overset{(d)}{\geq} \sum_{\bar\varphi_n,t}\sum_{x^n:\bar\varphi(x^n)=\bar\varphi_n}\sum_{y:n_y=t}\sum_{p\in P} p(x^n y)\log\frac{\sum_{u^n,v:\bar\varphi(u^n)=\bar\varphi_n,n_v=t}1}{\sum_{w^n,z:\bar\varphi(w^n)=\bar\varphi_n,n_z=t} q_z(w^n)} - H(p)\\
&= \sum_{\bar\varphi_n,t}\sum_{x^n:\bar\varphi(x^n)=\bar\varphi_n}\sum_{y:n_y=t}\sum_{p\in P} p(x^n y)\log\frac{1}{{q''}_y(x^n)} - H(p)\\
&= \sum_{p\in P} r_n(q'',p),
\end{aligned}$$

For all sequences $x^n y$ with the same $\bar\varphi(x^n)$ and $n_y(x^n)$, $\sum_{p\in P} p(x^n y)$ is the same. Hence, applying log-sum inequality results in (d). By Lemma 2.9, every $p \in P$ has the same $r_n^{\text{nat}}(p)$, hence subtracting $r_n^{\text{nat}}(p)$ from both sides of Equation (2.8) results in

$$\max_{p\in P}\left(r_n(q,p) - r_n^{\text{nat}}(p)\right) \geq \max_{p\in P}\left(r_n(q'',p) - r_n^{\text{nat}}(p)\right).$$

Hence for the optimal estimator q,

$$\begin{aligned}
r_n^{\text{nat}}(\Delta_k) &= \max_{p\in\Delta_k}\left(r_n(q,p) - r_n^{\text{nat}}(p)\right)\\
&= \max_{P\in\mathbb{P}_\sigma}\left(\max_{p\in P}\left(r_n(q,p) - r_n^{\text{nat}}(p)\right)\right)\\
&\geq \max_{P\in\mathbb{P}_\sigma}\left(\max_{p\in P}\left(r_n(q'',p) - r_n^{\text{nat}}(p)\right)\right)\\
&= \max_{p\in\Delta_k}\left(r_n(q'',p) - r_n^{\text{nat}}(p)\right)\\
&= r_n(q'',\Delta_k).
\end{aligned}$$

Thus q'' is an optimal estimator and furthermore it is natural, hence the lemma. □

Thus there is an optimal estimator that achieves $r_n^{\text{nat}}(\Delta_k)$ and is natural. In Equation (2.7), taking maximum over all distributions p and minimum over all estimators q results in

Lemma 2.11. For a natural estimator q let $\hat{S}_t = \sum_{x:N_x=t} q_x$, then

$$r_n^{\text{nat}}(q, \Delta_k) = \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})].$$

Furthermore,

$$r_n^{\text{nat}}(\Delta_k) = \min_{\hat{S}} \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})].$$

Lemmas 2.7, 2.10, and 2.11 yield Theorem 2.4.

2.6 Lower bounds

To lower bound $r_n^{\mathbb{P}_\sigma}(\Delta_k)$ it is sufficient to lower bound $r_n^{\mathbb{P}_\sigma}(\mathcal{P})$ for any subset $\mathcal{P} \subseteq \Delta_k$. We construct a subset $\mathcal{P}$ by considering a set of distributions $\{p^{\mathbf{v}} : \mathbf{v} \in \{-1, 1\}^{m-1}\}$ and all their possible permutations. The lower bound argument uses Fano's inequality and Gilbert Varshamov bounds.

We choose $\mathcal{P}$ to be the set of distributions whose probability multiset are close to that of a distribution p^0, where p^0 is defined as follows.

Let c be a sufficiently large constant. Let m be the largest odd number less than $\min(k, (n/(c^2 \log^2 n))^{1/3})$. Let p^0 be the following distribution. For $1 \leq i \leq m-1$,

$$p_i^0 = \frac{\log n}{6n} \sqrt{\frac{c^2 n}{m}} \left(\sqrt{\frac{n}{c^2 m \log^2 n}} + i \right)$$

and $p_m^0 = 1 - \sum_{i=1}^{m-1} p_i^0$. Observe that for all $1 \leq i \leq m-1$, $1/(6m) \leq p_i^0 \leq 1/(3m)$ and $p_m^0 \geq 2/3$.

We choose the close-by distributions as follows. Let $\epsilon = \sqrt{\frac{c^*}{mn}}$, where c^* is some sufficiently small constant. For a binary vector $\mathbf{v} \in \{-1, 1\}^{m-1}$, let $p^{\mathbf{v}}$ be the distribution such that $p_i^{\mathbf{v}} = p_i^0 + \mathbf{v}_i \epsilon$ for $1 \leq i \leq m-1$ and $p^{\mathbf{v}}(m) = 1 - \sum_{i=1}^{m-1} p_i^{\mathbf{v}}$. Note that by the properties of p^0 and ϵ, $p^{\mathbf{v}}$ is a valid distribution for every $\mathbf{v}$. Let $\mathcal{C}$ be the largest subset of $\{-1, 1\}^{m-1}$ such that for every $\mathbf{v} \in \mathcal{C}$, $\sum_i \mathbf{v}_i = 0$ and for every pair $\mathbf{v}, \mathbf{v}' \in \mathcal{C}$, $\sum_i |\mathbf{v}_i - \mathbf{v}'_i| \geq c'(m-1)$ for some constant c'. The following variation of Gilbert Varshamov lemma lower bounds size of $\mathcal{C}$.

Lemma 2.12. There exists a set of vectors $\mathcal{C}$ over $\{-1, 1\}^{m-1}$ of size $2^{c''\cdot(m-1)}$ such that the minimum hamming distance between any two vectors is $\geq c'(m-1)$ for some universal constants $c' > 0, c'' > 0$ and $\sum_i \mathbf{v}_i = 0$ for all $\mathbf{v} \in \mathcal{C}$.

Let $\mathcal{P}' = \{p^{\mathbf{v}} : \mathbf{v} \in \mathcal{C}\}$ and $P_{\mathbf{v}} = \{p^{\mathbf{v}}(\sigma(\cdot)) : \sigma \in \Sigma^{m-1}\}$ be the set of all permutations of a distribution $p^{\mathbf{v}}$, i.e., all distributions with the same multiset as $p^{\mathbf{v}}$. Let

$$\mathcal{P} = \cup_{\mathbf{v}\in\mathcal{C}} P_{\mathbf{v}}.$$

We first bound the regret of the induced permutation class $P_{\mathbf{v}}$ that contains all permutations of a distribution $p^{\mathbf{v}}$.

Lemma 2.13. For every induced permutation class $P_{\mathbf{v}}$,

$$r_n(P_{\mathbf{v}}) \leq \frac{1}{n}.$$

Proof. We prove the bound by constructing an estimator q. Consider the estimator q which sorts the multiplicities and assigns the i^{th}-frequently occurred symbol probability $p_i^{\mathbf{v}}$. Since this is a natural estimator, it occurs the same loss for all distributions in $P_{\mathbf{v}}$ and hence,

$$\begin{aligned}
r_n(P_{\mathbf{v}}) &\leq \max_{p\in P_{\mathbf{v}}} \mathbb{E}[D(p||q)] \\
&= \mathbb{E}[D(p^{\mathbf{v}}||q)] \\
&\overset{(a)}{\leq} \Pr(\exists i, j : N_i > N_j,\, p_i^{\mathbf{v}} < p_j^{\mathbf{v}}) \log n \\
&\overset{(b)}{\leq} \binom{m}{2} e^{-2\log n} \log n \\
&\leq \frac{1}{n}.
\end{aligned}$$

(a) follows from the fact that the estimator makes an error only if two multiplicities cross over and if it does make an error, the maximum KL divergence is at most $\log(p_{\max}/p_{\min}) \leq \log n$. Since probabilities for any two symbols i and j differ by at least $\frac{\log n}{6n} \cdot \sqrt{\frac{c^2 n}{m}}$ and the probabilities themselves lie between $1/(6m)$ and $1/(3m)$, by choosing a sufficiently large c, the cross over probability can be bounded by $e^{-2\log n}$ using the Chernoff bound and hence (b). □

We now lower bound the KL divergence between $p^{\mathbf{v}}$ and $p^{\mathbf{v}'}$ for every pair of vectors $\mathbf{v}$ and $\mathbf{v}'$. Let the Hamming distance between two vectors $\mathbf{v}$ and $\mathbf{v}'$ be $||\mathbf{v} - \mathbf{v}'||_1 = \sum_{i=1}^{m-1} |\mathbf{v}_i - \mathbf{v}'_i|$.

Lemma 2.14. For two distributions $p^{\mathbf{v}}$ and $p^{\mathbf{v}'}$ in $\mathcal{P}'$,

$$\frac{1}{8}\left(c'\sqrt{\frac{mc^*}{n}}\right)^2 \leq \frac{1}{2}\left|\left|p^{\mathbf{v}} - p^{\mathbf{v}'}\right|\right|_1^2 \leq D(p^{\mathbf{v}}||p^{\mathbf{v}'}) \leq \frac{48mc^*}{n}.$$

Proof.

$$\begin{aligned}
D(p^{\mathbf{v}}||p^{\mathbf{v}\prime}) &\overset{(a)}{\leq} \sum_{i=1}^{m} \frac{(p_i^{\mathbf{v}} - p_i^{\mathbf{v}\prime})^2}{p_i^{\mathbf{v}\prime}} \\
&\overset{(b)}{\leq} 2\sum_{i=1}^{m} \frac{(p_i^{\mathbf{v}} - p_i^{\mathbf{v}\prime})^2}{p_i^0} \\
&\leq 2\sum_{i=1}^{m-1} \frac{(\mathbf{v}_i - \mathbf{v}_i')^2(\sqrt{c^*/nm})^2}{1/(6m)} \\
&\leq 12\sum_{i=1}^{m-1} \frac{(\mathbf{v}_i - \mathbf{v}_i')^2 c^*}{n} \\
&\leq 24\sum_{i=1}^{m-1} \frac{|\mathbf{v}_i - \mathbf{v}_i'| c^*}{n} \\
&= \frac{24\,||\mathbf{v} - \mathbf{v}'||_1\, c^*}{n} \\
&\leq \frac{48mc^*}{n}.
\end{aligned}$$

(a) follows from bounding the KL divergence by the Chi-squared distance and (b) follows from the fact that $\epsilon \ll 1/m$. For the lower bound,

$$\begin{aligned}
D(p^{\mathbf{v}}||p^{\mathbf{v}'}) &\overset{(a)}{\geq} \frac{1}{2}\left|\left|p^{\mathbf{v}} - p^{\mathbf{v}'}\right|\right|_1^2 \\
&= \frac{1}{2}\left(\frac{||\mathbf{v} - \mathbf{v}'||_1\sqrt{c^*}}{\sqrt{mn}}\right)^2 \\
&\overset{(b)}{\geq} \frac{1}{2}\left(\frac{c'(m-1)\sqrt{c^*}}{\sqrt{mn}}\right)^2 \\
&\overset{(c)}{\geq} \frac{1}{8}\left(c'\sqrt{\frac{mc^*}{n}}\right)^2,
\end{aligned}$$

where (a) follows from Pinsker's inequality, (b) follows by construction, and $m-1 \geq 2$ and hence (c). □

We now state Fano's inequality for distribution estimation.

Lemma 2.15. Let $p^1, p^2, \dots p^{r+1}$ be distributions such that $D(p^i||p^j) \leq \beta$ and $||p^i - p^j||_1 \geq \alpha$, for all i, j. For any estimator q,

$$\sup_i \mathbb{E}_i[\left|\left|p^i - q\right|\right|_1] \geq \frac{\alpha}{2}\left(1 - \frac{n\beta + \log 2}{\log r}\right).$$

We now have all the tools for the lower bound.

Proof of Theorem 2.5. For every permutation subclass $P_{\mathbf{v}}$ in $\mathcal{P}$, by Lemma 2.13

$$r_n(P_{\mathbf{v}}) \leq \frac{1}{n}.$$

Thus,

$$\begin{aligned}
r_n^{\mathbb{P}_\sigma}(\mathcal{P}) &= \min_q \max_{\mathbf{v}} \big(\max_{p \in P_{\mathbf{v}}} r_n(q,p) - r_n(P_{\mathbf{v}})\big) \\
&\geq \min_q \max_{\mathbf{v}} \Big(\max_{p \in P_{\mathbf{v}}} r_n(q,p) - \frac{1}{n}\Big) \\
&= \min_q \max_{p \in \mathcal{P}} r_n(q,p) - \frac{1}{n} \\
&= \min_q \max_{p \in \mathcal{P}} \mathbb{E}[D(p||q)] - \frac{1}{n} \\
&\overset{(a)}{\geq} \min_q \max_{p \in \mathcal{P}'} \mathbb{E}[D(p||q)] - \frac{1}{n} \\
&\overset{(b)}{\geq} \min_q \max_{p \in \mathcal{P}'} \mathbb{E}\left[\frac{||p-q||_1^2}{2}\right] - \frac{1}{n} \\
&\overset{(c)}{\geq} \min_q \max_{p \in \mathcal{P}'} \frac{1}{2}\mathbb{E}\left[||p-q||_1\right]^2 - \frac{1}{n} \\
&\overset{(d)}{\geq} \Omega\left(\frac{m}{n}\right) - \frac{1}{n} \\
&\geq \Omega\left(\frac{m}{n}\right).
\end{aligned}$$

$\mathcal{P}' \subset \mathcal{P}$, hence (a). (b) follows from Pinsker's inequality and (c) follows from convexity. By construction, for every pair of distributions in $\mathcal{P}'$, $\beta = D(p||p') \leq 48c^*m/n$ and $\alpha = ||p - p'||_1 \geq \Omega(\sqrt{m/n})$ (Lemma 2.14). Furthermore by Lemma 2.12, $\mathcal{P}'$ has $r + 1 = 2^{c''(m-1)}$ distributions. Setting c^* to be a sufficiently small constant and applying Lemma 2.15 to $\mathcal{P}'$ with the above values of α, β, and r results in (d). Substituting the value of m in the above equation results in the Theorem. □

Acknowledgement

Chapter 2 is adapted from Alon Orlitsky and Ananda Theertha Suresh, "Competitive distribution estimation: Why is Good-Turing good", *Neural Information Processing Systems (NIPS),* 2015 [37].

Chapter 3

Combined-probability mass estimation

3.1 Introduction

In the last chapter, we related the problem of competitive distribution estimation to the problem of min-max combined-probability estimation, where the combined probability mass

$$S_t \stackrel{\text{def}}{=} S_t(X^n) \stackrel{\text{def}}{=} \sum_{x:N_x=t} p_x$$

is the sum of probabilities of symbols appearing t times. For instance, if $p_a = .3$, $p_b = .1$, $p_n = .35$, $p_s = .15$, and the sum of all other letter probabilities is .1, then for b, a, n, a, n, a, s, $S_1 = p_b + p_s = .25$, $S_2 = p_n = .35$, $S_3 = p_a = .3$, and $S_0 = p_c + p_d + \ldots + p_z = .1$.

In this chapter we compute the min-max combined-probability estimation and provide a linear estimator that achieves it. We also discuss its implications to the problem of pattern prediction.

3.2 Previous results

The problem of combined-probability mass estimation was first studied by [14], who noted that reasonable estimators assign the same probability to all sym-

bols appearing the same number of times in a sample. Let $\mathbb{1}_{N_x=t}$ be the indicator function that is 1 iff $N_x = t$. Recall that Φ_t denotes the number of symbols appearing t times in a sample of size n. The most natural estimator for S_t is the empirical frequency estimator that estimates S_t by

$$E_t = \frac{t}{n} \cdot \Phi_t.$$

The Good-Turing estimator proposed in [14], estimates S_t by

$$G_t \stackrel{\text{def}}{=} \frac{t+1}{n} \cdot \Phi_{t+1}. \tag{3.1}$$

The Good-Turing estimator is an important tool in a number of language processing applications,*e.g.,* [10]. However for several decades it defied rigorous analysis, partly because of the dependencies between N_x for different x's. First theoretical results were provided by [19]. Using McDiarmid's inequality [38], they showed that for all 0, with probability $\geq 1 - \delta$,

$$|G_t - S_t| = \mathcal{O}\left(\sqrt{\frac{\log(3/\delta)}{n}}\left(t + 1 + \log\frac{n}{\delta}\right)\right).$$

Note that this bound, like all subsequent ones in this chapter, holds uniformly, namely applies to all support sets $\mathcal{X}$ and all distributions p over $\mathcal{X}$.

To express this and subsequent results more succinctly, we will use several abbreviations. Recall that $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ hide poly-logarithmic factors in n and $1/\delta$. For a random variable X, we will use

$$X \underset{\delta}{=} \widetilde{\mathcal{O}}(\alpha) \quad \text{to abbreviate} \quad \Pr\left(X \neq \widetilde{\mathcal{O}}(\alpha)\right) \leq \delta,$$

and similarly $X \underset{\delta}{=} \widetilde{\Omega}(\alpha)$ for $\Pr\left(X \neq \widetilde{\Omega}(\alpha)\right) < \delta$. For example, the above bound becomes

$$|G_t - S_t| \underset{\delta}{=} \widetilde{\mathcal{O}}\left(\frac{t+1}{\sqrt{n}}\right).$$

As could be expected, most applications require simultaneous approximation of S_t over a wide range of t's. For example, as shown in Section 4.1, classification requires approximating $S_0, \ldots, S_n$ to within a small ℓ_1 distance, while prediction requires approximation to within a small KL-Divergence.

[20] improved the Good-Turing bound and combined it with the empirical estimator to obtain an estimator G' with ℓ_∞ convergence,

$$||G' - S||_\infty \stackrel{\text{def}}{=} \max_{0 \le t \le n} |G'_t - S_t| \underset{\delta}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n^{0.4}}\right).$$

Subsequently, [39] considered ℓ_1 convergence for a subclass of distributions where all symbols probabilities are proportional to $1/n$, namely for some constants c_1, c_2, all probabilities p_x are in the range $[c_1/n, c_2/n]$. Recently, [40] showed that the Good-Turing estimator is not uniformly multiplicatively consistent over all distributions, and described a class of distributions for which it is.

3.3 New results

In practice, often the Good-Turing estimator is used for small multiplicities and empirical estimators are used for large multiplicities. We analyze this estimator and bound its regret. For a symbol appearing t times in a random sequence X^n, we assign probability C_t/Φ_t,

$$C_t = \begin{cases} \Phi_t \cdot \frac{t}{nN} & \text{if } t \ge t_0, \\ (\Phi_{t+1} + 1) \cdot \frac{t+1}{nN} & \text{else,} \end{cases}$$

where N is the normalization factor to ensure that $\sum_{t=0}^{\infty} C_t = 1$ and We set $t_0 \propto n^{1/3}$ later. Similar to our experiments, we have modified the Good-Turing estimator to $(\Phi_{t+1} + 1) \cdot \frac{t+1}{n}$, thus ensuring that we never assign a non-zero probability. However, unlike our experiments, where we decided between empirical and Good-Turing estimators depending on if $\Phi_{t+1} \ge t$, for our proofs we just decide it based on t for convenience. We remark that in our experiments the estimator in Section 5.6 performed better than the one above. For this estimator, we prove that

Theorem 3.1. For any k and n, upon observing $n' \sim \text{poi}(n)$ samples,

$$\max_{p \in \Delta_k} \mathbb{E}[D(S||C)] \le \frac{3 + o_n(1)}{n^{1/3}},$$

and hence by Pinsker's inequality,

$$\max_{p \in \Delta_k} \mathbb{E}[||S - C||_1] \le \frac{\sqrt{6} + o_n(1)}{n^{1/6}}.$$

We then show that the above bound are tight in that no simple combination of G_t and the empirical estimator E_t can approximate S_t better. The proof is provided in Section 3.5.3.

Lemma 3.2. For every n, there is a distribution such that

$$\sum_{t=0}^{n} \min\left(|E_t - S_t|, |G_t - S_t|\right) \underset{1/n}{=} \widetilde{\Omega}\left(\frac{1}{n^{1/6}}\right).$$

In Subsections 3.6.1–3.6.3, we construct a new estimator F'_t and show that it estimates S_t better than G_t and essentially as well as any other estimator. A closer inspection of Good and Turing's intuition in [9] shows that the average probability of a symbol appearing t times is

$$\frac{S_t}{\Phi_t} \approx \frac{t+1}{n} \cdot \frac{\mathbb{E}[\Phi_{t+1}]}{E[\Phi_t]}. \tag{3.2}$$

If we were given the values of the $\mathbb{E}[\Phi_t]$'s, we could use this equation to estimate the S_t's. Since we are not given these values, Good-Turing (3.1) approximates the expectation ratio by just Φ_{t+1}/Φ_t. However, while Φ_t and Φ_{t+1} are by definition unbiased estimators of their expectations $\mathbb{E}[\Phi_t]$ and $\mathbb{E}[\Phi_{t+1}]$ respectively, their variance is high, leading to a probability estimation G_t that may be far from S_t.

In Section 3.6.2 of $\mathbb{E}[\Phi_t]$ by expressing it as a linear combination of the values of $\Phi_{t'}$ for t' near t. Lemma 3.22 shows that an appropriate choice of the smoothing coefficients yields an estimate $\widehat{\mathbb{E}[\Phi_t]}$ that approximates $\mathbb{E}[\Phi_t]$ well.

Incorporating this estimate into Equation (3.2), yields a new estimator F_t. Combining it with the empirical and Good-Turing estimators for different ranges of t and Φ_t, we obtain a modified estimator F'_t that has a small KL divergence from S_t, and hence by Pinsker's inequality, also small ℓ_1 distance uniformly over all distributions.

Theorem 3.3. For every distribution and every n, the proposed estimator F' satisfies

$$D(S||F') \underset{1/n}{=} \widetilde{\mathcal{O}}_n\left(\min\left(\frac{1}{n^{1/2}}, \frac{k}{n}\right)\right),$$

and hence by Pinsker's inequality,

$$||F' - S||_1 \underset{1/n}{=} \widetilde{\mathcal{O}}_n\left(\min\left(\frac{1}{n^{1/4}}, \frac{\sqrt{k}}{\sqrt{n}}\right)\right).$$

In Section 3.6.5 we show that the proposed estimator is optimal. An estimator is *label-invariant*, often called *canonical*, if its estimate of S_t remains unchanged under all permutations of the symbol labels. For example, its estimate of S_1 will be the same for the sample a, a, b, b, c as it is for u, u, v, v, w. Clearly all reasonable estimators are label-invariant.

Theorem 3.4. For any label-invariant estimator $\widehat{S}$, there is a distribution such that

$$D(S||\widehat{S}) \underset{1/n}{=} \widetilde{\Omega}\left(\frac{1}{n^{1/2}}\right)$$

hence by Pinsker's inequality,

$$||\widehat{S} - S||_1 \underset{1/n}{=} \widetilde{\Omega}\left(\frac{1}{n^{1/4}}\right).$$

Finally we note that the estimator F'_t can be computed in time linear in n. Also, observe that while the difference between ℓ_1 distance of $1/n^{1/6}$ and $1/n^{1/4}$ may seem small, an equivalent formulation of the results would ask for the number of samples needed to estimate within a ℓ_1 distance ϵ. Good-Turing and empirical frequency would require $(1/\epsilon)^6$ samples, while the estimator we construct needs $(1/\epsilon)^4$ samples. For $\epsilon = 1\%$, the difference between the two is a factor of 10,000.

The rest of the chapter is organized as follows: In Section 3.4, we introduce Poisson sampling, a tool that simplifies the analysis and present some preliminary results. In Section 3.5, we analyze the performance of the simple combination of Good-Turing and empirical estimators. In Section 3.6 we motivate and propose the improved estimator and in Section 3.7 we present its analysis. We Finally remark that Theorem 3.3 implies a faster algorithm for *pattern prediction* which we define and prove in Section 3.8.

3.4 Poisson sampling and preliminaries

The analysis of combined-probability estimators rely on computing the variances and expectations of prevalences. A standard tool to simplify the analysis is Poisson sampling.

In the standard sampling method, where a distribution is sampled n times, the multiplicities are dependent. Analysis of functions of dependent random variables requires various concentration inequalities, which often complicates the proofs. A useful approach to make them independent and hence simplify the analysis is to do Poisson sampling. The distribution is sampled a random n' times, where n' is a Poisson random variable with parameter n.

The following fact, mentioned without proof states that the multiplicities are independent under Poisson sampling.

Lemma 3.5 ([41])**.** If a distribution p is sampled *i.i.d.* poi(n) times, then the number of times symbol x appears is an independent Poisson random variable with mean np_x, namely, $\Pr(N_x = t) = \frac{e^{-np_x}(np_x)^t}{t!}$.

We first show the following lemma, which shows that proving properties for poi(n) sampling implies properties for sampling the distribution exactly n times. Hence in the rest of the chapter, we prove the properties of an estimator under Poisson sampling.

Lemma 3.6 ([41])**.** If when a distribution is sampled poi(n) times, a certain property holds with probability $\geq 1 - \delta$, then when the distribution is sampled exactly n times, the property holds with probability $\geq 1 - \delta \cdot e\sqrt{n}$.

Proof. If a distribution is sampled $n' = \text{poi}(n)$ times, with probability $e^{-n}\frac{n^n}{n!} \geq \frac{1}{e\sqrt{n}}$, $n' = n$. Conditioned on the fact that $n' = n$, Poisson sampling is same as sampling the distribution exactly n times. Therefore, if P fails with probability $> \delta \cdot e\sqrt{n}$ with exactly n samples, then P fails with probability $> \delta$ when sampled poi(n) times. □

To illustrate the advantages of Poisson sampling, we first show that Good-Turing estimator is unbiased under Poisson sampling. We use this fact to get a better understanding of the proposed estimator.

Lemma 3.7. For every distribution p and every t,

$$\mathbb{E}[G_t] = \frac{t+1}{n}\mathbb{E}[\Phi_{t+1}] = \mathbb{E}[S_t].$$

Proof. The proof follows from the fact that each multiplicity is a Poisson random variable under Poisson sampling.

$$\begin{aligned}\mathbb{E}[S_t] &= \mathbb{E}\Big[\sum_x p_x \cdot \mathbb{1}_{N_x=t}\Big] \\ &= \sum_x p_x \cdot e^{-np_x}\frac{(np_x)^t}{t!} \\ &= \frac{t+1}{n}\sum_x e^{-np_x}\frac{(np_x)^{t+1}}{(t+1)!} = \frac{t+1}{n}\mathbb{E}[\Phi_{t+1}].\end{aligned}$$

□

The next lemma bounds the variance of any linear estimator in terms of its coefficients.

Lemma 3.8. For every distribution p,

$$\mathrm{Var}\left(\sum_x\sum_t \mathbb{1}_{N_x=t}f(x,t)\right) \le \sum_x\sum_t \mathbb{E}[\mathbb{1}_{N_x=t}]f^2(x,t).$$

Proof. By Poisson sampling, the multiplicities are independent. Furthermore the variance of sum of independent random variables is the sum of their variances. Hence,

$$\begin{aligned}\mathrm{Var}\left(\sum_x\sum_t \mathbb{1}_{N_x=t}f(x,t)\right) &= \sum_x \mathrm{Var}\left(\sum_t \mathbb{1}_{N_x=t}f(x,t)\right) \\ &\le \sum_x \mathbb{E}\left[\left(\sum_t \mathbb{1}_{N_x=t}f(x,t)\right)^2\right] \\ &\overset{(a)}{=} \sum_x \mathbb{E}\left[\sum_t (\mathbb{1}_{N_x=t}f(x,t))^2\right] \\ &\overset{(b)}{=} \sum_x\sum_t \mathbb{E}[\mathbb{1}_{N_x=t}]f^2(x,t).\end{aligned}$$

For $t \neq t'$, $\mathbb{E}[\mathbb{1}_{N_x=t}\mathbb{1}_{N_x=t'}] = 0$ and hence (a). (b) uses the fact that $\mathbb{1}_{N_x=t}$ is an indicator random variable. □

The above two lemmas immediately imply that for any n and t.

$$\mathrm{Var}(S_t) \le \frac{(t+1)(t+2)}{n^2}\cdot\mathbb{E}[\Phi_{t+2}]. \tag{3.3}$$

$$\mathbb{E}\left[\left(S_t - \frac{(t+1)\Phi_{t+1}}{n}\right)^2\right] \le \frac{(t+1)(t+2)\mathbb{E}[\Phi_{t+2}]}{n^2} + \frac{(t+1)^2\mathbb{E}[\Phi_{t+1}]}{n^2}. \tag{3.4}$$

3.5 Regret bound on the Good-Turing estimator

We now prove the performance of the simple Good-Turing estimator under Poisson sampling. We first relate the KL regret to a chi-squared like distance between the combined probability mass S and the un-normalized estimate.

Lemma 3.9. For any distribution $p \in \Delta_k$,

$$\mathbb{E}[D(S||C)] \leq \sum_{t=0}^{t_0-1} \mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] + \sum_{t=t_0}^{\infty} \mathbb{E}\left[\frac{(S_t - t\Phi_t/n)^2}{\Phi_t t/n}\right].$$

Proof. Let $C'_t = C_t N$. Since $\log(1+y) \leq y$, $\sum_{t=0}^{\infty} C_t = 1$, and $\sum_{t=0}^{\infty} C'_t = N$,

$$\begin{aligned}
D(S||C) &= \sum_{t=0}^{\infty} S_t \log \frac{S_t}{C_t} \\
&= \sum_{t=0}^{\infty} S_t \log \frac{N S_t}{C'_t} \\
&= \sum_{t=0}^{\infty} S_t \log \frac{S_t}{C'_t} + \sum_{t=0}^{\infty} S_t \log N \\
&= \sum_{t=0}^{\infty} S_t \log \left(1 + \frac{S_t - C'_t}{C'_t}\right) + \log N \\
&\leq \sum_{t=0}^{\infty} S_t \left(\frac{S_t - C'_t}{C'_t}\right) + \log N \\
&= \sum_{t=0}^{\infty} (S_t - C'_t) \left(\frac{S_t - C'_t}{C'_t}\right) + \sum_{t=0}^{\infty} C'_t \left(\frac{S_t - C'_t}{C'_t}\right) + \log N \\
&= \sum_{t=0}^{\infty} (S_t - C'_t) \left(\frac{S_t - C'_t}{C'_t}\right) + \sum_{t=0}^{\infty} (S_t - C'_t) + \log N \\
&= \sum_{t=0}^{\infty} \frac{(S_t - C'_t)^2}{C'_t} + 1 - N + \log N \\
&\leq \sum_{t=0}^{\infty} \frac{(S_t - C'_t)^2}{C'_t} \\
&= \sum_{t=0}^{t_0-1} \frac{(S_t - C'_t)^2}{C'_t} + \sum_{t=t_0}^{\infty} \frac{(S_t - C'_t)^2}{C'_t}.
\end{aligned}$$

Taking expectations on both sides and substituting C'_t results in the lemma. □

3.5.1 Empirical estimators

All of our results including the next lemma hold for all distributions in Δ_k and hence stated without any condition on the underlying distribution.

Lemma 3.10. For any n and t_0,

$$\sum_{t=t_0}^{\infty} \mathbb{E}\left[\frac{(S_t - t\Phi_t/n)^2}{\Phi_t t/n}\right] \leq \frac{1}{t_0}.$$

Proof.

$$\begin{aligned}
\sum_{t=t_0}^{\infty} \frac{(S_t - t\Phi_t/n)^2}{\Phi_t t/n} &\leq \sum_{t=t_0}^{\infty} \frac{(S_t - t\Phi_t/n)^2}{\Phi_t t_0/n} \\
&\overset{(a)}{\leq} \sum_{t=t_0}^{\infty} \sum_{x} \mathbb{1}_{N_x=t} \frac{(p_x - t/n)^2}{t_0/n} \\
&= \sum_{x} \sum_{t=t_0}^{\infty} \mathbb{1}_{N_x=t} \frac{(p_x - t/n)^2}{t_0/n} \\
&\leq \sum_{x} \sum_{t=0}^{\infty} \mathbb{1}_{N_x=t} \frac{(p_x - t/n)^2}{t_0/n}.
\end{aligned}$$

(a) follows from the fact that $\frac{(\sum_{x=1}^{m} a_x)^2}{m} \leq \sum_{i=1}^{m} a_x^2$ for $a_x = \mathbb{1}_{N_x=t}(p_x - t/n)$ and $m = \Phi_t$. Taking expectations on both sides,

$$\begin{aligned}
\sum_{t=t_0}^{\infty} \mathbb{E}\left[\frac{(S_t - t\Phi_t/n)^2]}{\Phi_t t/n}\right] &\leq \sum_{x} \frac{\mathbb{E}[\sum_{t=0}^{\infty} \mathbb{1}_{N_x=t}(p_x - t/n)^2]}{t_0/n} \\
&\leq \sum_{x} \frac{p_x/n}{t_0/n} \\
&= \frac{1}{t_0},
\end{aligned}$$

where the second inequality follows from observing that $\mathbb{E}[\sum_{t=0}^{\infty} \mathbb{1}_{N_x=t}(p_x - t/n)^2]$ is the variance of a Binomial random variable with parameters n and p_x. □

3.5.2 Good-Turing estimators

To bound the regret corresponding to the Good-Turing estimator, we need few auxiliary results. The next lemma relates $\mathbb{E}[\Phi_{t+1}]$ to $\mathbb{E}[\Phi_t]$.

Lemma 3.11. For any n and $t \geq 1$,

$$\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}[\Phi_t]\left(\frac{2}{t}\log n + \frac{t}{t+1}\right) + \frac{1}{t+1}.$$

Proof. Let $r \geq \frac{t}{t+1}$.

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1}] &= \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=t+1}\right] \\
&= \sum_x e^{-np_x}\frac{(np_x)^{t+1}}{(t+1)!} \\
&= \sum_x \frac{n}{t+1}\cdot e^{-np_x}\frac{(np_x)^t}{t!}p_x \\
&= \sum_{x:np_x\leq r(t+1)} \frac{n}{t+1}\cdot e^{-np_x}\frac{(np_x)^t}{t!}p_x + \sum_{x:np_x>r(t+1)} \frac{n}{t+1}\cdot e^{-np_x}\frac{(np_x)^t}{t!}p_x \\
&\overset{(a)}{\leq} r\sum_{x:np_x\leq r(t+1)} e^{-np_x}\frac{(np_x)^t}{t!} + \sum_{x:np_x>r(t+1)} \frac{n}{t+1}e^{-r(t+1)}\frac{(r(t+1))^t}{t!}p_x \\
&\leq r\sum_x e^{-np_x}\frac{(np_x)^t}{t!} + \sum_x \frac{n}{t+1}e^{-r(t+1)}\frac{(r(t+1))^t}{t!}p_x \\
&\overset{(b)}{\leq} r\sum_x e^{-np_x}\frac{(np_x)^t}{t!} + \sum_x \frac{n}{t+1}e^{-rt/2}p_x \\
&\leq r\mathbb{E}[\Phi_t] + \frac{n}{t+1}e^{-\frac{rt}{2}}.
\end{aligned}$$

(a) follows from the fact that second term is a decreasing as a function of np_x in the range $[r(t+1), \infty)$. (b) follows from the fact that

$$e^{-r(t+1)}\frac{(r(t+1))^t}{t!} = e^{-rt}r^t \cdot e^{-t}\frac{(t+1)^t}{t!} \leq e^{-rt}r^t \leq e^{-rt/2}.$$

Choosing $r = \frac{2}{t}\log n + \frac{t}{t+1}$, yields

$$\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}[\Phi_t]\left(\frac{2}{t}\log n + \frac{t}{t+1}\right) + \frac{1}{t+1}.$$

□

The final auxiliary lemma bounds the inverse moment of Poisson binomial distributions.

Lemma 3.12. Let X_i for $1 \leq i \leq n$ be Bernoulli random variables, then

$$\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} X_i + 1}\right] \leq \frac{1}{\sum_{i=1}^{n} \mathbb{E}[X_i]}.$$

Proof. Let $r_i = \mathbb{E}[X_i]$. We show that of all tuples $r_1, r_2, \ldots, r_n$ such that $\sum_{i=1}^{n} r_i = nr$, the one that maximizes the expectation is $r_i = r, \forall i$. Suppose for some i, j, $r_i > r_j$, we show that if we decrease r_i and increase r_j keeping the sum same, then the expectation increases. Let $Y = 1 + \sum_{k \notin \{i,j\}} X_k$. For any instance of X^n, taking expectation with respect to only X_i and X_j.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{X_i + X_j + Y}\right] &= \frac{(1-r_i)(1-r_j)}{Y} + \frac{r_i(1-r_j) + (1-r_i)r_j}{Y+1} + \frac{r_i r_j}{Y+2} \\
&= \frac{1}{Y} + (r_i + r_j)\left(\frac{1}{Y+1} - \frac{1}{Y}\right) + r_i r_j \frac{2}{Y(Y+1)(Y+2)}.
\end{aligned}$$

Thus if we decrease r_i and increase r_j (keeping $r_i + r_j$ fixed), then $r_i r_j$ increases and hence the expectation increases. Hence the maximum occurs when $r_i = r_j$ for all i, j and

$$\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} X_i + 1}\right] \leq \mathbb{E}\left[\frac{1}{Z+1}\right],$$

where Z is a binomial random variable with parameters n and $r = \sum_{i=1}^{n} \mathbb{E}[X_i]/n$.

The expectation can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Z+1}\right] &= \sum_{j=0}^{n} \frac{1}{j+1} \binom{n}{j} r^j (1-r)^{n-j} \\
&= \frac{1}{(n+1)r} \sum_{j=0}^{n} \binom{n+1}{j+1} r^{j+1} (1-r)^{n+1-(j+1)} \\
&\leq \frac{1}{(n+1)r} \\
&\leq \frac{1}{nr} \\
&= \frac{1}{\sum_{i=1}^{n} \mathbb{E}[X_i]}.
\end{aligned}$$

□

Using the above lemma, we first bound the expectation of ${S_t}^2/(\Phi_{t+1} + 1)$.

Lemma 3.13. For any n and t, if $\mathbb{E}[\Phi_{t+1}] > 2$, then

$$\mathbb{E}\left[\frac{{S_t}^2}{\Phi_{t+1}+1}\right] \leq \frac{\mathbb{E}[{S_t}^2]}{\mathbb{E}[\Phi_{t+1}]-1}.$$

Proof. We first observe that for any x,

$$\mathbb{E}[\mathbb{1}_{N_x=t+1}] = e^{-np_x}\frac{(np_x)^{t+1}}{(t+1)!} \leq e^{-t-1}\frac{(t+1)^{t+1}}{(t+1)!} \leq \frac{1}{e}. \tag{3.5}$$

Since $S_t = \sum_x p_x \mathbb{1}_{N_x=t}$ and $\Phi_{t+1} = \sum_x \mathbb{1}_{N_x=t+1}$,

$$\frac{{S_t}^2}{\Phi_{t+1}+1} = \frac{\sum_x \sum_y p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{N_y=t}}{\sum_z \mathbb{1}_{N_z=t+1}+1} = \sum_x \sum_y \frac{p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{N_y=t}}{\sum_{z:z\neq x, z\neq y}\mathbb{1}_{N_z=t+1}+1},$$

where the equality follows from the fact that symbol cannot appear both t and $t+1$ times thus only one of $\mathbb{1}_{N_x=t}$ and $\mathbb{1}_{N_x=t+1}$ can be 1. The numerator and the denominator of the terms on RHS are independent of each other, hence

$$\begin{aligned}
\mathbb{E}\left[\frac{p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{N_y=t}}{\sum_z \mathbb{1}_{N_z=t+1}+1}\right] &= \mathbb{E}\left[\frac{p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{N_y=t}}{\sum_{z:z\neq x, z\neq y}\mathbb{1}_{N_z=t+1}+1}\right] \\
&= \mathbb{E}\left[p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{N_y=t}\right]\mathbb{E}\left[\frac{1}{\sum_{z:z\neq x, z\neq y}\mathbb{1}_{N_z=t+1}+1}\right] \\
&\overset{(a)}{\leq} \frac{\mathbb{E}\left[p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{N_y=t}\right]}{\sum_{z:z\neq x, z\neq y}\mathbb{E}[\mathbb{1}_{N_z=t+1}]} \\
&\overset{(b)}{\leq} \frac{\mathbb{E}\left[p_x p_y \mathbb{1}_{N_x=t}\mathbb{1}_{(N_y=t)}\right]}{\mathbb{E}[\Phi_{t+1}-1]},
\end{aligned}$$

(a) follows from Lemma 3.12 and (b) follows from Equation (3.5) as

$$\sum_{z:z\neq x, z\neq y}\mathbb{E}[\mathbb{1}_{N_z=t+1}] = \sum_z \mathbb{E}[\mathbb{1}_{N_z=t+1}] - \mathbb{E}[\mathbb{1}_{N_x=t+1}] - \mathbb{E}[\mathbb{1}_{N_y=t+1}] \geq \mathbb{E}[\Phi_{t+1}] - 1.$$

Summing over x and y results in the lemma. □

We now have all the tools to bound the error of the Good-Turing estimator. We divide the set of values into two groups, depending on the value of $\mathbb{E}[\Phi_{t+1}]$.

Lemma 3.14. For any n and t if $\mathbb{E}[\Phi_{t+1}] \leq 2$, then

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] \leq \frac{5t}{n} + \frac{4\log n}{n}\left(\frac{t+2}{t+1}\right) + \frac{6}{n}.$$

Proof. Let $Z = S_t - (t+1)\Phi_{t+1}/n$.

$$\begin{aligned}
&\mathbb{E}\left[\left(Z - \frac{t+1}{n}\right)^2\right] \\
&\overset{(a)}{=} \mathbb{E}[Z^2] + \frac{(t+1)^2}{n^2} \\
&\overset{(b)}{\leq} \frac{(t+1)(t+2)\mathbb{E}[\Phi_{t+2}]}{n^2} + \frac{(t+1)^2\mathbb{E}[\Phi_{t+1}]}{n^2} + \frac{(t+1)^2}{n^2} \\
&\overset{(c)}{\leq} 2\frac{(t+1)(t+2)}{n^2} \cdot \left(\frac{2\log n}{t+1} + \frac{t+1}{t+2}\right) + \frac{(t+1)(t+2)}{n^2(t+2)} + \frac{3(t+1)^2}{n^2}.
\end{aligned}$$

Lemma 3.7 implies Z is a zero mean random variable and hence (a). Equation (3.4) implies (b) and (c) follows by Lemma 3.11 and the fact that $\mathbb{E}[\Phi_{t+1}] \leq 2$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\frac{(Z-(t+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] &\leq \frac{\mathbb{E}[(Z-(t+1)/n)^2]}{(t+1)/n} \\
&\leq \frac{2(t+2)}{n} \cdot \left(\frac{2\log n}{t+1} + \frac{t+1}{t+2}\right) + \frac{1}{n} + \frac{3(t+1)}{n} \\
&= \frac{5t}{n} + \frac{4\log n(t+2)}{n(t+1)} + \frac{6}{n}.
\end{aligned}$$

□

Lemma 3.15. For any n and t if $\mathbb{E}[\Phi_{t+1}] > 2$, then

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] \leq \frac{5t}{n} + \frac{4\log n}{n}\left(\frac{t+2}{t+1}\right) + \frac{6}{n}.$$

Proof.

$$\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n} = \frac{{S_t}^2}{(\Phi_{t+1}+1)(t+1)/n} + \frac{(t+1)(\Phi_{t+1}+1)}{n} - 2S_t.$$

Thus by Lemma 3.7,

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] = \mathbb{E}\left[\frac{{S_t}^2}{(\Phi_{t+1}+1)(t+1)/n}\right] - \frac{(t+1)(\mathbb{E}[\Phi_{t+1}]-1)}{n}. \tag{3.6}$$

By Lemmas 3.13, 3.7, and Equation (3.3),

$$\begin{aligned}
\mathbb{E}\left[\frac{{S_t}^2}{(\Phi_{t+1}+1)(t+1)/n}\right] &\leq \frac{\mathbb{E}[{S_t}^2]}{\mathbb{E}[\Phi_{t+1}-1](t+1)/n} \\
&\leq \frac{t+1}{n}\frac{\mathbb{E}[\Phi_{t+1}]^2}{\mathbb{E}[\Phi_{t+1}-1]} + \frac{t+2}{n}\frac{\mathbb{E}[\Phi_{t+2}]}{\mathbb{E}[\Phi_{t+1}-1]}.
\end{aligned}$$

Substituting the above equation in Equation (3.6) and simplifying,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(S_t-(t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right]\\
&\leq \frac{(t+1)\mathbb{E}[\Phi_{t+1}]+(t+2)\mathbb{E}[\Phi_{t+2}]}{n\mathbb{E}[\Phi_{t+1}-1]}+\frac{t+1}{n}\\
&\overset{(a)}{\leq} 2\frac{(t+1)\mathbb{E}[\Phi_{t+1}]+(t+2)\mathbb{E}[\Phi_{t+2}]}{n\mathbb{E}[\Phi_{t+1}]}+\frac{t+1}{n}\\
&\overset{(b)}{\leq} 2\left(\frac{t+1}{n}+\frac{t+2}{n}\left(\frac{2\log n}{t+1}+\frac{t+1}{t+2}+\frac{1}{2(t+2)}\right)\right)+\frac{t+1}{n}\\
&=\frac{5t}{n}+\frac{4\log n}{n}\left(\frac{t+2}{t+1}\right)+\frac{6}{n}.
\end{aligned}$$

Since $\mathbb{E}[\Phi_{t+1}]\geq 2$, $\mathbb{E}[\Phi_{t+1}]-1\geq \mathbb{E}[\Phi_{t+1}]/2$ and hence (a). Lemma 3.11 implies (b). □

Combining the above two lemmas results in

Lemma 3.16. For any $t_0\geq 1$,

$$\sum_{t=0}^{t_0-1}\mathbb{E}\left[\frac{(S_t-(t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right]\leq \frac{5t_0^2}{2n}+\frac{4\log n}{n}\left(t_0+\log t_0+1\right)+\frac{7t_0}{2n}.$$

Proof. By Lemmas 3.14 and 3.15, regardless of the value of $\mathbb{E}[\Phi_{t+1}]$,

$$\mathbb{E}\left[\frac{(S_t-(t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right]\leq \frac{5t}{n}+\frac{4\log n}{n}\left(\frac{t+2}{t+1}\right)+\frac{6}{n}.$$

Summing the above expression for $0\leq t\leq t_0-1$ results in the lemma. □

Substituting the results from Lemmas 3.10 and 3.16 in Lemma 3.9,

$$\mathbb{E}[D(S||\hat{S})]\leq \frac{1}{t_0}+\frac{5t_0^2}{2n}+\frac{4\log n}{n}\left(t_0+\log t_0+1\right)+\frac{7t_0}{2n}.$$

Substituting $t_0=n^{1/3}/5^{1/3}$ results in Theorem 3.1.

$$\begin{aligned}
r^{\text{nat}}_{\text{poi}(n)}(q',\Delta_k)&\leq \max_{p\in\Delta_k}\mathbb{E}[D(S||\hat{S})]\\
&\leq \frac{2.6}{n^{1/3}}+\frac{2.4\log n(n^{1/3}+\log n+1)}{n}+\frac{2.1}{n^{2/3}}\\
&\leq \frac{3+o_n(1)}{n^{1/3}}.
\end{aligned}$$

3.5.3 Limitations of Good-Turing and empirical estimators

To understand the limitations of Good-Turing and empirical estimators, we first prove an upper bound on the estimation error of Good-Turing and empirical estimators. We then provide a simple example to see why these upper bounds are tight. Finally, we outline a proof sketch for the lower bound on the performance of Good-Turing and empirical estimators.

We now state two simple upper bounds on the estimation error of Good-Turing and empirical estimators. Proofs of variations of these lemmas are in [20]. We give simple proofs in Section 3.7.2 and 3.7.3 using Bernstein's inequality and Chernoff bound.

Lemma 3.17 (Empirical estimator)**.** For every distribution p and every $t \geq 1$,

$$|S_t - E_t| \underset{\delta}{=} \mathcal{O}\left(\Phi_t \frac{\sqrt{t+1}\log\frac{n}{\delta}}{n}\right).$$

Lemma 3.18 (Good-Turing estimator)**.** For every distribution p and every t, if $\mathbb{E}[\Phi_t] \geq 1$, then

$$|S_t - G_t| \underset{\delta}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_{t+1}]+1}\frac{(t+1)\log^2\frac{n}{\delta}}{n}\right).$$

The next sample shows the tightness of these results.

Example 3.19. Let $U[k]$ be the uniform distribution over k symbols, and let the sample size be $n \gg k$. The expected multiplicity of each symbol is $\frac{n}{k}$, and by properties of binomial distributions, the multiplicity of any symbol is $> \frac{n}{k} + \sqrt{\frac{n}{k}}$ with probability ≥ 0.1. Also, for every multiplicity t, $S_t = \Phi_t/k$.

- The empirical estimate $E_t = \Phi_t\frac{t}{n}$. For $t \geq \frac{n}{k} + \sqrt{\frac{n}{k}}$, the error is $\Phi_t\sqrt{\frac{1}{nk}} \approx \Phi_t\frac{\sqrt{t}}{n}$.
- The Good-Turing estimate $G_t = \Phi_{t+1}\frac{t+1}{n}$ and it does not depend on Φ_t. Therefore, if two sequences have same Φ_{t+1}, but different Φ_t then Good-Turing makes an error in at least one of the sequences. It can be shown that, the typical error is $\sqrt{\mathbb{E}[\Phi_t]}\frac{1}{k} \approx \sqrt{\mathbb{E}[\Phi_t]}\frac{t}{n}$, as the standard deviation of Φ_t is $\sqrt{\mathbb{E}[\Phi_t]}$.

The errors in the above example are very close to the upper bounds in Lemma 3.17 and Lemma 3.18. It can be shown that the following p achieves the lower bound in Lemma 3.2. Let p be a distribution with $\frac{\sqrt{n}}{\log^3 n}$ symbols with probability $p_i \stackrel{\text{def}}{=} \frac{n^{1/3}\log^3 n}{cn} + i\frac{n^{1/6}\log^3 n}{cn}$ for $1 \le i \le n^{1/6}$. c is chosen such that the sum of probabilities adds up to 1. It can be shown that p has the following properties.

- Let $\mathcal{R} \stackrel{\text{def}}{=} \cup_{i=1}^{n^{1/6}}[np_i + n^{1/6}, np_i + 2n^{1/6}]$. For every $t \in \mathcal{R}$, $\mathbb{E}[\Phi_t] = \widetilde{\Theta}(n^{1/3})$.
- Since the probabilities are $\widetilde{\Theta}\left(\frac{n^{1/3}}{n}\right)$, symbols occur with multiplicity $\widetilde{\Theta}(n^{1/3})$ with high probability.
- The distribution is chosen such that both empirical and Good-Turing bounds in Lemmas 3.17 and 3.18 are tight.

Hence for each $t \in \mathcal{R}$, both the Good-Turing and empirical estimators makes an error of

$$\widetilde{\Omega}\left(\frac{t\sqrt{\mathbb{E}[\Phi_t]}}{n}\right) = \widetilde{\Omega}\left(\frac{\sqrt{t}\mathbb{E}[\Phi_t]}{n}\right) = \widetilde{\Omega}\left(\frac{\sqrt{n^{1/3}}n^{1/3}}{n}\right) = \widetilde{\Omega}\left(\frac{1}{n^{1/2}}\right).$$

Number of multiplicities in the range $\mathcal{R}$ is $n^{1/6} \cdot n^{1/6} = n^{1/3}$. Adding the error over all the multiplicities yields an total error of $\widetilde{\Omega}\left(\frac{1}{n^{1/2}}\right) \cdot n^{1/3} = \widetilde{\Omega}\left(\frac{1}{n^{1/6}}\right)$. Adding over all the multiplicities yields the desired result. Adding this error over all multiplicities yields the result.

3.6 Analysis outline for the improved estimator

3.6.1 A genie-aided estimator

To motivate the proposed estimator we first describe an intermediate genie-aided estimator. In the next section, we remove the genie assumption. Although by Lemma 3.7 Good-Turing estimator is unbiased, it has a large variance. It does not use the fact that Φ_t symbols appear t times, as illustrated in Example 3.19.

To overcome these limitations, imagine for a short while that a genie gives us the values of $\mathbb{E}[\Phi_t]$ for all t. We can then define the *genie-aided* estimator,

$$\widehat{S}_t = \Phi_t \frac{t+1}{n} \frac{\mathbb{E}[\Phi_{t+1}]}{\mathbb{E}[\Phi_t]}.$$

We observe few properties of $\widehat{S}_t$. By Lemma 3.7

$$\mathbb{E}[\widehat{S}_t] = \mathbb{E}[G_t] = \mathbb{E}[S_t],$$

and hence $\widehat{S}_t$ is an unbiased estimator of S_t. It is linear in Φ_t and hence shields against the variance of Φ_{t+1}. For a uniform distribution with support size k, it is easy to see that $\widehat{S}_t = \Phi_t \frac{1}{k} = S_t$. For a general distribution, we quantify the error of this estimator in the next lemma, whose proof is given in Section 3.7.4.

Lemma 3.20 (Genie-aided estimator)**.** For every distribution p and every $t \geq 1$, if $\mathbb{E}(\Phi_t) \geq 1$, then

$$\left| S_t - \Phi_t \frac{t+1}{n} \frac{\mathbb{E}[\Phi_{t+1}]}{\mathbb{E}[\Phi_t]} \right| \underset{\delta}{=} \mathcal{O}\left(\frac{\sqrt{\mathbb{E}[\Phi_t] t} \log^2 \frac{n}{\delta}}{n} \right).$$

Recall that the error of E_t and G_t are $\widetilde{\mathcal{O}}\left(\frac{\sqrt{t}\Phi_t}{n}\right)$ and $\widetilde{\mathcal{O}}\left(\frac{\sqrt{\mathbb{E}[\Phi_{t+1}]t}}{n}\right)$, respectively. In Section A we show that $\mathbb{E}[\Phi_{t+1}] = \widetilde{\mathcal{O}}(\mathbb{E}[\Phi_t])$. Hence errors of both Good-Turing and empirical estimators are linear in one of t and Φ_t and sub-linear in the other. By comparison, the genie-aided estimator achieves the smaller exponent of both estimators, and has smaller error than both. It is advantageous to use such an estimator when both t and Φ_t are $\geq \text{polylog}(n/\delta)$. In the next section, we replace the genie assumption by a good estimate of $\frac{\mathbb{E}[\Phi_{t+1}]}{\mathbb{E}[\Phi_t]}$.

3.6.2 Estimating the ratio of expected values

We now develop estimator for the ratio $\frac{\mathbb{E}[\Phi_{t+1}]}{\mathbb{E}[\Phi_t]}$ from the observed sequence. Let $\widehat{\mathbb{E}[\Phi_{t+1}]}$, $\widehat{\mathbb{E}[\Phi_t]}$ be the estimates of $\mathbb{E}[\Phi_{t+1}]$ and $\mathbb{E}[\Phi_t]$ respectively. A natural choice for the estimator $\widehat{\mathbb{E}[\Phi_t]}$ is a linear estimator of the form $\sum_t h_t \Phi_t$. One can use tools from approximation theory such as Bernstein polynomials [42] to find such a linear approximation. However a naive application of these tools is not sufficient, and instead, we exploit properties of Poisson functionals.

If we can approximate $\mathbb{E}[\Phi_t]$ and $\mathbb{E}[\Phi_{t+1}]$ to a multiplicative factor of $1 \pm \delta_1$ and $1 \pm \delta_2$, respectively, then a naive combination of the two yields an approximation of the ratio to a multiplicative factor of $1 \pm (|\delta_1| + |\delta_2|)$. However, as is evident from the proofs in Section 3.7.6, if we choose different estimators for the numerator and the denominator, we can estimate the ratio accurately. Therefore, the estimates of $\mathbb{E}[\Phi_t]$, while calculating S_t and S_{t-1}, are different. For ease of notation we use $\widehat{\mathbb{E}[\Phi_t]}$ for both the cases. The usage becomes clear from the context.

We estimate $\mathbb{E}[\Phi_{t_0}]$ as a linear combination $\sum_{i=-r}^{r} \gamma_r(i)\Phi_{t_0+i}$ of the $2r+1$ nearest Φ_t's. The coefficients $\gamma_r(i)$ are chosen to minimize to estimator's variance and bias. We show that if $\max_i |\gamma_r(i)|$ is small, then the variance is small, and that for a low bias the coefficients $\gamma_r(i)$ need to be symmetric, namely $\gamma_r(-i) = \gamma_r(i)$, and the following function should be small when $x \sim 1$,

$$B_r(x) \stackrel{\text{def}}{=} \gamma_r(0) + \sum_{i=1}^{r} \gamma_r(i) \left(x^i + x^{-i}\right) - 1.$$

To satisfy these requirements, we choose the coefficients according to the polynomial

$$\gamma_r(i) = \frac{r^2 - r\alpha_r|i| - \beta_r i^2}{r^2 + 2\sum_{j=1}^{r}(r^2 - r\alpha_r|j| - \beta_r j^2)},$$

where α_r and β_r are chosen so that $\sum_{i=1}^{r} \gamma_r(i) i^2 = 0$ and $\gamma_r(r) = 0$.

The next lemma bounds $B_r(x)$ for the estimator with co-efficients γ_r and is used to prove that the bias of the proposed estimator is small. It is proved in Section 3.7.5.

Lemma 3.21. If $r|(x-1)| \leq \min(1, x)$, then

$$|B_r(x)| = \mathcal{O}(r(x-1))^4.$$

The estimators for $\mathbb{E}[\Phi_{t_0}]$ and $\mathbb{E}[\Phi_{t_0+1}]$ are as follows. Let $r_{t_0} = \left\lfloor \frac{\sqrt{t_0}}{\log n (\Phi_{t_0}\sqrt{t_0})^{1/11}} \right\rfloor$. Let $\mathcal{S}_r^{t_0} = \{t \mid |t - t_0| \leq r\}$. Then,

$$\widehat{\mathbb{E}[\Phi_{t_0+1}]} = \sum_{t \in \mathcal{S}_{r_{t_0}}^{t_0+1}} \gamma_{r_{t_0}}(|t_0 + 1 - t|) \frac{t_0 a_t^{t_0}}{t_0 + 1} \Phi_t,$$

$$\widehat{\mathbb{E}[\Phi_{t_0}]} = \sum_{t \in \mathcal{S}_{r_{t_0}}^{t_0}} \gamma_{r_{t_0}}(|t_0 - t|) a_t^{t_0} \Phi_t.$$

where, $a_t^{t_0} = \frac{t!}{t_0!} t_0^{t_0-t}$ and is used for simplifying the analysis. Note that $\widehat{\mathbb{E}[\Phi_t]}$ used to calculate S_t and S_{t-1} are different. r_{t_0} is chosen to minimize the bias variance trade-off. The following lemma quantifies the quality of approximation of the ratio of $\mathbb{E}[\Phi_{t_0+1}]$ and $\mathbb{E}[\Phi_{t_0}]$. The proof is involved and uses Lemma 3.21. It is given in Section 3.7.6.

Lemma 3.22. For every distribution p, if $t_0 \geq \log^2 n$ and $\frac{1}{\log n}\left(\frac{t_0}{\log^2 n}\right)^5 \geq E[\Phi_{t_0}] \geq \log^2 \frac{n}{\delta}$, then

$$\left| \frac{\widehat{\mathbb{E}[\Phi_{t_0+1}]}}{\widehat{\mathbb{E}[\Phi_{t_0}]}} - \frac{\mathbb{E}[\Phi_{t_0+1}]}{\mathbb{E}[\Phi_{t_0}]} \right| \underset{\delta}{=} \mathcal{O}\left(\frac{\log^2 \frac{n}{\delta}}{\sqrt{t_0}(\mathbb{E}[\Phi_{t_0}]\sqrt{t_0})^{4/11}} \right),$$

and if $E[\Phi_{t_0}] > \frac{1}{\log n}\left(\frac{t_0}{\log^2 n}\right)^5$ then,

$$\left| \frac{\widehat{\mathbb{E}[\Phi_{t_0+1}]}}{\widehat{\mathbb{E}[\Phi_{t_0}]}} - \frac{\mathbb{E}[\Phi_{t_0+1}]}{\mathbb{E}[\Phi_{t_0}]} \right| \underset{\delta}{=} \mathcal{O}\left(\frac{\log^2 \frac{n}{\delta}}{\sqrt{\mathbb{E}[\Phi_{t_0}]}} \right).$$

3.6.3 Proposed estimator

Substituting the estimators for $\mathbb{E}[\Phi_t]$ and $\mathbb{E}[\Phi_{t+1}]$ in the genie-aided estimator we get the proposed estimator as

$$F_t = \Phi_t \frac{t+1}{n} \frac{\widehat{\mathbb{E}[\Phi_{t+1}]}}{\widehat{\mathbb{E}[\Phi_t]}}.$$

As mentioned before, for small values of Φ_t, empirical estimator performs well, and for small values of t Good-Turing performs well. Therefore, we propose the following (unnormalized) estimator that uses estimator F_t for t and $\Phi_t \geq \text{polylog}(n)$.

$$F_t'^{\text{un}} = \begin{cases} \max\left(G_0, \frac{1}{n}\right) & \text{if } t = 0, \\ E_t & \text{if } \Phi_t \leq \log^2 n, \\ \max\left(G_t, \frac{1}{n}\right) & \text{if } t \leq \log^2 n \text{ and } \Phi_t > \log^2 n, \\ \min\left(\max\left(F_t, \frac{1}{n^3}\right), 1\right) & \text{otherwise.} \end{cases}$$

Letting $N \overset{\text{def}}{=} \sum_{t=0}^{n} F_t'^{\text{un}}$, the normalized estimator is then $F_t' \overset{\text{def}}{=} \frac{1}{N} F_t'^{\text{un}}$. Note that the Good-Turing and F_t may assign 0 probability to S_t even though $\Phi_t \neq 0$.

To avoid infinite log loss and KL Divergence between the distribution and the estimate, both estimators are slightly modified by taking $\max\left(G_t, \frac{1}{n}\right)$ instead of G_t and $\min\left(\max\left(F_t, \frac{1}{n^3}\right), 1\right)$ instead of F_t so as not to assign 0 or ∞ probability mass to any multiplicity. Such modifications are common in prediction and compression, *e.g.,* [43].

3.6.4 Proof sketch of Theorem 3.3

To prove Theorem 3.3, we will analyze the unnormalized estimator $F_t'^{\text{un}}$ and prove that $|N-1| \underset{10n^{-2}}{=} \widetilde{\mathcal{O}}(n^{-1/4})$ and use that to prove the desired result for the normalized estimator F_t'. We first show that the estimation error for every multiplicity is small. The proof is in Section 3.7.7.

Lemma 3.23. For every distribution p, $|S_0 - F_0'^{\text{un}}| \underset{4n^{-3}}{=} \mathcal{O}\left(\frac{\log^2 n}{\sqrt{n}}\right)$, and for all $t \geq 1$,

$$|S_t - F_t'^{\text{un}}| \underset{4n^{-3}}{=} \mathcal{O}\left(\frac{\min(\sqrt{\Phi_t}(t+1), \Phi_t^{7/11}\sqrt{t+1})}{n \log^{-3} n}\right).$$

The error probability in the above equation is $4n^{-3}$ can be generalized to any poly$(1/n)$. We have chosen the above error to achieve the over all error in Theorem 3.3 to be n^{-1}. Note that the error of F_t' is smaller than both Good-Turing and empirical estimators up to polylog(n) factors. Using Lemma 3.23, we show that $N \approx 1$ in the following lemma. It is proved in Section 3.7.8.

Lemma 3.24. For every distribution p,

$$|N-1| \underset{10n^{-2}}{=} \widetilde{\mathcal{O}}\left(\min\left(\frac{1}{n^{1/4}}, \frac{\sqrt{k}}{\sqrt{n}}\right)\right).$$

Using the bounds on $N-1$ in Lemma 3.24 and bounds on $|S_t - F_t'^{\text{un}}|$ in Lemma 3.23 and maximizing the KL divergence, we prove Theorem 3.3 in Section 3.7.9.

3.6.5 Lower bounds on estimation

We now lower bound the rate of convergence. We construct an explicit distribution such that with probability $\geq 1 - n^{-1}$ the total variation distance is

$\widetilde{\Omega}(n^{-1/4})$. By Pinsker's inequality, this implies that the KL divergence is $\widetilde{\Omega}(n^{-1/2})$. Note that since distance is $\widetilde{\Omega}(n^{-1/4})$ with probability close to 1, the expected distance is also $\widetilde{\Omega}(n^{-1/4})$.

Let p be a distribution with $n_i \stackrel{\text{def}}{=} \sqrt{\frac{\pi}{2}} i \log^{1.5} n$ symbols with probability $p_i \stackrel{\text{def}}{=} \frac{\lfloor i^2 \log^3 n \rfloor}{n}$, and n_i symbols with probability $p_i + \frac{i}{n}$, for $c_1 \frac{n^{1/4}}{\log^{9/8} n} \leq i \leq c_2 \frac{n^{1/4}}{\log^{9/8} n}$. c_1 and c_2 are constants such that the sum of probabilities is 1. We sketch the proof here.

Proof sketch of Theorem 3.4. The distribution p has the following properties.

- Let $\mathcal{R} = \cup_i \{np_i, np_i + 1 \ldots np_i + i\}$ for $c_1 \frac{n^{1/4}}{\log^{9/8} n} \leq i \leq c_2 \frac{n^{1/4}}{\log^{9/8} n}$. For every $t \in \mathcal{R}$, $\Pr(\Phi_t = 1) \geq 1/3$.
- If $\Phi_t = 1$, then the symbol that has appeared t times has probability p_i or $p_i + \frac{i}{n}$ with almost equal probability.
- Label-invariant estimators cannot distinguish between the two cases, and hence incur an error of $\widetilde{\Omega}(i/n) = \widetilde{\Omega}(n^{-3/4})$ for a constant fraction of multiplicities $t \in \mathcal{R}$.

The total number of multiplicities in $\mathcal{R}$ is $n^{1/4} \cdot n^{1/4} = n^{1/2}$. Multiplying by the error for each multiplicity yields the bound $\widetilde{\Omega}(n^{-1/4})$. □

3.7 Proofs for the improved estimator

3.7.1 Bounds on linear estimators

In this section, we prove error bounds for linear estimators that are used to simplify other proofs in the chapter. We first show that the difference of expected values of consecutive Φ_t's is bounded.

Lemma 3.25. For every distribution p and every t,

$$|\mathbb{E}[\Phi_t] - \mathbb{E}[\Phi_{t+1}]| = \mathcal{O}\left(\mathbb{E}[\Phi_t] \max\left(\frac{\log n}{t+1}, \sqrt{\frac{\log n}{t+1}}\right)\right) + \frac{1}{n}.$$

Proof. We consider the two cases $t+1 \geq \log n$ and $t+1 < \log n$ separately. Consider the case when $t+1 \geq \log n$. We first show that

$$|\mathbb{E}[\mathbb{1}_{N_x=t}] - \mathbb{E}[\mathbb{1}_{N_x=t+1}]| = e^{-np_x}\frac{(np_x)^t}{t!}\left|1-\frac{np_x}{t+1}\right| \leq 5e^{-np_x}\frac{(np_x)^t}{t!}\sqrt{\frac{\log n}{t+1}} + \frac{2}{n^3}. \tag{3.7}$$

The first equality follows by substituting $\mathbb{E}[\mathbb{1}_{N_x=t}] = e^{-np_x}(np_x)^t/t!$. For the inequality, note that if $|np_x - t - 1|^2 \leq 25(t+1)\log n$, then the inequality follows. If not, then by the Chernoff bound

$$\mathbb{E}[\mathbb{1}_{N_x=t}] = \Pr(t_x = t) \leq n^{-3}$$

and hence

$$|\mathbb{E}[\mathbb{1}_{N_x=t}] - \mathbb{E}[\mathbb{1}_{N_x=t+1}]| \leq \mathbb{E}[\mathbb{1}_{N_x=t}] + \mathbb{E}[\mathbb{1}_{N_x=t+1}] \leq 2/n^3.$$

By definition,

$$\mathbb{E}[\Phi_t] - \mathbb{E}[\Phi_{t+1}] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=t}] - \mathbb{E}[\mathbb{1}_{N_x=t+1}].$$

Substituting,

$$\begin{aligned}
|\mathbb{E}[\Phi_t] - \mathbb{E}[\Phi_{t+1}]| &\leq \sum_x |\mathbb{E}[\mathbb{1}_{N_x=t}] - \mathbb{E}[\mathbb{1}_{N_x=t+1}]| \\
&\overset{(a)}{=} \sum_x e^{-np_x}\frac{(np_x)^t}{t!}\left|1-\frac{np_x}{t+1}\right| \\
&= \sum_{x:np_x\leq 1} e^{-np_x}\frac{(np_x)^t}{t!}\left|1-\frac{np_x}{t+1}\right| + \sum_{x:np_x>1} e^{-np_x}\frac{(np_x)^t}{t!}\left|1-\frac{np_x}{t+1}\right| \\
&\overset{(b)}{\leq} \sum_{x:np_x\leq 1}\frac{np_x}{t!} + \sum_{x:np_x>1} 5e^{-np_x}\frac{(np_x)^t}{t!}\sqrt{\frac{\log n}{t+1}} + \frac{2}{n^3} \\
&\leq \frac{1}{n^2} + \mathcal{O}\left(\mathbb{E}[\Phi_t]\sqrt{\frac{\log n}{t+1}}\right) + \frac{2n}{n^3} \leq \mathcal{O}\left(\mathbb{E}[\Phi_t]\sqrt{\frac{\log n}{t+1}}\right) + \frac{1}{n}.
\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}[\mathbb{1}_{N_x=t}] = e^{-np_x}(np_x)^t/t!$. (b) follows from the fact that $np_x \leq 1$ in the first summation and Equation (3.7). The proof for the case $t+1 < \log n$ is similar and hence omitted. □

Next we prove a concentration inequality for any linear estimator f.

Lemma 3.26. Let $r \le \sqrt{\frac{t_0}{\log n}}$, $t_0 \ge \log n$, and $f = \sum_{t \in \mathcal{S}_r^{t_0}} c_t \Phi_t$. For every distribution p if $\mathbb{E}[\Phi_{t_0}] \ge \log \frac{1}{\delta}$, then

$$|f - \mathbb{E}[f]| \underset{\delta}{=} \mathcal{O}\left(\max_{t \in \mathcal{S}_r^{t_0}} |c_t| \sqrt{\mathbb{E}[\Phi_{t_0}](2r+1)\log\frac{1}{\delta}}\right).$$

Proof. By Lemma 3.8,

$$\begin{aligned}
\mathrm{Var}(f) &\le \sum_{t \in \mathcal{S}_r^{t_0}} \sum_x c_t^2 \mathbb{E}[\mathbb{1}_{N_x = t}] \\
&\le \left(\max_{t \in \mathcal{S}_r^{t_0}} c_t\right)^2 \sum_{t \in \mathcal{S}_r^{t_0}} \sum_x \mathbb{E}[\mathbb{1}_{N_x = t}] \\
&\overset{(a)}{=} \left(\max_{t \in \mathcal{S}_r^{t_0}} c_t\right)^2 \sum_{t \in \mathcal{S}_r^{t_0}} \mathbb{E}[\Phi_t] \\
&= \mathcal{O}\left(\left(\max_{t \in \mathcal{S}_r^{t_0}} c_t\right)^2 (2r+1)\mathbb{E}[\Phi_{t_0}]\right).
\end{aligned}$$

Substituting $\sum_x \mathbb{E}[\mathbb{1}_{N_x=t}] = \mathbb{E}[\Phi_t]$ results in (a). The last equality follows by repeatedly applying Lemma 3.25. Changing one of the multiplicities changes f by at-most $\max_{t \in \mathcal{S}_r^{t_0}} |c_t|$. Applying Bernstein's inequality with the above calculated bounds on variance, $M = \max_{t \in \mathcal{S}_r^{t_0}} |c_t|$, and $\sum_i \epsilon_i = 0$ yields the lemma. □

Next we prove a concentration bound for Φ_t in the next lemma.

Lemma 3.27. For every distribution p and every multiplicity t, if $\mathbb{E}[\Phi_t] \ge \log \frac{1}{\delta}$, then

$$|\Phi_t - \mathbb{E}[\Phi_t]| \underset{\delta}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_t]\log\frac{1}{\delta}}\right).$$

Proof. Since $\Phi_t = \sum_x \mathbb{1}_{N_x=t}$, by Lemma 3.8, $\mathrm{Var}(\Phi_t) \le \mathbb{E}[\Phi_t]$. Furthermore $|\mathbb{1}_{N_x=t} - \mathbb{E}(\mathbb{1}_{N_x=t})| \le 1$. Applying Bernstein's inequality with $M = 1$, $\mathrm{Var}(\Phi_t) \le \mathbb{E}[\Phi_t]$, and $\sum_i \epsilon_i = 0$ proves the lemma. □

3.7.2 Proof of Lemma 3.17

Let $\epsilon = \frac{20\sqrt{t+1}\log\frac{n}{\delta}}{n}$. Since $\varphi_t = \sum_x \mathbb{1}_{N_x=t}$ and $S_t = \sum_x p_x \mathbb{1}_{N_x=t}$,

$$\Pr\left(\left|S_t - \Phi_t \frac{t}{n}\right| \ge \Phi_t \epsilon\right) \le \Pr\left(\exists\, x \text{ s.t. } \left|p_x - \frac{t}{n}\right| > \epsilon, \mathbb{1}_{N_x=t} = 1\right).$$

If $p_x \geq \frac{t}{n} + \epsilon$, then by the Chernoff bound $\Pr(\mathbb{1}_{N_x=t} = 1) \leq \delta/2n$. Therefore by the union bound,

$$\Pr\left(\exists\ x \text{ s.t. } \quad p_x - \frac{t}{n} > \epsilon, \mathbb{1}_{N_x=t} = 1\right) \leq n\frac{\delta}{2n} \leq \frac{\delta}{2}.$$

Now consider the set of symbols such that $p_x \leq \frac{t}{n} - \epsilon$. Since $p_x \geq 0$, we have $t \geq 20\sqrt{t+1}\log\frac{n}{\delta}$. Group symbols x with probability $\leq 1/4n$ in to smallest number of groups such that $\Pr(g) \leq 1/n$ for each group g. By Poisson sampling, for each group g, $N_g = \sum_{x\in g} N_x$ and N_g is a Poisson random variable with mean $\Pr(g)$. Observe that for any two (or more) symbols x and x',

$$\Pr(N_x \geq t \vee N_{x'} \geq t) \leq \Pr(N_x + N_{x'} \geq t).$$

Therefore

$$\begin{aligned}
&\Pr\left(\exists\ x \text{ s.t. } \ \frac{t}{n} - p_x > \epsilon, \mathbb{1}_{N_x=t} = 1\right) \\
&\leq \Pr\left(\exists\ x \text{ s.t. } \ N_x \geq t, p_x \leq \frac{t}{n} - \epsilon\right) \\
&\leq \Pr\left(\exists\ g \text{ s.t. } \ N_g \geq t \vee \exists x \text{ s.t. } \quad N_x \geq t, \frac{1}{4n} \leq p_x \leq \frac{t}{n} - \epsilon\right).
\end{aligned}$$

It is easy to see that the number of groups and the number of symbols with probabilities $\geq 1/4n$ is at most $n+1+4n \leq 6n$. Therefore by the union bound and the Chernoff bound the above probability is $\leq \delta/2$. Adding the error probabilities for cases $p_x \geq \frac{t}{n} + \epsilon$ and $p_x \leq \frac{t}{n} - \epsilon$ results in the lemma. ■

3.7.3 Proof of Lemma 3.18

By Lemma 3.7, $\mathbb{E}\left[S_t - \Phi_{t+1}\frac{t+1}{n}\right] = 0$. Recall that $S_t = \sum_x p_x \mathbb{1}_{N_x=t}$ and $\Phi_{t+1} = \sum_x \mathbb{1}_{N_x=t+1}$. Hence by Lemma 3.8 (stated and proved in Section 3.7),

$$\begin{aligned}
\mathrm{Var}\left(S_t - \Phi_{t+1}\frac{t+1}{n}\right) &\leq \sum_x \mathbb{E}[\mathbb{1}_{N_x=t}]p_x^2 + \mathbb{E}[\mathbb{1}_{N_x=t+1}]\frac{(t+1)^2}{n^2} \\
&\overset{(a)}{=} \sum_x \mathbb{E}[\mathbb{1}_{N_x=t+2}]\frac{(t+1(t+2)}{n^2} + \mathbb{E}[\mathbb{1}_{N_x=t+1}]\frac{(t+1)^2}{n^2} \\
&\overset{(b)}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_{t+1}]+1)(t+1)^2\log n}{n^2}\right).
\end{aligned}$$

$\mathbb{E}[\mathbb{1}_{N_x=t}] = e^{-np_x}(np_x)^t/t!$ and $\mathbb{E}[\mathbb{1}_x^{t+2}] = e^{-np_x}(np_x)^{t+2}/t+2!$, and hence (a). (b) follows from Lemma 3.25 (stated and proved in Section 3.7) and the fact that $\sum_x \mathbb{E}[\mathbb{1}_{N_x=t+2}] = \mathbb{E}[\Phi_{t+2}]$. By the proof of Lemma 3.17,

$$\Pr\left(\exists\, x \text{ s.t. } \left|p_x - \frac{t}{n}\right| > \frac{20\sqrt{t+1}\log\frac{n}{\delta'}}{n}, \mathbb{1}_{N_x=t} = 1\right) \leq \delta'.$$

Choosing $\delta' = \delta/2$ we get $\forall x$,

$$\left|\mathbb{1}_{N_x=t}p_x - \mathbb{1}_{N_x=t+1}\frac{t+1}{n}\right| = \mathcal{O}\left(\frac{\sqrt{t+1}\log\frac{n}{\delta}}{n} + \frac{t+1}{n}\right)$$

with probability $1-\delta/2$. The lemma follows from Bernstein's inequality with $M = \mathcal{O}\left(\frac{\sqrt{t+1}\log\frac{n}{\delta}}{n} + \frac{t+1}{n}\right)$, $\sum_i \epsilon_i = \delta/2$, and above calculated bound on the variance. ■

3.7.4 Proof of Lemma 3.20

By Lemma 3.7,

$$\mathbb{E}[S_t] - \mathbb{E}[\Phi_t]\frac{t+1}{n}\frac{\mathbb{E}[\Phi_{t+1}]}{\mathbb{E}[\Phi_t]} = 0.$$

We now bound the variance. By definition, $S_t = \sum_x p_x \mathbb{1}_{N_x=t}$ and $\Phi_{t+1} = \sum_x \mathbb{1}_{N_x=t+1}$. Using Lemma 3.8,

$$\begin{aligned}
&\mathrm{Var}\left(S_t - \frac{(t+1)\Phi_t}{n}\frac{\mathbb{E}[\Phi_{t+1}]}{\mathbb{E}[\Phi_t]}\right)\\
&\leq \sum_x \mathbb{E}[\mathbb{1}_{N_x=t}]\left(p_x - \frac{(t+1)\mathbb{E}[\Phi_{t+1}]}{n\mathbb{E}[\Phi_t]}\right)^2\\
&= \sum_x \mathbb{E}[\mathbb{1}_{N_x=t}]\left(p_x - \frac{t+1}{n} + \frac{(t+1)(\mathbb{E}[\Phi_t] - \mathbb{E}[\Phi_{t+1}])}{n\mathbb{E}[\Phi_t]}\right)^2\\
&\overset{(a)}{\leq} \sum_x 2\mathbb{E}[\mathbb{1}_{N_x=t}]\left(p_x - \frac{t+1}{n}\right)^2 + 2\mathbb{E}[\mathbb{1}_{N_x=t}]\left(\frac{(\mathbb{E}[(\Phi_{t+1}] - \mathbb{E}[\Phi_t])(t+1)}{n\mathbb{E}[\Phi_t]}\right)^2\\
&\overset{(b)}{=} \mathcal{O}\left(\frac{\mathbb{E}[\Phi_t]t\log^2 n}{n^2}\right),
\end{aligned}$$

where (a) follows from the fact that $(x+y)^2 \leq 2x^2 + 2y^2$. Similar to the proof of Lemma 3.25, one can show that the first term in (a) is $\mathcal{O}\left(\frac{\mathbb{E}[\Phi_t]t\log^2 n}{n^2}\right)$. The second term can be bounded by $\mathcal{O}\left(\frac{\mathbb{E}[\Phi_t]t\log^2 n}{n^2}\right)$ using Lemma 3.25, hence (b). We

now bound the maximum value of each individual term in the summation. By the proof of Lemma 3.17,

$$\Pr\left(\exists\, x \text{ s.t. } \left|p_x - \frac{t}{n}\right| > \frac{c\sqrt{t+1}\log\frac{n}{\delta'}}{n}, \mathbb{1}_{N_x=t} = 1\right) \leq \delta' \tag{3.8}$$

Choosing $\delta' = \delta/2$ we get that with probability $1 - \delta/2$, $\forall x$

$$\begin{aligned}
\mathbb{1}_{N_x=t}\left|p_x - \frac{(t+1)\mathbb{E}[\Phi_{t+1}]}{n\mathbb{E}[\Phi_t]}\right| &\leq \mathbb{1}_{N_x=t}\left|p_x - \frac{t+1}{n}\right| + \left|\frac{(t+1)\mathbb{E}[\Phi_{t+1}] - \mathbb{E}[\Phi_t]}{n\mathbb{E}[\Phi_t]}\right| \\
&\overset{(a)}{=} \mathcal{O}\left(\frac{\sqrt{t+1}\log\frac{n}{\delta}}{n} + \frac{(t+1)\log n}{n}\right) \\
&= \mathcal{O}\left(\frac{(t+1)\log\frac{n}{\delta}}{n}\right).
\end{aligned}$$

where the (a) follows from Lemma 3.25 and Equation (3.8). The lemma follows from Bernstein's inequality with the calculated variance, $M = \mathcal{O}\left(\frac{(t+1)\log\frac{n}{\delta}}{n}\right)$, and $\sum_i \epsilon_i = \delta/2$. ■

3.7.5 Proof of Lemma 3.21

By assumption, $|r(x-1)| \leq \min(1, x)$. Hence $|r\ln x| < 2|r(x-1)|$ and $|r\ln x| \leq 1$. Therefore

$$\begin{aligned}
|B_r(x)| &= \left|1 - \gamma_r(0) - \sum_{i=1}^{r}\gamma_r(i)2\cosh(i\ln x)\right| \\
&= \left|1 - \gamma_r(0) - 2\sum_{i=1}^{r}\gamma_r(i)\Big(1 + \frac{(i\ln x)^2}{2!} + \frac{(i\ln x)^4}{4!} + \frac{(i\ln x)^6}{6!} + \cdots\Big)\right| \\
&\overset{(a)}{=} \left|2\sum_{i=1}^{r}\gamma_r(i)\Big(\frac{(i\ln x)^4}{4!} + \frac{(i\ln x)^6}{6!} + \cdots\Big)\right| \\
&\overset{(b)}{\leq} 2\sum_{i=1}^{r}\left|\gamma_r(i)\right|2\frac{(i\ln x)^4}{4!},
\end{aligned}$$

where in (a) we use that $\gamma_r(0) + 2\sum_{i=1}^{r}\gamma_r(i) = 1$ and $\sum_{i=1}^{r}\gamma_r(i)i^2 = 0$. (b) follows from the fact that $|r\ln x| \leq 1$. Now using $r|\ln(x)| \leq 2r|x-1|$, and $|\gamma_r(i)| = \mathcal{O}\left(\frac{1}{r+1}\right)$ (can be shown), the result follows. ■

3.7.6 Proof of Lemma 3.22

The proof is technically involved and we prove it in steps. We first observe the following property of $a_t^{t_0}$. The proof follows from the definition.

Lemma 3.28. For every distribution p and multiplicities t, t_0,

$$a_t^{t_0}\mathbb{E}[\mathbb{1}_{N_x=t}] = \mathbb{E}[\mathbb{1}_{N_x=t_0}]\left(\frac{np_x}{t_0}\right)^{t-t_0}.$$

Next we bound $\widehat{\mathbb{E}[\Phi_t]}-\mathbb{E}[\Phi_t]$. The proposed estimators for $\mathbb{E}[\Phi_t]$ and $\mathbb{E}[\Phi_{t+1}]$ have positive bias. Hence we analyze $\widehat{\mathbb{E}[\Phi_t]} - \widehat{\mathbb{E}[\Phi_{t+1}]}$ to prove tighter bounds for the ratio.

Lemma 3.29. Let $r \le \frac{\sqrt{t_0}}{\log n}$ and $t_0 \ge \log n$. For every distribution p, if $\mathbb{E}[\Phi_{t_0}] \ge \log\frac{1}{\delta}$, then

$$\left|\widehat{\mathbb{E}[\Phi_{t_0}]} - \mathbb{E}[\Phi_{t_0}]\right| \underset{\delta}{=} \mathcal{O}\left(\frac{r^4\log^2 n\mathbb{E}[\Phi_{t_0}]}{t_0^2} + \sqrt{\frac{\mathbb{E}[\Phi_{t_0}]\log\frac{1}{\delta}}{r+1}}\right),$$

and

$$\left|\widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]} - \mathbb{E}[\Phi_{t_0} - \Phi_{t_0+1}]\right| \underset{\delta}{=} \mathcal{O}\left(\frac{r^4\mathbb{E}[\Phi_{t_0}]\log^{2.5} n}{t_0^{2.5}} + \frac{\sqrt{\mathbb{E}[\Phi_{t_0}]\log\frac{1}{\delta}}}{(r+1)^{1.5}}\right).$$

Proof. By triangle inequality,

$$\left|\widehat{\mathbb{E}[\Phi_{t_0}]} - \mathbb{E}[\Phi_{t_0}]\right| \le \left|\widehat{\mathbb{E}[\Phi_{t_0}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]}]\right| + \left|\mathbb{E}[\Phi_{t_0}] - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]}]\right|.$$

We first bound $|\widehat{\mathbb{E}[\Phi_{t_0}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]}]|$.

Since $r \le \sqrt{t_0}$ it can show that $a_t^{t_0} \le e$ and $|\gamma_r(|t-t_0|)| = \mathcal{O}((r+1)^{-1})$. Therefore each coefficient in $\widehat{\mathbb{E}[\Phi_{t_0}]}$ is $\mathcal{O}((r+1)^{-1})$. Hence by Lemma 3.26 (stated and proved in Section 3.7),

$$\left|\widehat{\mathbb{E}[\Phi_{t_0}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]}]\right| \underset{\delta}{=} \mathcal{O}\left(\sqrt{\frac{\mathbb{E}[\Phi_{t_0}]\log\frac{1}{\delta}}{r+1}}\right).$$

Next we bound the bias, *i.e.*, $\left|\mathbb{E}[\Phi_{t_0}] - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]}]\right|$. Recall that $a_t^{t_0}\mathbb{E}[\mathbb{1}_{N_x=t}] = \mathbb{E}[\mathbb{1}_{N_x=t_0}]\left(\frac{np_x}{t_0}\right)^{t-t_0}$. Therefore by the linearity of expectation and the definition of $B_r(x)$,

$$\mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]}] - \mathbb{E}[\Phi_{t_0}] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=t_0}]B_r\left(\frac{np_x}{t_0}\right).$$

For $r = 0$, the bias is 0. For $r \geq 1$, by the Chernoff bound and the grouping argument similar to that in the proof of empirical estimator 3.17, it can be shown that there is a constant c such that if $|np_x - t_0| \geq c\sqrt{t_0 \log n}$, then

$$\sum_{x\in\mathcal{X}} \mathbb{E}[\mathbb{1}_{N_x=t_0}]B_r\left(\frac{np_x}{t_0}\right) \leq \frac{1}{n^3}.$$

If not, then by Lemma 3.21,

$$B_r\left(\frac{np_x}{t_0}\right) = \mathcal{O}\left(\frac{r^4\log^2 n}{t_0^2}\right).$$

Bounding $\mathbb{E}[\mathbb{1}_{N_x=t_0}]B_r\left(\frac{np_x}{t_0}\right)$ for each alphabet x and using the fact that $\mathbb{E}[\Phi_{t_0}] \geq \log\frac{1}{\delta}$, we get

$$\left| E[\widehat{\mathbb{E}[\Phi_{t_0}]}] - \mathbb{E}[\Phi_{t_0}]\right| = \sum_x \mathbb{E}[\mathbb{1}_{N_x=t_0}]\mathcal{O}\left(\frac{r^4\log^2 n}{t_0^2}\right) + \frac{1}{n^3} = \mathcal{O}\left(\mathbb{E}[\Phi_{t_0}]\frac{r^4\log^2 n}{t_0^2}\right).$$

The first part of the lemma follows by the union bound. The proof of the second part is similar. We will prove the concentration of $\widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]}$ and then quantify the bias. We first bound the coefficients in $\widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]}$. The coefficient of Φ_t is bounded by

$$a_t^{t_0}\frac{|\gamma_r(|t_0+1-t|)|}{t_0+1} + a_t^{t_0}\left|\gamma_r(|t_0+1-t|) - \gamma_r(|t_0-t|)\right| = \mathcal{O}\left(\frac{1}{(r+1)^2}\right).$$

Applying Lemma 3.26, we get

$$\left|\widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]}]\right| \underset{\delta}{=} \mathcal{O}\left(\frac{\sqrt{\mathbb{E}[\Phi_{t_0}]\log\frac{1}{\delta}}}{(r+1)^{1.5}}\right).$$

Next we bound the bias.

$$\mathbb{E}[\widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]}] - \mathbb{E}[\Phi_{t_0} - \Phi_{t_0+1}] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=t_0}]\left(1 - \frac{np_x}{t_0+1}\right)B_r\left(\frac{np_x}{t_0}\right).$$

As before, bounding $\mathbb{E}[\mathbb{1}_{N_x=t_0}]\left(1-\frac{np_x}{t_0+1}\right)B_r\left(\frac{np_x}{t_0}\right)$ for each x yields the lemma. □

Now we have all the tools to prove Lemma 3.22.

Proof of Lemma 3.22. If $|\Delta b| \leq 0.9b$, then

$$|\frac{a+\Delta a}{b+\Delta b} - \frac{a}{b}| \leq \frac{\mathcal{O}(\Delta b)a}{b^2} + \frac{\mathcal{O}(\Delta a)}{b}.$$

Let $b = \mathbb{E}[\Phi_{t_0}]$, $a = \mathbb{E}[\Phi_{t_0+1} - \Phi_{t_0}]$, $\Delta b = \widehat{\mathbb{E}[\Phi_{t_0+1}]} - \mathbb{E}[\Phi_{t_0}]$ and $\Delta a = \widehat{\mathbb{E}[\Phi_{t_0}]} - \widehat{\mathbb{E}[\Phi_{t_0+1}]} - \mathbb{E}[\Phi_{t_0} - \Phi_{t_0+1}]$. By Lemma 3.29, if $\mathbb{E}[\Phi_{t_0}] \geq \log^2 \frac{n}{\delta}$ and $t_0 \geq r^2 \log^{1.5} n$, then $|\Delta b| \leq 0.9b$. Therefore by Lemma 3.29, Lemma 3.25, and the union bound,

$$\left| \frac{\widehat{\mathbb{E}[\Phi_{t_0+1}]}}{\widehat{\mathbb{E}[\Phi_{t_0}]}} - \frac{\mathbb{E}[\Phi_{t_0+1}]}{\mathbb{E}[\Phi_{t_0}]} \right| \underset{2\delta'}{=} \mathcal{O}\left(\frac{r^4 \log^{2.5} n}{t_0^{2.5}} + \frac{\log^{0.5} \frac{n}{\delta'}}{(r+1)^{1.5}\sqrt{\mathbb{E}[\Phi_{t_0}]}} \right). \tag{3.9}$$

By Lemma 3.27 (stated and proved in Section 3.7), if $\mathbb{E}[\Phi_{t_0}] \geq \log^2 \frac{n}{\delta}$, then with probability $1 - \delta/2$, $\Phi_{t_0} \in [0.5\mathbb{E}[\Phi_{t_0}], 2\mathbb{E}[\Phi_{t_0}]]$. Hence,

$$r_{t_0} \in \mathcal{R} \stackrel{\text{def}}{=} \left[\left\lfloor \frac{\sqrt{t_0}}{(2\mathbb{E}[\Phi_{t_0}]\sqrt{t_0})^{1/11} \log n} \right\rfloor, \left\lfloor \frac{\sqrt{t_0}}{(0.5\mathbb{E}[\Phi_{t_0}]\sqrt{t_0})^{1/11} \log n} \right\rfloor \right].$$

Therefore if we prove the concentration bounds for all $r \in \mathcal{R}$, the lemma would follow by the union bound. If $\max_r \mathcal{R} < 1$, then substituting $r = 0$ in Equation (3.9) yields the result for the case $\mathbb{E}[\Phi_{t_0}] \geq \frac{2}{\log n} \left(\frac{t_0}{\log^2 n} \right)^5$. If $\min_r \mathcal{R} \geq 1$, then substituting $r = \Theta\left(\frac{\sqrt{t}}{(\mathbb{E}[\Phi_{t_0}]\sqrt{t_0})^{1/11} \log n} \right)$ in Equation (3.9) yields the result for the case $\mathbb{E}[\Phi_{t_0}] \leq \frac{0.5}{\log n} \left(\frac{t_0}{\log^2 n} \right)^5$. A similar analysis proves the result for the case $1 \in \mathcal{R}$. Choosing $\delta' = \delta/2$ in Equation (3.9) and using the union bound we get the total error probability $\leq \delta$. □

3.7.7 Proof of Lemma 3.23

The proof uses the bound on the error of F_t, which is given below.

Lemma 3.30. For every distribution p and $t \geq \log^2 n$, if $\frac{1}{\log n} \left(\frac{t}{\log^2 n} \right)^5 \geq E[\Phi_t] \geq \log^2 \frac{n}{\delta}$, then

$$|S_t - F_t| \underset{2\delta}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_t]\sqrt{t})^{7/11} \log^2 \frac{n}{\delta}}{n} + \frac{\sqrt{\mathbb{E}[\Phi_t]t} \log^2 \frac{n}{\delta}}{n} \right),$$

and if $E[\Phi_t] \geq \frac{1}{\log n}\left(\frac{t}{\log^2 n}\right)^5$, then

$$|S_t - F_t| \underset{2\delta}{=} \mathcal{O}\left(\frac{t\sqrt{\mathbb{E}[\Phi_t]}\log^2\frac{n}{\delta}}{n} + \frac{\sqrt{\mathbb{E}[\Phi_t]t}\log^2\frac{n}{\delta}}{n}\right).$$

Proof. is a simple application of triangle inequality and the union bound. It follows from Lemmas 3.20 and 3.22. □

Lemma 3.23. We first show that $\mathbb{E}[\Phi_t]$ and $\mathbb{E}[\Phi_{t+1}]$ in the bounds of Lemmas 3.30 and 3.18 can be replaced by Φ_t. By Lemma 3.25, if $\mathbb{E}[\Phi_{t+1}] \geq 1$,

$$|\mathbb{E}[\Phi_t] - \mathbb{E}[\Phi_{t+1}]| = \mathcal{O}\left(\mathbb{E}[\Phi_t]\max\left(\frac{\log n}{t+1}, \sqrt{\frac{\log n}{t+1}}\right)\right) + \frac{1}{n} = \mathcal{O}\left(\mathbb{E}[\Phi_t]\log n\right).$$

Hence $\mathbb{E}[\Phi_{t+1}] = \mathcal{O}(\mathbb{E}[\Phi_t]\log n)$. Hence by Lemma 3.18, for $\mathbb{E}[\Phi_t] \geq 1$,

$$|S_t - G_t| \underset{0.5n^{-3}}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_{t+1}]+1}\frac{(t+1)\log^2 n}{n}\right) = \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_t]}\frac{(t+1)\log^3 n}{n}\right).$$

Furthermore by Lemma 3.27, if $\mathbb{E}[\Phi_t] \leq 0.5\log^2 n$, then $\Phi_t \leq \log^2 n$ with probability $\geq 1 - 0.5n^{-3}$, and we use the empirical estimator. Therefore with probability $\geq 1-0.5n^{-3}$, F_t and G_t are used only if $\mathbb{E}[\Phi_t] \geq 0.5\log^2 n$. If $\mathbb{E}[\Phi_t] \geq 0.5\log^2 n$, then by Lemma 3.27 $\mathbb{E}[\Phi_t] \underset{0.5n^{-3}}{=} \mathcal{O}(\Phi_t)$. Therefore by the union bound, if $\Phi_t \geq \log^2 n$, then

$$|S_t - G_t| \underset{n^{-3}}{=} \mathcal{O}\left(\sqrt{\Phi_t}\frac{(t+1)\log^3 n}{n}\right).$$

Similarly by Lemma 3.30, for $t \geq \log^2 n$ and $\Phi_t \geq \log^2 n$, if $\frac{1}{\log n}\left(\frac{t}{\log^2 n}\right)^5 \geq E[\Phi_t] \geq \log^2 n$, then

$$|S_t - F_t| \underset{0.5n^{-3}}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_t]\sqrt{t})^{7/11}\log^2 n}{n} + \frac{\sqrt{\mathbb{E}[\Phi_t]t}\log^2 n}{n}\right) \underset{n^{-3}}{=} \mathcal{O}\left(\frac{\Phi_t^{7/11}\sqrt{t}\log^2 n}{n}\right),$$

and if $E[\Phi_t] \geq \frac{1}{\log n}\left(\frac{t}{\log^2 n}\right)^5$, then

$$|S_t - F_t| \underset{0.5n^{-3}}{=} \mathcal{O}\left(\frac{t\sqrt{\mathbb{E}[\Phi_t]}\log^2 n}{n} + \frac{\sqrt{\mathbb{E}[\Phi_t]t}\log^2 n}{n}\right) \underset{n^{-3}}{=} \mathcal{O}\left(\frac{t\sqrt{\Phi_t}\log^2 n}{n}\right).$$

Using the above mentioned modified versions of Lemmas 3.18, 3.30 and Lemma 3.17, it can be easily shown that the lemma is true for $t \geq 1$. By Lemma 3.18,

$$|F_0'^{\text{un}} - S_0| \underset{n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{\sqrt{\Phi_1}}{n}\right).$$

By the Chernoff bound with probability $\geq 1 - e^{-n/4}$, $\Phi_1 \leq n' \leq 2n$. Hence,

$$|F_0'^{\text{un}} - S_0| \underset{4n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right).$$

Note that the error probabilities are not optimized. □

3.7.8 Proof of Lemma 3.24

By triangle inequality, $|N-1| = |\sum_t F_t'^{\text{un}} - S_t| \leq \sum_t |F_t'^{\text{un}} - S_t|$. By Lemma 3.23, for $t = 0$, $|S_0 - F_0'^{\text{un}}| \underset{4n^{-3}}{=} \widetilde{\mathcal{O}}\left(n^{-1/2}\right)$. We now use Lemma 3.23 to bound $|F_t'^{\text{un}} - S_t|$ for $t \geq 1$. Since $\sum_t t\Phi_t = n'$ is a Poisson random variable with mean n, $\Pr(\sum_t t\Phi_t \leq 2n) \geq 1 - e^{-n/4}$. For $t \geq 1$, applying Cauchy Schwarz inequality repeatedly with the above constraints we get

$$\begin{aligned}
|N-1| &\underset{10n^{-2}}{=} \sum_{t=1}^{2n} \mathcal{O}\left(\min\left(\frac{\Phi_t^{7/11}\sqrt{t}}{n}, \frac{\sqrt{\Phi_t}t}{n}\right)\text{polylog}(n)\right) \\
&= \sum_{t=1}^{2n} \widetilde{\mathcal{O}}\left(\frac{\sqrt{t}}{n}\Phi_t^{7/11}\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{2n}\frac{t\Phi_t}{n}\sum_{t=1}^{2n}\frac{\Phi_t^{3/11}}{n}}\right) \overset{(a)}{=} \widetilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{2n}\frac{\Phi_t^{1/2}}{n}}\right) \\
&= \widetilde{\mathcal{O}}\left(\min\left(\sqrt{\sqrt{\sum_{t=1}^{2n}\frac{\Phi_t t}{n}\sum_{t=1}^{2n}\frac{1}{nt}}, \frac{\sqrt{k}}{\sqrt{n}}\right)\right) \\
&= \widetilde{\mathcal{O}}\left(\min\left(\frac{1}{n^{1/4}}, \frac{\sqrt{k}}{\sqrt{n}}\right)\right).
\end{aligned}$$

Φ_t takes only integer values, hence (a). Note that by the union bound, the error probability is bounded by

$$\Pr\left(\sum_t t\Phi_t > 2n\right) + \sum_{t=0}^{2n}\Pr\left(|S_t - F_t'^{\text{un}}| \neq \widetilde{\mathcal{O}}\left(\min\left(\frac{\Phi_t^{7/11}\sqrt{t}}{n}, \frac{\sqrt{\Phi_t}t}{n}\right)\right)\right).$$

By the concentration of Poisson random variables (discussed above) the first term is $\leq e^{-n/4}$. By Lemma 3.23, the second term is $2n(4n^{-3})$. Hence the error probability is bounded by $e^{-n/4} + 2n(4n^{-3}) \leq 10n^{-2}$.

3.7.9 Proof of Theorem 3.3

It is easy to show that if $\Phi_t > \log^2 n$, with probability $\geq 1 - n^{-3}$, $\max(G_t, 1/n) = G_t$ and $\min(\max(F_t, n^{-3})1) = F_t$. For the clarity of proofs we ignore these modifications and add an additional error probability of n^{-3}.

Recall that $F'_t = \frac{F'^{\text{un}}_t}{N}$. By Jensen's inequality,

$$\sum_t S_t \log \frac{S_t}{F'_t} \leq \log \sum_t \frac{M_t^2}{F'_t}.$$

Furthermore

$$\sum_t \frac{M_t^2}{F'_t} = 1 + \frac{(S_t - F'_t)^2}{F'_t}.$$

Substituting $F'_t = F'^{\text{un}}_t/N$ and rearranging, we get

$$\sum_t \frac{(S_t - F'_t)^2}{F'_t} \leq 2(N-1)^2 + \sum_t 2N \frac{(S_t - F'^{\text{un}}_t)^2}{F'^{\text{un}}_t}.$$

By Lemma 3.24, $N = 1 + \widetilde{\mathcal{O}}(n^{-1/4})$. Therefore,

$$\sum_t \frac{(S_t - F'_t)^2}{F'_t} = \widetilde{\mathcal{O}}\left(\min\left(\frac{1}{n^{1/2}}, \frac{k}{n}\right)\right) + \sum_t \mathcal{O}\left(\frac{(S_t - F'^{\text{un}}_t)^2}{F'^{\text{un}}_t}\right).$$

To bound the second term in the above equation, we bound $|F'^{\text{un}}_t - S_t|$ and F'^{un}_t separately. We first show that $F'^{\text{un}}_t \underset{n^{-3}}{=} \widetilde{\Omega}\left(\frac{t\Phi_t}{n}\right)$.

If empirical estimator is used for estimation, then $F'^{\text{un}}_t = \Phi_t \frac{t}{n}$. If Good-Turing or F_t is used, then $\Phi_t \geq \log^2 n$. If $\mathbb{E}[\Phi_t] \leq 0.5 \log^2 n$, then $\Pr(\Phi_t \geq \log^2 n) \leq 0.5n^{-3}$. If $\mathbb{E}[\Phi_t] \geq 0.5 \log^2 n$, then using Lemma 3.25 and Lemma 3.22 it can be shown that $F'^{\text{un}}_t \underset{0.5n^{-3}}{=} \widetilde{\Omega}\left(\frac{t\Phi_t}{n}\right)$. By the union bound, $F'^{\text{un}}_t \underset{n^{-3}}{=} \widetilde{\Omega}\left(\frac{t\Phi_t}{n}\right)$.

Now using bounds on $|F'^{\text{un}}_t - S_t|$ from Lemma 3.23 and the fact that $F'^{\text{un}}_t = \widetilde{\Omega}(\Phi_t t/n)$, we bound the KL divergence. Observe that $\sum_t t\Phi_t = n'$ is a Poisson random variable with mean n, therefore

$$\Pr(\sum_t t\Phi_t \leq 2n) \geq 1 - e^{-n/4}.$$

Applying Cauchy Schwarz inequality repeatedly with the above constraint and

using bounds on $|F_t'^{\text{un}} - S_t|$ (Lemma 3.23) and F_t' we get

$$\begin{aligned}\sum_{t=1}^{2n} \frac{(S_t - F_t'^{\text{un}})^2}{F_t'^{\text{un}}} &\underset{2n(4n^{-3}+n^{-3})}{=} \sum_{t=1}^{2n} \mathcal{O}\left(\min\left(\frac{t}{n}, \frac{\Phi_t^{3/11}}{n}\right)\text{polylog}(n)\right) \\ &= \sum_{t=1}^{2n} \widetilde{\mathcal{O}}\left(\frac{\Phi_t^{1/2}}{n}\right) \\ &= \widetilde{\mathcal{O}}\left(\min\left(\sqrt{\sum_{t=1}^{2n}\frac{\Phi_t t}{n}\sum_{t=1}^{2n}\frac{1}{nt}}, \frac{k}{n}\right)\right) \\ &= \widetilde{\mathcal{O}}\left(\min\left(\frac{1}{n^{1/2}}, \frac{k}{n}\right)\right).\end{aligned}$$

For $t = 0$, by Lemma 3.18, $(S_0 - F_0'^{\text{un}})^2 \underset{n^{-3}}{=} \mathcal{O}\left(\frac{\Phi_1 \text{polylog}(n)}{n^2} + \frac{\text{polylog}(n)}{n^2}\right)$ and hence,

$$\frac{(S_0 - F_0'^{\text{un}})^2}{F_0'^{\text{un}}} \underset{n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

Similar to the proof of Lemma 3.24, by the union bound the error probability is at most

$$e^{-n/4} + 10n^{-2} + 2n(4n^{-3} + n^{-3}) + n^{-3} + n^{-3} \le 22n^{-2} \le e^{-1}n^{-1.5}$$

for $n \ge 4000$. Hence with $\text{poi}(n)$ samples, error probability is $\le e^{-1}n^{-1.5}$. Therefore by Lemma 3.6, with exactly n samples, error probability is $\le n^{-1}$.

3.8 Prediction

3.8.1 Background

Probability estimation can be naturally applied to prediction and compression. Upon observing a sequence $X^i \stackrel{\text{def}}{=} X_1, \ldots, X_i$ generated *i.i.d.* according to some distribution $p \in \mathcal{D}_{\mathcal{X}}$, we would like to form an estimate $q(x|x^i)$ of $p(x)$ to minimize a cumulative loss $\sum_{i=1}^n f_p(q(X_{i+1}|X^i), X_{i+1})$ see for example [44, 45].

The most commonly used loss is *log-loss*,

$$f_p(q(x_{i+1}|x^i), x_{i+1}) = \log(q(x_{i+1}|x^i)/p(x_{i+1})).$$

Its numerous applications include compression, *e.g.,* [45], MDL principle, *e.g.,* [46], and learning theory, *e.g.,* [47]. Its expected value is the KL-divergence between the underlying distribution p and the prediction q.

Again we consider label-invariant predictors that use only ordering and frequency of symbols, not the specific labels. Following [22], after observing n samples, we assign probability to each of the previously-observed symbols, and to observing a new symbol new. For example, if after three samples, the sequence observed is aba, we assign the probabilities $q(a|aba)$, $q(b|aba)$, and $q(\text{new}|\text{aba})$ that reflects the probability at which we think a symbol other than a or b will appear. These three probabilities must add to 1. Furthermore, if the sequence is bcb, then the probability we assign to b must be the same as the probability we previously assigned to a.

Equivalently, [22] defined the *pattern* of a sequence to be the sequence of integers, where the i^{th} new symbol appearing in the original sequence is replaced by the integer i. For example, the pattern of aba is 121. We use Ψ^n and to denote a length-n pattern, and Ψ_i to denote its ith element.

The prediction problem is now that of estimating $\Pr(\Psi_{n+1}|\Psi^n)$, where if Ψ^n consists of m distinct symbols then the distribution is over $[m+1]$, and $m+1$ reflects a new symbol. For example, after observing 121, we assign probabilities to 1, 2, and 3.

3.8.2 Previous results

[22] proved that the Good-Turing estimator achieves constant per-symbol worst-case log-loss, and constructed two sequential estimators with diminishing worst-case log-loss: a computationally efficient estimator with log-loss $\mathcal{O}(n^{-1/3})$, and a high complexity estimator with log-loss $\mathcal{O}(n^{-1/2})$. [48] constructed a low-complexity block estimator for patterns with worst-case per-symbol log-loss of $\mathcal{O}(n^{-1/2})$. For expected log-loss, [49] improved this bound to $\mathcal{O}(n^{-3/5})$ and [50] further improved it to $\widetilde{\mathcal{O}}(n^{-2/3})$, but their estimators are computationally inefficient.

3.8.3 New results

Using Theorem 3.3, we obtain a computationally efficient predictor q that achieves expected log-loss of $\widetilde{\mathcal{O}}(n^{-1/2})$. Let F'_t be the estimator proposed in Section 3.6.3. Let $q(\Psi_{n+1}|\Psi^n) = \frac{F'_t}{\Phi_t}$ if Ψ_{n+1} appears t times in Ψ^n, and F'_0, if it is Ψ_{n+1} is a new symbol. The following corollary bounds the predictor's performance.

Corollary 3.31. For every distribution p,

$$\mathbb{E}_p[D(p(\Psi_{n+1}|\Psi^n)||q(\Psi_{n+1}|\Psi^n)] = \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/2}}\right).$$

Proof. By definition $\Pr(\Psi^n) \stackrel{\text{def}}{=} \sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)$. Let ψ appear t times in Ψ^n. Using the fact that sampling is *i.i.d.*, and the definition of pattern, each of the Φ_t integers (in the pattern) are equally likely to appear as Ψ_{n+1}. This leads to,

$$P(\Psi^{n+1}, \Psi_{n+1} = \psi) = \sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)\frac{S_t(x^n)}{\Phi_t},$$

and hence

$$\Pr(\Psi_{n+1}|\Psi^n) = \frac{\sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)\frac{S_t(x^n)}{\Phi_t}}{\sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)}.$$

Any label-invariant estimator including the proposed estimator assigns identical values for F'_t to all sequences with the same pattern. Hence

$$\begin{aligned}
&\mathbb{E}\left[\sum_t S_t \log \frac{S_t}{F'_t}\right] \\
&= \sum_{x^n} p(x^n) \sum_t S_t(x^n) \log \frac{S_t(x^n)}{F'_t(x^n)} \\
&= \sum_{\Psi^n} \sum_t \sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n) S_t(x^n) \log \frac{p(x^n) S_t(x^n)}{p(x^n) F'_t(x^n)} \\
&\stackrel{(a)}{\geq} \sum_{\Psi^n} \sum_t \Big(\sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n) S_t(x^n) \Big) \log \frac{\Big(\sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n) S_t(x^n) \Big)}{\Big((\sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n)) F'_t(x^n) \Big)} \\
&= \sum_{\Psi^n} \sum_t \Big(P(\Psi^n) P(\Psi_{n+1}|\Psi^n) \Big) \log \frac{P(\Psi^{n+1})}{P(\Psi^n) F'_t} \\
&= \mathbb{E}_{\Psi^n \sim P}\left[\sum_{\Psi_{n+1}=1}^{m+1} P(\Psi_{n+1}|\Psi^n) \log \left(\frac{P(\Psi_{n+1}|\Psi^n)}{q(\Psi_{n+1}|\Psi^n)} \right) \right],
\end{aligned}$$

where in (a) we used the log-sum inequality and the fact that our estimator F'_t is identical for all sequences with the same pattern. □

Acknowledgement

Chapters 3 is adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Optimal probability estimation with applications to prediction and classification", *Conference on Learning Theory (COLT)*, 2013 [21]; and Alon Orlitsky and Ananda Theertha Suresh, "Competitive distribution estimation: Why is Good-Turing good", *Neural Information Processing Systems (NIPS)*, 2015 [37].

Chapter 4

Competitive classification

4.1 Introduction

4.1.1 Background

Classification is one of the most studied problems in machine learning and statistics [51]. Given two *training* sequences X^n and Y^n, drawn *i.i.d.* according to two distributions p and q respectively, we would like to associate a new *test* sequence Z^m drawn *i.i.d.* according to one of p and q with the training sequence that was generated by the same distribution.

It can be argued that natural classification algorithms are *label invariant*, namely, their decisions remain the same under all one-one symbol relabellings, *e.g.,* [52]. For example, if given training sequences `abb` and `cbc`, and a classifier associates `b` with `abb`, then given `utt` and `gtg`, it must associate `t` with `utt`.

Our objective is to derive a *competitive* classifier whose error is close to the best possible by any label-invariant classifier, uniformly over all (p, q). Namely, a single classifier whose error probability differs from that of the best classifier for the given (p, q) by a quantity that diminishes to 0 at a rate determined by the sample size n alone, and is independent of p and q.

4.1.2 Previous results

A number of classifiers have been studied in the past, including the likelihood-ratio, generalized-likelihood, and Chi-Square tests. However while they perform well when the number of samples is large, none of them is uniformly competitive with all label-invariant classifiers.

When $m = \Theta(n)$, classification can be related to the problem of *closeness testing* that asks whether two sequences X^n and Y^n are generated by the same or different distributions. Over the last decade, closeness testing has been considered by a number of researchers. [53] showed that testing if the distributions generating X^n and Y^n are identical or are at least δ apart in ℓ_1 distance requires $n = \widetilde{\mathcal{O}}(k^{2/3})$ samples where the constant depends on δ. [32] took a competitive view of closeness testing and derived a test whose error is $\leq \epsilon e^{\mathcal{O}(n^{2/3})}$ where ϵ is the error of the best label-invariant protocol for this problem, designed in general with knowledge of p and q.

Their result shows that if the optimal closeness test requires n samples to achieve an error $\leq \epsilon$, then the proposed test achieves the same error with $\widetilde{\mathcal{O}}(n^3)$ samples. [33] improved it to $\widetilde{\mathcal{O}}(n^{3/2})$ and proved a lower bound of $\widetilde{\Omega}(n^{7/6})$ samples.

4.1.3 New results

We consider the case where $m = 1$, namely the test data is a single sample. Many machine-learning problems are defined in this regime, for example, we are given the DNA sequences of several individuals and need to decide whether or not they are susceptible to a certain disease *e.g.,* [54].

It may seem that when $m = 1$, the best classifier is a simple majority classifier that associates Z with the sequence X^n or Y^n where Z appears more times. Perhaps surprisingly, the next example shows that this is not the case.

Example 4.1. Let $p = U[n]$ and $q = U[2n]$ be the uniform distributions over $\{1, \ldots, n\}$ and $\{1, \ldots, 2n\}$, and let the test symbol Z be generated according to $U[n]$ or $U[2n]$ with equal probability. We show that the empirical classifier, that associates Z with the sample in which it appeared more times, entails a constant

additional error more than the best achievable.

The probability that Z appears in both X^n and Y^n is a constant. And in all these cases, the optimal label-invariant test that knows p and q assigns Z to $U[n]$, namely X^n, because $p(Z) = 1/n > 1/2n = q(Z)$. However, with constant probability, Z appears more times in Y^n than in X^n, and then the empirical classifier associates Z with the wrong training sample, incurring a constant error above that of the optimal classifier.

Using probability-estimation techniques, we derive a uniformly competitive classifier. Before stating our results we formally define the quantities involved. Recall that $X^n \sim p$ and $Y^n \sim q$. A classifier S is a mapping $S : \mathcal{X}^* \times \mathcal{X}^* \times \mathcal{X} \to \{\mathtt{x}, \mathtt{y}\}$, where $S(\overline{x}, \overline{y}, z)$ indicates whether z is generated by the same distribution as $\overline{x}$ or $\overline{y}$. For simplicity we assume that $Z \sim p$ or q with equal probability, but this assumption can be easily relaxed. The error probability of a classifier S with n samples is

$$\mathcal{E}^S_{p,q}(n) = \frac{1}{2}\Pr\left(S(X^n, Y^n, Z) = \mathtt{y} | Z \sim p\right) + \frac{1}{2}\Pr\left(S(X^n, Y^n, Z) = \mathtt{x} | Z \sim q\right).$$

Let $\mathcal{S}$ be the collection of label-invariant classifiers. For every p, q, let $\mathcal{E}^{S_{p,q}}_{p,q}(n) = \min_{S\in\mathcal{S}} \mathcal{E}^S_{p,q}(n)$ be the lowest error achieved for (p, q) by any label-invariant classifier, where the classifier $S_{p,q}$ achieving $\mathcal{E}^{S_{p,q}}_{p,q}(n)$ is typically designed with prior knowledge of (p, q).

We construct a linear-time label-invariant classifier S whose error is close to $\mathcal{E}^{S_{p,q}}_{p,q}(n)$. We first extend the ideas developed in the previous section to pairs of sequences and develop an estimator $F'^p_{t,t'}$, and then use this estimator to construct a classifier whose extra error is $\widetilde{\mathcal{O}}(n^{-1/5})$.

Theorem 4.2. For all (p, q), there exists a classifier S such that

$$\mathcal{E}^S_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/5}}\right).$$

In Section 4.2 we state the classifier that has extra error $\widetilde{\mathcal{O}}(n^{-1/5})$ and prove Theorem 4.2. In Section 4.2.6 we also provide a non-tight lower bound for the problem and show that for any classifier S, there exist (p, q), such that $\mathcal{E}^S_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\Omega}\left(n^{-1/3}\right)$.

4.2 Label-invariant classification

In this section, we extend the combined-probability estimator to joint-sequences and propose a competitive classifier. First introduce profiles, a sufficient statistic for label-invariant classifiers. Then we relate the problem of classification to that of estimation in joint sequences. Motivated by the techniques in probability estimation, we then develop a joint-sequence probability estimator and prove its convergence rate, thus proving an upper bound on the error of the proposed classifier. Finally we prove a non-tight lower bound of $\widetilde{\Omega}(n^{-1/3})$.

4.2.1 Joint-profiles

Let the training sequences be X^n and Y^n and the test sequence be Z^1. It is easy to see that a sufficient statistic for label invariant classifiers is the *joint profile* φ of X^n, Y^n, Z^1, that counts how many elements appeared any given number of times in the three sequences [32]. For example, for $\overline{X} = aabcd$, $\overline{Y} = bacde$ and $Z = a$, the profiles are $\varphi(\overline{X}, \overline{Y}) = \{(2,1),(1,1),(1,1),(0,1)\}$ and $\varphi(\overline{X}, \overline{Y}, Z) = \{(2,1,1),(1,1,0),(1,1,0),(1,1,0),(0,1,0)\}$. $\varphi(\overline{X}, \overline{Y})$ indicates that there is one symbol appearing twice in first sequence and once in second, two symbols appearing once in both and so on. The profiles for three sequences can be understood similarly. Any label invariant test is only a function of the joint profile.

By definition, the probability of a profile is the sum of the probabilities of all sequences with that profile *i.e.,* for profiles of $(\overline{x}, \overline{y}, z)$,

$$\Pr(\varphi) = \sum_{\overline{x},\overline{y},z|\varphi(\overline{x},\overline{y},z)} \Pr(\overline{x}, \overline{y}, z).$$

$\Pr(\varphi)$ is difficult to compute due to the permutations involved. Various techniques to compute profile probabilities are studied in [55]. Still the proposed classifier we derive runs in linear time.

4.2.2 Classification via estimation

Let $n_x(\overline{x},\overline{y})$ denote the number of multiplicities symbol x in $(\overline{x},\overline{y})$. Let

$$S^p_{t,t'}(\overline{x},\overline{y}) \stackrel{\text{def}}{=} \sum_{x:n_x(\overline{x},\overline{y})=(t,t')} p_x$$

be the sum of the probabilities of all elements in p such that $n_x(\overline{x},\overline{y}) = (t,t')$. $S^q_{t,t'}(\overline{x},\overline{y})$ is defined similarly.

Let $\varphi = \varphi(\overline{x},\overline{y})$ be the joint profile of $(\overline{x},\overline{y})$. If z is generated according to p, then the probability of observing the joint profile $\varphi(\overline{x},\overline{y},z)$, where z is an element appearing t and t' times respectively in $\overline{x}$ and $\overline{y}$ is

$$\begin{aligned}\Pr{}^p(\varphi(\overline{x},\overline{y},z)) &= \sum_{\overline{x},\overline{y}|\varphi(\overline{x},\overline{y})=\varphi} P(\overline{x})Q(\overline{y})S^p_{t,t'}(\overline{x},\overline{y}),\\ &= \Pr(\varphi(\overline{x},\overline{y}))\mathbb{E}_\varphi[S^p_{t,t'}],\end{aligned}$$

where $\mathbb{E}_\varphi[S^p_{t,t'}] \stackrel{\text{def}}{=} \mathbb{E}[S^p_{t,t'}|\Phi=\varphi]$ is the expected value of $S^p_{t,t'}$ given that φ is the profile.

When the two distributions are known and the observed joint profile is $\varphi(\overline{x},\overline{y},z)$, then the classification problem becomes a hypothesis testing problem. The optimal solution to the hypothesis testing when both hypotheses are equally likely is the one that assigns higher probability to the observation (joint profile in our case). So the optimal classifier is

$$\begin{aligned}&\Pr{}^p(\varphi(\overline{x},\overline{y},z)) \gtrless_q \Pr{}^q(\varphi(\overline{x},\overline{y},z))\\ \Rightarrow\quad &\mathbb{E}_\varphi[S^p_{t,t'}] \gtrless_q \mathbb{E}_\varphi[S^q_{t,t'}].\end{aligned}$$

We will develop variants of F'_t for joint profiles, denoted by $F'^p_{t,t'}$, and $F'^q_{t,t'}$. We use these estimators in place of the expected values. Our classifier S assigns z to $\overline{x}$ if $F'^p_{t,t'} > F'^q_{t,t'}$ and to $\overline{y}$ if $F'^p_{t,t'} < F'^q_{t,t'}$. Ties are broken at random. There is an additional error in classification with respect to the optimal label-invariant classifier when $\mathbb{E}_\varphi[S^p_{t,t'}] < \mathbb{E}_\varphi[S^q_{t,t'}]$ but $F'^p_{t,t'} \geq F'^q_{t,t'}$ or vice versa.

Let $\mathbb{1}^\epsilon_{t,t'}$ be an indicator random variable that is 1 if

$$|\mathbb{E}_\varphi[S^p_{t,t'}] - \mathbb{E}_\varphi[S^q_{t,t'}]| \leq \sum_{s\in\{p,q\}} |F'^s_{t,t'} - \mathbb{E}_\varphi[S^s_{t,t'}]|. \tag{4.1}$$

It is easy to see that if there is an additional error, then $\mathbb{1}^{\epsilon}_{t,t'} = 1$. Using these conditions the following lemma provides a bound on the additional error with respect to the optimal.

Lemma 4.3 (Classification via estimation)**.** For every (p, q) and every classifier S,

$$\mathcal{E}^{S}_{p,q}(n) \leq \mathcal{E}^{S_{p,q}}_{p,q}(n) + \sum_{t,t'} \sum_{t \in \{p,q\}} \mathbb{E}[\mathbb{1}^{\epsilon}_{t,t'} |F'^{t}_{t,t'} - S^{t}_{t,t'}|].$$

Proof. For a joint profile φ, S assigns z to the wrong hypothesis, if $F'^{p}_{t,t'} > F'^{q}_{t,t'}$ and $\mathbb{E}_{\varphi}[S^{p}_{t,t'}] < \mathbb{E}_{\varphi}[S^{q}_{t,t'}]$ or vice versa. Hence $\mathbb{1}^{\epsilon}_{t,t'} = 1$. If $\mathbb{1}^{\epsilon}_{t,t'} = 1$, then the increase in error is $\Pr(\varphi)\mathbb{1}^{\epsilon}_{t,t'} |\mathbb{E}_{\varphi}[S^{p}_{t,t'}] - \mathbb{E}_{\varphi}[S^{q}_{t,t'}]|$. Using Equation (4.1) and summing over all profiles results in the lemma. □

In the next section we develop estimators for $S^{p}_{t,t'}$ and $S^{q}_{t,t'}$.

4.2.3 Conventional estimation and the proposed approach

Empirical and Good-Turing estimators can be naturally extended to joint sequences as $E^{p}_{t,t'} \stackrel{\text{def}}{=} \Phi_{t,t'} \frac{t}{n}$ and $G^{p}_{t,t'} \stackrel{\text{def}}{=} \Phi_{t+1,t'} \frac{t+1}{n}$. As with probability estimation, it is easy to come up with examples where the rate of convergence of these estimates is not optimal. The rate of convergence of Good-Turing and empirical estimators are quantified in the next lemma.

Lemma 4.4 (Empirical and Good-Turing for joint sequences)**.** For every (p, q) and t and t',

$$\left|S^{p}_{t,t'} - G^{p}_{t,t'}\right| \underset{n^{-4}}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_{t+1,t'}] + 1} \frac{(t+1)\log^2 n}{n}\right),$$

and if $\max(t, t') > 0$, then

$$\left|S^{p}_{t,t'} - E^{p}_{t,t'}\right| \underset{n^{-4}}{=} \mathcal{O}\left(\Phi_{t,t'} \frac{\sqrt{t+1}\log n}{n}\right).$$

Similar results hold for $S^{q}_{t,t'}$.

The proof of the above lemma is similar to those of Lemmas 3.17 and 3.18 and hence omitted. Note that the error probability in the above lemma can be

any polynomial in $1/n$. n^{-4} has been chosen to simplify the analysis. Motivated by combined-probability estimation, we propose $F^p_{t_0,t'_0}$ for joint sequences as

$$F^p_{t_0,t'_0} = \Phi_{t_0,t'_0} \frac{t_0+1}{n} \frac{\widehat{\mathbb{E}[\Phi_{t_0+1,t'_0}]}}{\widehat{\mathbb{E}[\Phi_{t_0,t'_0}]}},$$

where $\widehat{\mathbb{E}[\Phi_{t_0,t'_0}]}$ and $\widehat{\mathbb{E}[\Phi_{t_0+1,t'_0}]}$ are estimators for $\mathbb{E}[\Phi_{t_0,t'_0}]$ and $\mathbb{E}[\Phi_{t_0+1,t'_0}]$ respectively. Let $\mathcal{S}_r^{t_0,t'_0} = \{(t,t') \mid |t-t_0| \le r, |t'-t'_0| \le r\}$ and $r_{t_0} = \left\lfloor \frac{\sqrt{t_0}}{(t_0 \Phi_{t_0,t'_0})^{1/12} \log n} \right\rfloor$. The estimators $\widehat{\mathbb{E}[\Phi_{t_0,t'_0}]}$ and $\widehat{\mathbb{E}[\Phi_{t_0+1,t'_0}]}$ are given by

$$\widehat{\mathbb{E}[\Phi_{t_0,t'_0}]} = \sum_{t,t' \in \mathcal{S}_{r_{t_0}}^{t_0,t'_0}} c_{t,t'} \Phi_{t,t'}, \text{ and } \widehat{\mathbb{E}[\Phi_{t_0+1,t'_0}]} = \sum_{t,t' \in \mathcal{S}_{r_{t_0}}^{t_0+1,t'_0}} d_{t,t'} \Phi_{t,t'},$$

where

$$c_{t,t'} = \gamma_{r_{t_0}}(|t-t_0|)\gamma_{r_{t_0}}(|t'-t'_0|) a_t^{t_0} a_{t'}^{t'_0}$$

and

$$d_{t,t'} = \gamma_{r_{t_0}}(|t-t_0-1|)\gamma_{r_{t_0}}(|t'-t'_0|) \frac{t_0}{t_0+1} a_t^{t_0} a_{t'}^{t'_0}.$$

γ_r and $a_t^{t_0}$ are defined in Section 3.6. The estimator $F^q_{t_0,t'_0}$ can be obtained similarly.

The next lemma shows that the estimate for the ratio of $\mathbb{E}[\Phi_{t_0+1,t'_0}]$ and $\mathbb{E}[\Phi_{t_0,t'_0}]$ is close to the actual ratio. The proof is similar to that of Lemma 3.22 and hence omitted.

Lemma 4.5. For every (p,q) and every $t_0 \ge \log^2 n$, if $\frac{1}{t_0}\left(\frac{t_0}{\log^2 n}\right)^6 \ge \mathbb{E}[\Phi_{t_0,t'_0}] \ge \log^2 n$, then

$$\left| \frac{\widehat{\mathbb{E}[\Phi_{t_0+1,t'_0}]}}{\widehat{\mathbb{E}[\Phi_{t_0,t'_0}]}} - \frac{\mathbb{E}[\Phi_{t_0+1,t'_0}]}{\mathbb{E}[\Phi_{t_0,t'_0}]} \right| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\log^3 n}{\sqrt{t_0}(\mathbb{E}[\Phi_{t_0,t'_0}]t_0)^{1/3}} \right),$$

and if $\mathbb{E}[\Phi_{t_0,t'_0}] \ge \frac{1}{t_0}\left(\frac{t_0}{\log^2 n}\right)^6$, then

$$\left| \frac{\widehat{\mathbb{E}[\Phi_{t_0+1,t'_0}]}}{\widehat{\mathbb{E}[\Phi_{t_0,t'_0}]}} - \frac{\mathbb{E}[\Phi_{t_0+1,t'_0}]}{\mathbb{E}[\Phi_{t_0,t'_0}]} \right| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\log^3 n}{\sqrt{\mathbb{E}[\Phi_{t_0,t'_0}]}} \right).$$

Using the previous lemma, we bound the error of $F^p_{t,t'}$ in the next lemma. The proof is similar to that of Lemma 3.30 and hence omitted.

Lemma 4.6. For every (p, q) and $t \geq \log^2 n$, if $\frac{1}{t}\left(\frac{t}{\log^2 n}\right)^6 \geq \mathbb{E}[\Phi_{t,t'}] \geq \log^2 n$, then

$$\left|S^p_{t,t'} - F^p_{t,t'}\right| \underset{2n^{-4}}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_{t,t'}]^{2/3} t^{1/6} \log^3 n}{n} + \frac{\sqrt{\mathbb{E}[\Phi_{t,t'}]t}\log^2 n}{n}\right),$$

and if $\mathbb{E}[\Phi_{t,t'}] > \frac{1}{t}\left(\frac{t}{\log^3 n}\right)^6$, then

$$\left|S^p_{t,t'} - F^p_{t,t'}\right| \underset{2n^{-4}}{=} \mathcal{O}\left(\frac{t\sqrt{\mathbb{E}[\Phi_{t,t'}]}\log^3 n}{n} + \frac{\sqrt{\mathbb{E}[\Phi_{t,t'}]t}\log^2 n}{n}\right).$$

Similar results hold for $S^q_{t,t'}$.

4.2.4 Competitive classifier

The proposed classifier is given below. It estimates $S^p_{t,t'}$ (call it $F'^p_{t,t'}$) and $S^q_{t,t'}$ (call it $F'^q_{t,t'}$) and assigns z to the hypothesis that has the higher estimate. Let t and t' be the multiplicities of the z in $\overline{x}$ and $\overline{y}$ respectively. If $|t - t'| \geq \sqrt{t + t'}\log^2 n$, then the classifier uses empirical estimates. Since t and t' are far apart, by the Chernoff bound such an estimate provides us good bounds for the purposes of classification. In other cases, it uses the estimate with the lowest error bounds, given by Lemma 4.4 for $E^p_{t,t'}, G^p_{t,t'}$, and Lemma 4.6 for $F^p_{t,t'}$. We also set $F'^p_{t,t'} = \min(F^p_{t,t'}, 1)$ and $F'^q_{t,t'} = \min(F^q_{t,t'}, 1)$, to help in the analysis and ensure that the estimates are always ≤ 1.

Classifier $S(\overline{x}, \overline{y}, z)$

Input: Two sequences $\overline{x}$ and $\overline{y}$ and a symbol z.

Output: `x` or `y`.

1. Let $t = t_z(\overline{x})$ and $t' = t_z(\overline{y})$.
2. If $\max(t, t') = 0$, then $F'^p_{t,t'} = G^p_{t,t'}$ and $F'^q_{t,t'} = G^q_{t,t'}$.
3. If $\max(t, t') > 0$ and $|t - t'| \geq \sqrt{t + t'} \log^2 n$ or $\Phi_{t,t'} \leq \log^2 n$, then $F'^p_{t,t'} = E^p_{t,t'}$ and $F'^q_{t,t'} = E^q_{t,t'}$.
4. If $\max(t, t') > 0$, $|t - t'| < \sqrt{t + t'} \log^2 n$, and $\Phi_{t,t'} > \log^2 n$, then
   (a) If $t \geq 4 \log^4 n$, then $F'^p_{t,t'} = F^p_{t,t'}$ and $F'^q_{t,t'} = F^q_{t,t'}$.
   (b) If $t < 4 \log^4 n$, then $F'^p_{t,t'} = G^p_{t,t'}$ and $F'^q_{t,t'} = G^q_{t,t'}$.
5. Set $F'^p_{t,t'} = \min(F'^p_{t,t'}, 1)$ and $F'^q_{t,t'} = \min(F'^q_{t,t'}, 1)$.
6. If $F'^p_{t,t'} > F'^q_{t,t'}$, then return `x`. If $F'^p_{t,t'} < F'^q_{t,t'}$, then return `y`. If $F'^p_{t,t'} = F'^q_{t,t'}$ return `x` or `y` with equal probability.

4.2.5 Proof of Theorem 4.2

The analysis of the classifier is similar to that of the combined-probability estimation, and we outline few key steps. The error in estimating $S^p_{t,t'}$ (and $S^q_{t,t'}$) is quantified in the following lemma.

Lemma 4.7. For every (p, q), $|S^p_{0,0} - F'^p_{0,0}| \underset{10n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)$ and for $(t, t') \neq (0, 0)$ and $|t - t'| \leq \sqrt{t + t'} \log^2 n$,

$$|S^p_{t,t'} - F'^p_{t,t'}| \underset{10n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{\min\left(\Phi^{2/3}_{t,t'}\sqrt{t+1}, \Phi^{1/2}_{t,t'}(t+1)\right)}{n}\right).$$

Similar results hold for $S^q_{t,t'}$.

The analysis of the lemma is similar to that of Lemma 3.23 and hence omitted. We now prove Theorem 4.2 using the above set of results.

Proof of Theorem 4.2. Let $\mathcal{R} = \{(t,t') \mid |t-t'| \le \sqrt{t+t'}\log^2 n\}$. By Lemma 4.3, $\mathcal{E}^S_{p,q}(n)$ is at most

$$\mathcal{E}^{S_{p,q}}_{p,q}(n) + 2\max_p \left(\sum_{(t,t')\in\mathcal{R}} \mathbb{E}[\mathbb{1}^{\epsilon}_{t,t'}|F'^p_{t,t'} - S^p_{t,t'}|] + \sum_{(t,t')\in\mathcal{R}^c} \mathbb{E}[\mathbb{1}^{\epsilon}_{t,t'}|F'^p_{t,t'} - S^p_{t,t'}|] \right).$$

We first show that the second term is $\mathcal{O}(n^{-1.5})$. By Lemma 4.4,

$$|S^p_{t,t'} - E^p_{t,t'}| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\Phi_{t,t'}\sqrt{t}\log n}{n}\right) \quad \text{and} \quad |S^q_{t,t'} - E^q_{t,t'}| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\Phi_{t,t'}\sqrt{t'}\log n}{n}\right).$$

If $|t-t'| \ge \sqrt{t+t'}\log^2 n$, then

$$|S^p_{t,t'} - S^q_{t,t'}| \ge \frac{\Phi_{t,t'}\sqrt{t+t'}\log^2 n}{n}.$$

Hence $\mathbb{1}^{\epsilon}_{t,t'} \underset{2n^{-4}}{=} 0$. Since with $\text{poi}(n)$ samples, the bounds hold with probability $1-\mathcal{O}(n^{-4})$, by Lemma 3.6, with exactly n samples, they hold with probability $1-\mathcal{O}(n^{-3.5})$. Observe that (t,t') takes at most $n\cdot n = n^2$ values. Therefore, by the union bound $\Pr(\mathbb{1}^{\epsilon}_{t,t'} = 1) \le \mathcal{O}(n^{-1.5})$. Hence

$$\max_p \sum_{(t,t')\in\mathcal{R}^c} \mathbb{E}[|F'^p_{t,t'} - S^p_{t,t'}|] = \mathcal{O}(n^{-1.5}).$$

We now consider the case $(t,t') \in \mathcal{R}$. In Lemma 4.7, the bounds on $|F'^p_{t,t'} - S^p_{t,t'}|$ hold with probability $\ge 1-\mathcal{O}(n^{-3})$, with $\text{poi}(n)$ samples. Therefore by Lemma 3.6, with exactly n samples, they hold with probability $\ge 1-\mathcal{O}(n^{-2.5})$, *i.e.,*

$$|F'^p_{t,t'} - S^p_{t,t'}| \underset{\mathcal{O}(n^{-2.5})}{=} \widetilde{\mathcal{O}}\left(\frac{\Phi^{2/3}_{t,t'}(t+t')^{1/2}}{n}\right).$$

Observe that (t,t') takes at most $n\cdot n = n^2$ values, hence by the union bound, the probability that the above bound holds for all $(t,t') \in \mathcal{R}$ is at least $1-\mathcal{O}(n^{-0.5})$. Since $|F'^p_{t,t'} - S^p_{t,t'}| \le 1$, we get

$$\max_p \sum_{(t,t')\in\mathcal{R}} \mathbb{E}[|F'^p_{t,t'} - S^p_{t,t'}|] \le \sum_{(t,t')\in\mathcal{R}} \widetilde{\mathcal{O}}\left(\frac{\Phi^{2/3}_{t,t'}(t+t')^{1/2}}{n}\right) + \mathcal{O}\left(\frac{1}{n^{1/2}}\right).$$

Using techniques similar to those in the proofs Lemma 3.24 and Theorem 3.3, it can be shown that the above quantity is $\le \widetilde{\mathcal{O}}(n^{-1/5})$, thus proving the theorem. $\square$

4.2.6 Lower bound for classification

We show a non-tight converse for the additional error in this section.

Theorem 4.8. For any classifier S there exists (p, q) such that

$$\mathcal{E}^{S}_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\Omega}\left(\frac{1}{n^{1/3}}\right).$$

We construct a distribution q and a collection of distributions $\mathcal{P}$ such that for any distribution $p \in \mathcal{P}$, the optimal label-invariant classification error for (p, q) is $\frac{1}{2} - \Theta\left(\frac{1}{n^{1/3}\log n}\right)$. We then show that any label-invariant classifier incurs an additional error of $\widetilde{\Omega}(n^{-1/3})$ for at least one pair (p', q), where $p' \in \mathcal{P}$. Similar arguments have been used in [56, 57].

Let q be a distribution over $i = 1, 2, \ldots, \frac{n^{1/3}}{\log n}$ such that $q_i = \frac{3i^2 \log^3 n}{cn}$, and $c \leq 2$ is the normalization factor.

Let $\mathcal{P}$ to be a collection of $2^{\frac{n^{1/3}}{2\log n}}$ distributions. For every $p \in \mathcal{P}$, for all odd i, $p_i = q_i \pm \frac{i \log n}{n}$ and $p_{i+1} = q_{i+1} \mp \frac{i \log n}{n}$, such that, $p_i + p_{i+1} = q_i + q_{i+1}$. For every $p \in \mathcal{P}$. $||p - q||_1 = \Theta\left(\frac{1}{n^{1/3}\log n}\right)$. The next lemma states that every distribution $p \in \mathcal{P}$ and q can be classified by a label-invariant classifier with error $\frac{1}{2} - \Theta\left(\frac{1}{n^{1/3}\log n}\right)$.

Lemma 4.9. For every $p \in \mathcal{P}$ and q,

$$\mathcal{E}^{S_{p,q}}_{p,q}(n) = \frac{1}{2} - \Theta\left(\frac{1}{n^{1/3}\log n}\right).$$

Proof sketch of Theorem 4.8. We show that for any classifier S,

$$\max_{p \in \mathcal{P}} \mathcal{E}^{S}_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\Omega}(n^{-1/3})$$

for some $p \in \mathcal{P}$, thus proving the theorem. Since extra information reduces the error probability, we aid the classifier with a genie that associates the multiplicity with the probability of the symbol. Using ideas similar to [56, 50], one can show that the worst error probability of any classifier between q and the set of distribution $\mathcal{P}$ is lower bounded by error probability between q and any mixture on $\mathcal{P}$. We choose the mixture p_0 such that each $p \in \mathcal{P}$ is chosen uniformly at random. Therefore for any classifier S,

$$\max_{p} \mathcal{E}^{S}_{p,q}(n) \geq \sum_{\overline{x},\overline{y},z} \frac{\min\left(q(\overline{x})p_0(\overline{y}, z), p_0(\overline{y})q(\overline{x}, z)\right)}{2}.$$

Using techniques similar to [50], it can be shown that difference between above error and $\mathcal{E}_{p,q}^{S_{p,q}}(n)$ is $\widetilde{\Omega}(n^{-1/3})$. The proof is similar to the lower bounds in the previous chapter and hence omitted. □

Acknowledgement

Chapter 4 is adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Optimal probability estimation with applications to prediction and classification", *Conference on Learning Theory (COLT)*, 2013 [21].

Part II

How far can one predict?

Chapter 5

Estimating the unseen

5.1 Introduction

Species estimation is an important problem in numerous scientific disciplines. Initially used to estimate ecological diversity [58, 59, 60, 61], it was subsequently applied to assess vocabulary size [62, 63], database attribute variation [64], and password innovation [65]. Recently it has found a number of bio-science applications including estimation of bacterial and microbial diversity [66, 67, 68, 69], immune receptor diversity [70], and unseen genetic variations [71].

All approaches to the problem incorporate a statistical model, with the most popular being the *extrapolation model* introduced by Fisher, Corbet, and Williams [72] in 1943. It assumes that n independent samples $X^n \stackrel{\text{def}}{=} X_1, \ldots, X_n$ were collected from an unknown distribution p, and calls for estimating

$$U \stackrel{\text{def}}{=} U(X^n, X_{n+1}^{n+m}) \stackrel{\text{def}}{=} \left|\{X_{n+1}^{n+m}\} \backslash \{X^n\}\right|,$$

the number of hitherto unseen symbols that would be observed if m additional samples $X_{n+1}^{n+m} \stackrel{\text{def}}{=} X_{n+1}, \ldots, X_{n+m}$, were collected from the same distribution.

In 1956, Good and Toulmin [1] predicted U by a fascinating estimator that has since intrigued statisticians and a broad range of scientists alike [2]. To describe the Good-Toulmin estimator we need only a modicum of nomenclature.

The *prevalence* $\Phi_i \stackrel{\text{def}}{=} \Phi_i(X^n)$ of an integer $i \geq 0$ in X^n is the number of symbols appearing i times in X^n. For example, for X^7=*bananas*, $\Phi_1 = 2$ and

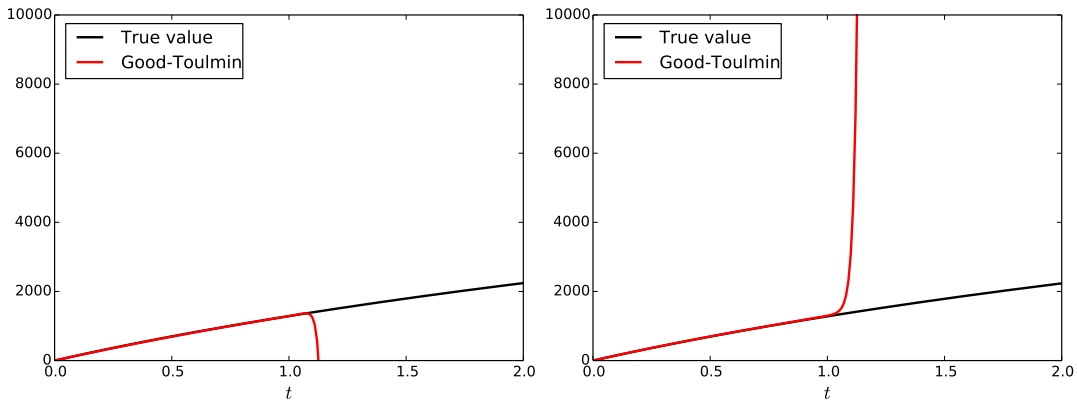


Figure 5.1: GT estimate as a function of t for two realizations random samples of size $n = 5000$ generated by a Zipf distribution $p_i \propto 1/(i+10)$ for $1 \le i \le 10000$.

$\Phi_2 = \Phi_3 = 1$, and in Corbet's table, $\Phi_1 = 118$ and $\Phi_2 = 74$. Let $t \stackrel{\text{def}}{=} \frac{m}{n}$ be the ratio of the number of future and past samples so that $m = tn$. Good and Toulmin estimated U by the surprisingly simple formula

$$U^{\text{GT}} \stackrel{\text{def}}{=} U^{\text{GT}}(X^n, t) \stackrel{\text{def}}{=} -\sum_{i=1}^{\infty} (-t)^i \Phi_i. \tag{5.1}$$

They showed that for all $t \le 1$, U^{GT} is nearly unbiased, and that while U can be as high as nt,*

$$\mathbb{E}(U^{\text{GT}} - U)^2 \lesssim nt^2,$$

hence in expectation, U^{GT} approximates U to within just $\sqrt{n}t$. Figure 5.1 shows that for the ubiquitous Zipf distribution, U^{GT} indeed approximates U well for all $t < 1$. Naturally, we would like to estimate U for as large a t as possible. However, as $t > 1$ increases, U^{GT} grows as $(-t)^i \Phi_i$ for the largest i such that $\Phi_i > 0$. Hence whenever any symbol appears more than once, U^{GT} grows super-linearly in t, eventually far exceeding U that grows at most linearly in t. Figure 5.1 also shows that for the same Zipf distribution, for $t > 1$ indeed U^{GT} does not approximate U at all.

To predict U for $t > 1$, Good and Toulmin [1] suggested using the *Euler transform* [73] that converts an alternating series into another series with the same

*For $a, b > 0$, denote $a \lesssim b$ or $b \gtrsim a$ if $\frac{a}{b} \le c$ for some universal constant c. Denote $a \asymp b$ if both $a \lesssim b$ and $a \gtrsim b$.

sum, and heuristically often converges faster. Interestingly, Efron and Thisted [62] showed that when the Euler transform of U^{GT} is truncated after v terms, it can be expressed as another simple linear estimator,

$$U^{\mathrm{ET}} \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} h_i^{\mathrm{ET}} \cdot \Phi_i,$$

where

$$h_i^{\mathrm{ET}} \stackrel{\mathrm{def}}{=} -(-t)^i \cdot \Pr\left(\mathrm{Bin}\left(v, \frac{1}{1+t}\right) \geq i\right),$$

and

$$\Pr\left(\mathrm{Bin}\left(v, \frac{1}{1+t}\right) \geq i\right) = \begin{cases} \sum_{j=i}^{v} \binom{v}{j} \frac{t^{v-j}}{(1+t)^v} & i \leq v, \\ 0 & i > v, \end{cases}$$

is the *binomial* tail probability that decays with i, thereby moderating the rapid growth of $(-t)^i$.

Over the years, U^{ET} has been used by numerous researchers in a variety of scenarios and a multitude of applications. Yet despite its wide-spread use and robust empirical results, no provable guarantees have been established for its performance or that of any related estimator when $t > 1$. The lack of theoretical understanding, has also precluded clear guidelines for choosing the parameter v in U^{ET}.

5.2 Approach and results

We construct a family of estimators that *provably* predict U optimally not just for constant $t > 1$, but all the way up to $t \propto \log n$. This shows that per each observed sample, we can infer properties of $\log n$ yet unseen samples. The proof technique is general and provides a disciplined guideline for choosing the parameter v for U^{ET} and, in addition, a modification that outperforms U^{ET}.

5.2.1 Smoothed Good-Toulmin (SGT) estimator

To obtain a new class of estimators, we too start with U^{GT}, but unlike U^{ET} that was derived from U^{GT} via analytical considerations aimed at improving the

convergence rate, we take a probabilistic view that controls the bias and variance of U^{GT} and balances the two to obtain a more efficient estimator.

Note that what renders U^{GT} inaccurate when $t > 1$ is not its bias but mainly its high variance due to the exponential growth of the coefficients $(-t)^i$ in (5.1); in fact U^{GT} is the unique unbiased estimator for all t and n in the closely related Poisson sampling model (see Section 5.3). Therefore it is tempting to truncate the series (5.1) at the ℓ^{th} term and use the partial sum as an estimator:

$$U^{\ell} \stackrel{\mathrm{def}}{=} -\sum_{i=1}^{\ell} (-t)^i \Phi_i. \tag{5.2}$$

However, for $t > 1$, it can be shown that for certain distributions most of the symbols typically appear ℓ times and hence the last term in (5.2) dominates, resulting in a large bias and inaccurate estimates regardless of the choice of ℓ (see Section 5.5.1 for a rigorous justification).

To resolve this problem, we truncate the Good-Toulmin estimator at a random location, denoted by an independent random nonnegative integer L, and average over the distribution of L, which yields the following estimator:

$$U^{\mathrm{L}} = \mathbb{E}_L \left[-\sum_{i=1}^{L} (-t)^i \Phi_i \right]. \tag{5.3}$$

The key insight is that since the bias of U^{ℓ} typically alternates signs as ℓ grows, averaging over different cutoff locations takes advantage of the cancellation and dramatically reduces the bias. Furthermore, the estimator (5.3) can be expressed simply as a linear combination of prevalences:

$$U^{\mathrm{L}} = \mathbb{E}_L \left[-\sum_{i\geq 1} (-t)^i \Phi_i \mathbb{1}_{i\leq L} \right] = -\sum_{i\geq 1} (-t)^i \Pr(L \geq i) \Phi_i. \tag{5.4}$$

We shall refer to estimators of the form (5.4) *Smoothed Good-Toulmin (SGT)* estimators and the distribution of L the *smoothing distribution*.

Choosing different smoothing distributions results a variety of linear estimators, where the tail probability $\Pr(L \geq i)$ compensates the exponential growth of $(-t)^i$ thereby stabilizing the variance. Surprisingly, though the motivation and approach are quite different, SGT estimators include U^{ET} in (5.1) as a special case

which corresponds to the binomial smoothing $L \sim \mathrm{Bin}(v, \frac{1}{1+t})$. This provides an intuitive probabilistic interpretation of U^{ET}, which was originally derived via Euler's transform and analytic considerations. As we show in the next section, this interpretation leads to the first theoretical guarantee for U^{ET} as well as improved estimators that are provably optimal.

5.2.2 Main results

Since U takes in values between 0 and nt, we measure the performance of an estimator U^{E} by the worst-case *normalized mean-square error (NMSE)*,

$$\mathcal{E}_{n,t}(U^{\mathrm{E}}) \stackrel{\mathrm{def}}{=} \max_{p} \mathbb{E}_p\left(\frac{U^{\mathrm{E}} - U}{nt}\right)^2.$$

Observe that this criterion conservatively evaluates the performance of the estimator for the worst possible distribution. The trivial estimator that always predicts $nt/2$ new elements has NMSE equal to $1/4$, and we would like to construct estimators with vanishing NMSE, which can estimate U up to an error that diminishes with n, regardless of the data-generating distribution; in particular, we are interested in the largest t for which this is possible.

Relating the bias and variance of U^{L} to the expectation of t^L and another functional we obtain the following performance guarantee for SGT estimators with appropriately chosen smoothing distributions.

Theorem 5.1. For Poisson or binomially distributed L with the parameters given in Table 5.1, for all $t \geq 1$ and $n \in \mathbb{N}$,

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \lesssim \frac{1}{n^{1/t}}.$$

Theorem 5.1 provides a principled way for choosing the parameter v for U^{ET} and the first provable guarantee for its performance, shown in Table 5.1. Furthermore, the result shows that a modification of U^{ET} with $q = \frac{2}{t+2}$ enjoys even faster convergence rate and, as experimentally demonstrated in Section 5.6, outperforms the original version of Efron-Thisted as well as other state-of-the-art estimators.

Table 5.1: NMSE of SGT estimators for three smoothing distributions. Since for any $t \geq 1$, $\log_3(1 + 2/t) \geq \log_2(1 + 1/t) \geq 1/t$, binomial smoothing with $q = 2/(2 + t)$ yields the best convergence rate.

Smoothing distribution	Parameters	$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \lesssim$
Poisson (r)	$r = \frac{1}{2t} \log_e \frac{n(t+1)^2}{t-1}$	$n^{-1/t}$
Binomial (v, q)	$v = \left\lceil \frac{1}{2} \log_2 \frac{nt^2}{t-1} \right\rceil$, $q = \frac{1}{t+1}$	$n^{-\log_2(1+1/t)}$
Binomial (v, q)	$v = \left\lceil \frac{1}{2} \log_3 \frac{nt^2}{t-1} \right\rceil$, $q = \frac{2}{t+2}$	$n^{-\log_3(1+2/t)}$

Furthermore, SGT estimators are essentially optimal as witnessed by the following matching minimax lower bound.

Theorem 5.2. There exist universal constant c, c' such that for any $t \geq c$, any $n \in \mathbb{N}$, and any estimator U^{E}

$$\mathcal{E}_{n,t}(U^{\mathrm{E}}) \gtrsim \frac{1}{n^{c'/t}}.$$

Theorems 5.1 and 5.2 determine the limit of predictability up to a constant multiple.

Corollary 5.3. For any $\delta > 0$,

$$\lim_{n\to\infty} \frac{\max\{t : \mathcal{E}_{n,t}(U^{\mathrm{E}}) < \delta \text{ for some } U^{\mathrm{E}}\}}{\log n} \asymp \frac{1}{\log \frac{1}{\delta}}.$$

The rest of the chapter is organized as follows: In Section 5.3, we describe the four statistical models commonly used across various scientific disciplines, namely, the multinomial, Poisson, hypergeometric, and Bernoulli product models. Among the four models Poisson is the simplest to analyze and hence in Sections 5.4 and 5.5, we first prove Theorem 5.1 for the Poisson model. Finally, in Section 5.6 we demonstrate the efficiency and practicality of our estimators on a variety of synthetic and data sets. In the next chapter, we prove similar results for the other three statistical models and prove lower bound for the multinomial and Poisson models.

5.3 Statistical models

The extrapolation paradigm has been applied to several statistical models. In all of them, an initial sample of size related to n is collected, resulting in a set S_{old} of observed elements. We consider collecting a new sample of size related to m, that would result in a yet unknown set S_{new} of observed elements, and we would like to estimate

$$|S_{\text{new}} \backslash S_{\text{old}}|,$$

the number of unseen symbols that will appear in the new sample. For example, for the observed sample `bananas` and future sample `sonatas`, $S_{\text{old}} = \{\mathtt{a}, \mathtt{b}, \mathtt{n}, \mathtt{s}\}$, $S_{\text{new}} = \{\mathtt{a}, \mathtt{n}, \mathtt{o}, \mathtt{s}, \mathtt{t}\}$, and $|S_{\text{new}} \backslash S_{\text{old}}| = |\{\mathtt{o}, \mathtt{t}\}| = 2$.

Four statistical models have been commonly used in the literature (cf. survey [60] and [61]), and our results apply to all of them. The first three statistical models are also referred as the *abundance models* and the last one is often referred to as the *incidence model* in ecology [61].

Multinomial: This is Good and Toulmin's original model where the samples are independently and identically distributed (i.i.d.), and the initial and new samples consist of exactly n and m elements respectively. Formally, $X^{n+m} = X_1, \ldots, X_{n+m}$ are generated independently according to an unknown discrete distribution of finite or even infinite support, $S_{\text{old}} = \{X^n\}$, and $S_{\text{new}} = \{X_{n+1}^{n+m}\}$.

Hypergeometric: This model corresponds to a sampling-without-replacement variant of the multinomial model. Specifically, X^{n+m} are drawn uniformly without replacement from an unknown collection of symbols that may contain repetitions, for example, an urn with some white and black balls. Again, $S_{\text{old}} = \{X^n\}$ and $S_{\text{new}} = \{X_{n+1}^{n+m}\}$.

Poisson: As in the multinomial model, the samples are also *i.i.d.*, but the sample sizes, instead of being fixed, are Poisson distributed. Formally, $N \sim \text{poi}(n)$, $M \sim \text{poi}(m)$, X^{N+M} are generated independently according to an unknown discrete distribution, $S_{\text{old}} = \{X^N\}$, and $S_{\text{new}} = \{X_{N+1}^{N+M}\}$.

Bernoulli-product: In this model we observe signals from a collection of independent processes over subset of an unknown set $\mathcal{X}$. Every $x \in \mathcal{X}$ is associated with an unknown probability $0 \leq p_x \leq 1$, where the probabilities do not necessarily sum to 1. Each sample X_i is a *subset* of $\mathcal{X}$ where symbol $x \in \mathcal{X}$ appears with probability p_x and is absent with probability $1 - p_x$, independently of all other symbols. $S_{\text{old}} = \cup_{i=1}^{n} X_i$ and $S_{\text{new}} = \cup_{i=n+1}^{n+m} X_i$.

For theoretical analysis in Sections 5.4 and 5.5 we use the Poisson sampling model as the leading example due to its simplicity. Later in the next chapter, we show that very similar results continue to hold for the other three models.

We close this section by discussing two problems that are closely related to the extrapolation model, namely, *support size estimation* and *missing mass estimation*, which correspond to $m = \infty$ and $m = 1$ respectively. Indeed, the probability that the next sample is new is precisely the expected value of U for $m = 1$, which is the goal in the basic Good-Turing problem [14, 74, 19, 37] discussed in the previous two chapters. On the other hand, any estimator U^{E} for U can be converted to a (not necessarily good) support size estimator by adding the number of observed symbols. Estimating the support size of an underlying distribution has been studied by both ecologists [58, 59, 60] and theoreticians [75, 76, 77, 78]; however, to make the problem non-trivial, all statistical models impose a lower bound on the minimum non-zero probability of each symbol, which is assumed to be known to the statistician. We discuss these estimators and their differences to our results in Section 5.4.3.

5.4 Preliminaries and the Poisson model

Throughout the chapter, we use standard asymptotic notation, e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, denote $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if $1/c \leq a_n/b_n \leq c$ for some universal constant $c > 0$. Let $\mathbb{1}_A$ denote the indicator random variable of an event A. Let $\text{Bin}(n, p)$ denote the binomial distribution with n trials and success probability p and let $\text{poi}(\lambda)$ denote the Poisson distribution with mean λ. All logarithms are with respect to the natural base unless otherwise specified.

Let p be a probability distribution over a discrete set $\mathcal{X}$, namely $p_x \geq 0$ for all $x \in \mathcal{X}$ and $\sum_{x\in\mathcal{X}} p_x = 1$. Recall that the sample sizes are Poisson distributed: $N \sim \text{poi}(n)$, $M \sim \text{poi}(m)$, and $t = \frac{m}{n}$. We abbreviate the number of unseen symbols by

$$U \stackrel{\text{def}}{=} U(X^N, X_{N+1}^{N+M}),$$

and we denote an estimator by $U^{\text{E}} \stackrel{\text{def}}{=} U^{\text{E}}(X^N, t)$.

Let N_x and $N_x{}'$ denote the multiplicity of a symbol x in the current samples and future samples, respectively. Let $\lambda_x \stackrel{\text{def}}{=} np_x$. Then a symbol x appears $N_x \sim \text{poi}(np_x) = \text{poi}(\lambda_x)$ times, and for any $i \geq 0$,

$$\mathbb{E}[\mathbb{1}_{N_x=i}] = e^{-\lambda_x}\frac{\lambda_x^i}{i!}.$$

Hence

$$\mathbb{E}[\Phi_i] = \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=i}\right] = \sum_x e^{-\lambda_x}\frac{\lambda_x^i}{i!}.$$

A helpful property of Poisson sampling is that the multiplicities of different symbols are independent of each other. Therefore, for any function $f(x,i)$,

$$\text{Var}\left(\sum_x f(x, N_x)\right) = \sum_x \text{Var}(f(x, N_x)).$$

Many of our derivations rely on these three equations. For example,

$$\mathbb{E}[U] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}] \cdot \mathbb{E}[\mathbb{1}_{N_x{}'>0}] = \sum_x e^{-\lambda_x} \cdot (1 - e^{-t\lambda_x}),$$

and

$$\begin{aligned}\text{Var}(U) &= \text{Var}\left(\sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_x{}'>0}\right) = \sum_x \text{Var}(\mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_x{}'>0}) \\ &\leq \sum_x \mathbb{E}\left[\mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_x{}'>0}\right] = \mathbb{E}\left[U\right].\end{aligned}$$

Note that these equations imply that the standard deviation of U is at most $\sqrt{\mathbb{E}[U]} \ll \mathbb{E}[U]$, hence U highly concentrates around its expectation, and estimating U and $\mathbb{E}[U]$ are essentially the same.

5.4.1 The Good-Toulmin estimator

Before proceeding with general estimators, we prove a few properties of U^{GT}. Under the Poisson model, U^{GT} is in fact the *unique* unbiased estimator for U.

Lemma 5.4 ([62])**.** For any distribution,

$$\mathbb{E}[U] = \mathbb{E}[U^{\mathrm{GT}}].$$

Proof.

$$\begin{aligned}\mathbb{E}[U] &= \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_x>0}\right] = \sum_x e^{-\lambda_x} \cdot \left(1 - e^{-t\lambda_x}\right) \\ &= -\sum_x e^{-\lambda_x} \cdot \sum_{i=1}^{\infty} \frac{(-t\lambda_x)^i}{i!} = -\sum_{i=1}^{\infty}(-t)^i \cdot \sum_x e^{-\lambda_x}\frac{\lambda_x^i}{i!} \\ &= -\sum_{i=1}^{\infty}(-t)^i \cdot \mathbb{E}[\Phi_i] = \mathbb{E}[U^{\mathrm{GT}}].\end{aligned}$$

□

Even though U^{GT} is unbiased for all t, for $t > 1$ it has high variance and hence does not estimate U well even for the simplest distributions.

Lemma 5.5. For any $t > 1$,

$$\lim_{n\to\infty} \mathcal{E}_{n,t}(U^{\mathrm{GT}}) = \infty.$$

Proof. Let p be the uniform distribution over two symbols a and b, namely, $p_a = p_b = 1/2$. First consider even n. Since $(U^{\mathrm{GT}} - U)^2$ is always nonnegative,

$$\mathbb{E}[(U^{\mathrm{GT}} - U)^2] \geq \Pr(N_a = N_b = n/2)(2(-t)^{n/2})^2 = \left(e^{-n/2}\frac{(n/2)^{n/2}}{(n/2)!}\right)^2 4t^n \geq \frac{4t^n}{e^2 n},$$

where we used the fact that $k! \leq (\frac{k}{e})^k\sqrt{k}e$. Hence for $t > 1$,

$$\lim_{n\to\infty} \frac{\mathbb{E}[(U^{\mathrm{GT}} - U)^2]}{(nt)^2} \geq \lim_{n\to\infty} \frac{4t^n}{e^2 n(nt)^2} = \infty.$$

The case of odd n can be shown similarly by considering the event $N_a = \lfloor n/2 \rfloor, N_b = \lceil n/2 \rceil$. □

5.4.2 General linear estimators

Following [62], we consider general linear estimators of the form

$$U^{\mathrm{h}} = \sum_{i=1}^{\infty} \Phi_i \cdot h_i, \tag{5.5}$$

which can be identified with a formal power series $h(y) = \sum_{i=1}^{\infty} \frac{h_i y^i}{i!}$. For example, U^{GT} in (5.1) corresponds to the function $h(y) = 1 - e^{-yt}$. The next lemma bounds the bias and variance of any linear estimator U^{h} using properties of the function h. In Section 5.5.2 we apply this result to the SGT estimator whose coefficients are of the specific form:

$$h_i = -(-t)^i \cdot \Pr\left(L \geq i\right).$$

Let $\Phi_+ \stackrel{\mathrm{def}}{=} \sum_{i=1}^{\infty} \Phi_i$ denote the number of observed symbols.

Lemma 5.6. The bias of U^{h} is

$$\mathbb{E}[U^{\mathrm{h}} - U] = \sum_x e^{-\lambda_x} \left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right),$$

and the variance satisfies

$$\mathrm{Var}(U^{\mathrm{h}} - U) \leq \mathbb{E}[\Phi_+] \cdot \sup_{i \geq 1} h_i^2 + \mathbb{E}[U].$$

Proof. Note that

$$\begin{aligned} U^{\mathrm{h}} - U &= \sum_{i=1}^{\infty} \Phi_i h_i - \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_{x'}>0} \\ &= \sum_{i=1}^{\infty} \sum_x \mathbb{1}_{N_x=i} \cdot h_i - \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_{x'}>0} \\ &= \sum_x \left(\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} \cdot h_i - \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_{x'}>0} \right). \end{aligned}$$

For every symbol x,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} \cdot h_i - \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_{x'}>0} \right] &= \sum_{i=1}^{\infty} e^{-\lambda_x} \frac{\lambda_x^i}{i!} \cdot h_i - e^{-\lambda_x} \cdot (1 - e^{-t\lambda_x}) \\ &= e^{-\lambda_x} \left(\sum_{i=1}^{\infty} \frac{\lambda_x^i h_i}{i!} - (1 - e^{-t\lambda_x}) \right) \\ &= e^{-\lambda_x} \left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right), \end{aligned}$$

from which (5.6) follows. For the variance, observe that for every symbol x,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{\infty}\mathbb{1}_{N_x=i}\cdot h_i-\mathbb{1}_{N_x=0}\cdot\mathbb{1}_{N_{x'}>0}\right) &\le \mathbb{E}\left[\left(\sum_{i=1}^{\infty}\mathbb{1}_{N_x=i}\cdot h_i-\mathbb{1}_{N_x=0}\cdot\mathbb{1}_{N_{x'}>0}\right)^2\right]\\ &\overset{(a)}{=}\mathbb{E}\left[\sum_{i=1}^{\infty}\mathbb{1}_{N_x=i}h_i^2\right]+\mathbb{E}[\mathbb{1}_{N_x=0}]\cdot\mathbb{E}[\mathbb{1}_{N_{x'}>0}]\\ &=\sum_{i=1}^{\infty}\mathbb{E}[\mathbb{1}_{N_x=i}]\cdot h_i^2+\mathbb{E}[\mathbb{1}_{N_x=0}]\cdot\mathbb{E}[\mathbb{1}_{N_{x'}>0}],\end{aligned}$$

where (a) follows as for every $i\neq j$, $\mathbb{E}[\mathbb{1}_{N_x=i}\mathbb{1}_{N_x=j}]=0$. Since the variance of a sum of independent random variables is the sum of variances,

$$\begin{aligned}\mathrm{Var}(U^{\mathrm{h}}-U) &\le \sum_x\sum_{i=1}^{\infty}\mathbb{E}[\mathbb{1}_{N_x=i}]h_i^2+\sum_x\mathbb{E}[\mathbb{1}_{N_x=0}]\cdot\mathbb{E}[\mathbb{1}_{N_{x'}>0}]\\ &=\sum_{i=1}^{\infty}\mathbb{E}[\Phi_i]\cdot h_i^2+\mathbb{E}[U]\\ &\le \mathbb{E}[\Phi_+]\cdot\sup_{i\ge1}h_i^2+\mathbb{E}[U].\end{aligned}$$

□

Lemma 5.6 enables us to reduce the estimation problem to a task on approximating functions. Specifically, in view of (5.6), the goal is to approximate $1-e^{-yt}$ by a function $h(y)$ whose derivatives at zero all have small magnitude.

5.4.3 Estimation via polynomial approximation and support size estimation

Approximation-theoretic techniques for estimating norms and other properties such as support size and entropy have been successfully used in the statistics literature. For example, estimating the L_p norms in Gaussian models [79, 80] and estimating entropy [78, 81] and support size [82] of discrete distributions. Among the aforementioned problems, support size estimation is closest to ours. Hence, we now discuss the difference between the approximation technique we use and the those used for support size estimation.

The support size of a discrete distribution p is

$$\mathrm{Supp}(p)=\sum_x\mathbb{1}_{p_x>0}. \tag{5.6}$$

At the first glance, estimating $\mathrm{Supp}(p)$ may appear similar to species estimation problem as one can convert a support size estimator $\hat{\mathrm{Supp}}$ to $\hat{U}$ by

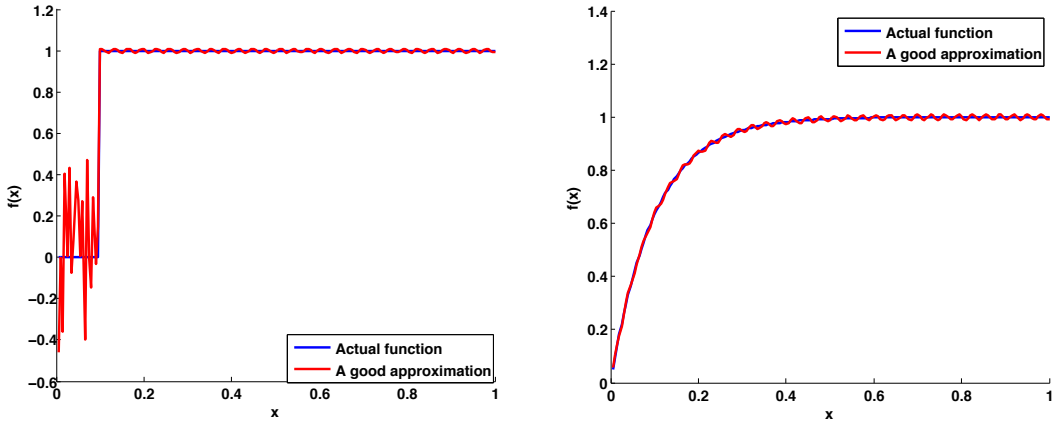
$$\hat{U} = \hat{\mathrm{Supp}} - \sum_{i=1}^{\infty} \Phi_i.$$

However, without any assumption on the distribution it is impossible to estimate the support size. For example, regardless how many samples are collected, there could be infinitely many symbols with arbitrarily small probabilities that will never be observed. A common assumption is therefore that the minimum non-zero probability of the underlying distribution p, denoted by $p^+_{\min}$, is at least $1/k$, for some known k. Under this assumption [76] used a linear programming estimator similar to the one in [62], to estimate the support size within an additive error of $k\epsilon$ with constant probability using $\Omega(\frac{k}{\log k}\frac{1}{\epsilon^2})$ samples. Based on best polynomial approximations recently [82] showed that the minimax risk of support size estimation satisfies

$$\min_{\hat{\mathrm{Supp}}} \max_{p: p^+_{\min} \geq 1/k} \mathbb{E}_p[(\hat{\mathrm{Supp}} - \mathrm{Supp}(p))^2] = k^2 \exp\left(-\Theta\left(\max\left\{\sqrt{\frac{k\log k}{n}}, \frac{k}{n}, 1\right\}\right)\right)$$

and that the optimal sample complexity of for estimating $\mathrm{Supp}(p)$ within an additive error of $k\epsilon$ with constant probability is in fact $\Theta(\frac{k}{\log k}\log^2\frac{1}{\epsilon})$. Note that the assumption $p^+_{\min} \geq 1/k$ is crucial for this result to hold for otherwise estimation is impossible; in contrast, as we show later, for species estimation no such assumptions are necessary. The intuition is that if there exist a large number of very improbable symbols, most likely they will not appear in the new samples anyway.

To estimate the support size, in view of (5.6) and the assumption $p^+_{\min} \geq 1/k$, the technique of [82] is to approximate the indicator function $y \mapsto \mathbb{1}_{y \geq 1/k}$ in the range $\{0\} \cup [1/k, \log k/n]$ using Chebyshev polynomials. Since by assumption no p_x lies in $(0, \frac{1}{k})$, the approximation error in this interval is irrelevant. For example, in Figure 5.2(a), the red curve is a useful approximation for the support size, even though it behaves badly over $(0, 1/k)$. To estimate the average number of unseen symbols U, in view of (5.6), we need to approximate $y \mapsto 1 - e^{-yt}$ over the entire $[0, \infty)$ as in, *e.g.,* Figure 5.2(b). Concurrent to our work, [36] proposed a linear programming algorithm to estimate U. However, their NMSE is $\mathcal{O}(\frac{t}{\log n})$



(a) (b)

Figure 5.2: (a) a good approximation for support size; (b) a good approximation for species estimation.

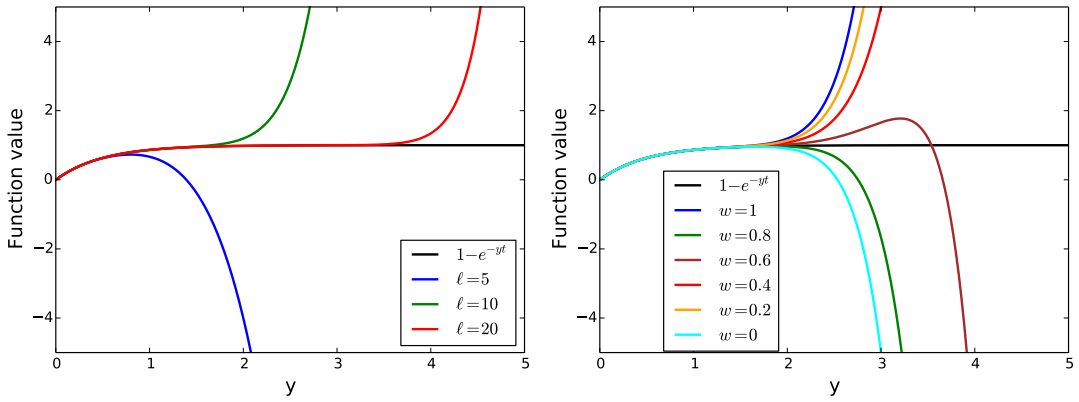
compared to the optimal result $\mathcal{O}(n^{-1/t})$ in Theorem 5.1, thus exponentially weaker for $t = o(\log n)$. Furthermore, the computational cost far exceeds those of our linear estimators.

5.5 Results for the Poisson model

In this section, we provide the performance guarantee for SGT estimators under the Poisson sampling model. We first show that the truncated GT estimators incurs a high bias. We then introduce the class of smoothed GT estimators obtained by averaging several truncated GT estimators and bound their mean squared error in Theorem 5.11 for an arbitrary smoothing distribution. We then apply this result to obtain NMSE bounds for Poisson and Binomial smoothing in Corollaries 5.12 and 5.13 respectively, which imply the main result (Theorem 5.1) announced in Section 5.2.2 for the Poisson model.

5.5.1 Why truncated Good-Toulmin does not work

Before we discuss the SGT estimator, we first show that the naive approach of truncating the GT estimator described in Section 5.2.1 leads to bad performance



(a) (b) .

Figure 5.3: (a) Taylor approximation for $t = 2$, (b) Averages of 10 and 11 term Taylor approximation for $t = 2$.

when $t > 1$. Recall from Lemma 5.6 that designing a good linear estimator boils to approximating $1 - e^{-yt}$ by an analytic function $h(y) = \sum_{i\geq 1} \frac{h_i}{i!} y^i$ such that all its derivatives at zero are small, namely, $\sup_{i\geq 1} |h_i|$ is small. The GT estimator corresponds to the perfect approximation

$$h^{\mathrm{GT}}(y) = 1 - e^{-yt};$$

however, $\sup_{i\geq 1} |h_i| = \max(t, t^\infty)$, which is infinity if $t > 1$ and leads to large variance. To avoid this situation, a natural approach is to use use the ℓ-term Taylor expansion of $1 - e^{-yt}$ at 0, namely,

$$h^\ell(y) = -\sum_{i=1}^{\ell} \frac{(-yt)^i}{i!}, \tag{5.7}$$

which corresponds to the estimator U^ℓ defined in (5.2). Then $\sup_{i\geq 1} |h_i| = t^\ell$ and, by Lemma 5.6, the variance is at most $n(t^\ell + t)$. Hence if $\ell \leq \log_t m$, the variance is at most $n(m+t)$. However, note that the ℓ-term Taylor approximation is a degree-ℓ polynomial which eventually diverges and deviates from $1 - e^{-yt}$ as y increases, thereby incurring a large bias. Figure 5.3(a) illustrates this phenomenon by plotting the function $1 - e^{-yt}$ and its Taylor expansion with $5, 10$, and 20 terms. Indeed, the next result rigorously shows that the NMSE of truncated GT estimator never vanishes:

Lemma 5.7. There exist a constant $c > 0$ such that for any $\ell \geq 0$, any $t > 1$ and any $n \in \mathbb{N}$,

$$\mathcal{E}_{n,t}(U^\ell) \geq \frac{c(t-1)^5}{t^4}.$$

Proof. To rigorously prove an impossibility result for the truncated GT estimator, we demonstrate a particular distribution under which the bias is large. Consider the uniform distribution over $n/(\ell+1)$ symbols, where ℓ is a non-zero even integer. By Lemma 5.6, for this distribution the bias is

$$\begin{aligned}
\mathbb{E}[U - U^\ell] &= \sum_x e^{-\lambda_x}(1 - e^{-\lambda_x t} - h(\lambda_x)) \\
&= \frac{n}{\ell+1} e^{-(\ell+1)} \left(1 - e^{-(\ell+1)t} + \sum_{i=1}^{\ell} \frac{(-(\ell+1)t)^i}{i!}\right) \\
&\geq \frac{n}{\ell+1} e^{-(\ell+1)} \left(\sum_{i=1}^{\ell} \frac{(-(\ell+1)t)^i}{i!}\right) \\
&\overset{(a)}{\geq} \frac{n}{\ell+1} e^{-(\ell+1)} \left(\frac{((\ell+1)t)^\ell}{\ell!} - \frac{((\ell+1)t)^{\ell-1}}{(\ell-1)!}\right) \\
&\geq \frac{n}{(\ell+1)} e^{-(\ell+1)} \frac{((\ell+1)t)^\ell}{\ell!} \cdot \frac{(t-1)}{t} \\
&\geq \frac{n}{3(\ell+1)^{3/2}} t^\ell \frac{(t-1)}{t} \geq \frac{n}{3 \cdot 2^{3/2}} \frac{t^\ell}{\ell^{3/2}} \frac{(t-1)}{t},
\end{aligned}$$

where (a) follows from the fact that $\frac{(-(\ell+1)t)^i}{i!}$ for $i = 1, \ldots, \ell$ is an alternating series with increasing magnitude of terms. Hence

$$\mathbb{E}[U - U^\ell] \geq \frac{n}{3 \cdot 2^{3/2}} \frac{(t-1)}{t} \min_{\ell \in \{2,4,\ldots\}} \frac{t^\ell}{\ell^{3/2}}.$$

For $t \geq 2$, the above minimum occurs at $\ell = 2$ and hence $\min_{\ell \in \{2,4,\ldots\}} \frac{t^\ell}{\ell^{3/2}} \geq \frac{(t-1)^{3/2}}{2^{3/2}}$. For $1 < t < 2$, using the fact that $e^y \geq ey$ for $y > 0$ and $\log t \geq (t-1)\log 2$ for $1 < t < 2$, we have $\min_{\ell \in \{2,4,\ldots\}} \frac{t^\ell}{\ell^{3/2}} \geq (\frac{2e \log t}{3})^{3/2} \geq (\frac{2e \log 2 (t-1)}{3})^{3/2}$. Thus for any even value of $\ell > 0$,

$$\mathbb{E}[U - U^\ell] \geq \frac{n(t-1)^{5/2}}{6.05t}.$$

A similar argument holds for odd values of ℓ and $\ell = 0$, showing that $|\mathbb{E}[U - U^\ell]| \gtrsim \frac{n(t-1)^{5/2}}{t}$ and hence the desired NMSE bound. □

5.5.2 Smoothing by random truncation

As we saw in the previous section, the ℓ-term Taylor approximation, where all the coefficients after the ℓ^{th} term are set to zero results in large bias. Instead, one can choose a weighted average of several Taylor series approximations, whose biases cancel each other leading to significant bias reduction. For example, in Figure 5.3(b), we plot

$$wh^{10} + (1-w)h^{11}$$

for various values of $w \in [0,1]$. Notice that the weight $w = 0.6$ leads to better approximation of $1 - e^{-yt}$ than both h^{10} and h^{11}.

A natural generalization of the above argument entails taking the weighted average of various Taylor approximations with respect to a given probability distribution over $\mathbb{Z}_+ \stackrel{\text{def}}{=} \{0, 1, 2, \ldots\}$. For a $\mathbb{Z}_+$-valued random variable L, consider the power series

$$h^{\text{L}}(y) = \sum_{\ell=0}^{\infty} \Pr(L = \ell) \cdot h^{\ell}(y),$$

where h^{ℓ} is defined in (5.7). Rearranging terms, we have

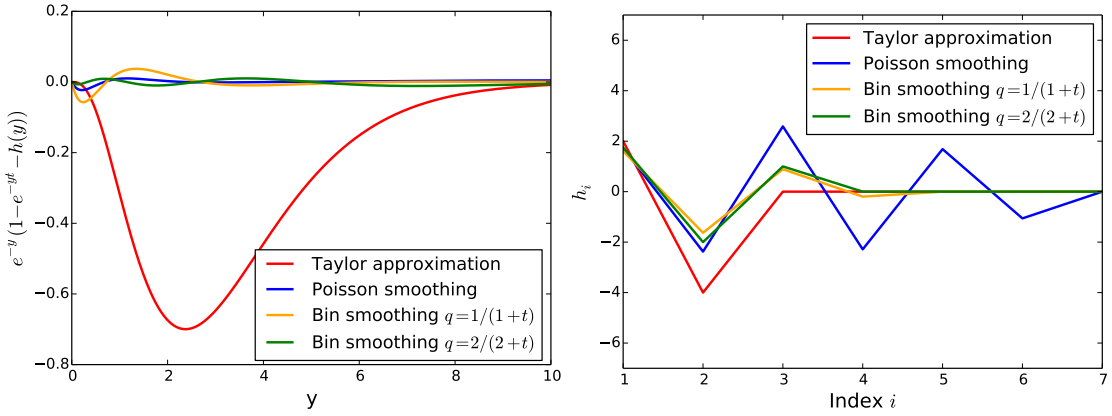
$$h^{\text{L}}(y) = \sum_{\ell=0}^{\infty} \Pr(L = \ell) \sum_{i=1}^{\ell} \frac{-(-yt)^i}{i!} = -\sum_{i=1}^{\infty} \frac{(-yt)^i}{i!} \Pr(L \geq i).$$

Thus, the linear estimator with coefficients

$$h_i^{\text{L}} = -(-t)^i \Pr(L \geq i), \tag{5.8}$$

is precisely the SGT estimator U^{L} defined in (5.4). Special cases of smoothing distributions include:

- $L = \infty$: This corresponds to the original Good-Toulmin estimator (5.1) without smoothing;
- $L = \ell$ deterministically: This leads to the estimator U^{ℓ} in (5.2) corresponding to the ℓ-term Taylor approximation;
- $L \sim \text{Bin}(v, 1/(1+t))$: This recovers the Efron-Thisted estimator (5.1), where v is a tuning parameter to be chosen.



(a) (b)

Figure 5.4: Comparisons of approximations of $h^{\text{L}}(\cdot)$ with $\mathbb{E}[L] = 2$ and $t = 2$. (a) $e^{-y}(1 - e^{-yt} - h^{\text{L}}(y))$ as a function of y. (b) Coefficients h_i^{L} as a function of index i.

We study the performance of linear estimators corresponding to the Poisson smoothing and the Binomial smoothing. To this end, we first systematically upper bound the bias and variance for any probability smoothing L. We plot the error that corresponds to each smoothing in Figure 5.4(a). Notice that the Poisson and binomial smoothings have significantly small error compared to the Taylor series approximation. The coefficients of the resulting estimator is plotted in Figure 5.4(b). It is easy so see that the maximum absolute value of the coefficient is higher for the Taylor series approximation compared to the Poisson or binomial smoothings.

Lemma 5.8. For a random variable L over $\mathbb{Z}_+$ and $t \geq 1$,

$$\text{Var}(U^{\text{L}} - U) \leq \mathbb{E}[\Phi_+] \cdot \mathbb{E}^2[t^L] + \mathbb{E}[U].$$

Proof. By Lemma 5.6, to bound the variance it suffices to bound the highest coefficient in h^{L}.

$$|h_i^{\text{L}}| \leq t^i \Pr(L \geq i) = t^i \sum_{j=i}^{\infty} \Pr(L = j) \leq \sum_{j=i}^{\infty} \Pr(L = j) t^j \leq \mathbb{E}[t^L]. \tag{5.9}$$

The above bound together with Lemma 5.6 yields the result. □

To bound the bias, we need few definitions. Let

$$g(y) \stackrel{\text{def}}{=} -\sum_{i=1}^{\infty} \frac{\Pr(L \geq i)}{i!} (-y)^i. \tag{5.10}$$

Under this definition, $h^{\text{L}}(y) = g(yt)$. We use the following auxiliary lemma to bound the bias.

Lemma 5.9. For any random variable L over $\mathbb{Z}_+$,

$$g(y) - (1 - e^{-y}) = -e^{-y} \int_0^y \mathbb{E}\left[\frac{(-s)^L}{L!}\right] e^s \mathrm{d}s.$$

Proof. Subtracting (5.10) from the Taylor series expansion of $1 - e^{-y}$,

$$\begin{aligned} g(y) - (1 - e^{-y}) &= \sum_{i=1}^{\infty} \frac{\Pr(L < i)}{i!}(-y)^i \\ &= \sum_{i=1}^{\infty} \sum_{j=0}^{i-1} \frac{(-y)^i}{i!} \Pr(L = j) \\ &= \sum_{j=0}^{\infty} \left(\sum_{i=j+1}^{\infty} \frac{(-y)^i}{i!} \right) \Pr(L = j). \end{aligned}$$

Note that $\sum_{i=j+1}^{\infty} \frac{z^i}{i!}$ can be expressed (via incomplete Gamma function) as

$$\sum_{i=j+1}^{\infty} \frac{z^i}{i!} = \frac{e^z}{j!} \int_0^z \tau^j e^{-\tau} \mathrm{d}\tau.$$

Thus by Fubini's theorem,

$$\begin{aligned} g(y) - (1 - e^{-y}) &= \sum_{j=0}^{\infty} \frac{e^{-y}}{j!} \int_0^{-y} \tau^j e^{-\tau} \mathrm{d}\tau \Pr(L = j) \\ &= e^{-y} \int_0^{-y} e^{-\tau} \mathrm{d}\tau \left(\sum_{j=0}^{\infty} \frac{\tau^j}{j!} \Pr(L = j) \right) \\ &= -e^{-y} \int_0^{y} e^{s} \mathrm{d}s \left(\sum_{j=0}^{\infty} \frac{(-s)^j}{j!} \Pr(L = j) \right) \\ &= -e^{-y} \int_0^{y} \mathbb{E}\left[\frac{(-s)^L}{L!}\right] e^s \mathrm{d}s. \end{aligned}$$

□

To bound the bias, we need one more definition. For a random variable L over $\mathbb{Z}_+$, let

$$\xi_L(t) \stackrel{\text{def}}{=} \max_{0 \le s < \infty} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| e^{-s/t},$$

Lemma 5.10. For a random variable L over $\mathbb{Z}_+$,

$$|\mathbb{E}[U^{\text{L}} - U]| \le (\mathbb{E}[\Phi_+] + \mathbb{E}[U]) \cdot \xi_L(t).$$

Proof. By Lemma 5.9,

$$\begin{aligned}|g(y) - (1 - e^{-y})| &\leq e^{-y} \int_0^y \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| e^s \mathrm{d}s \\ &\leq \max_{s \leq y} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| e^{-y} \int_0^y e^s \mathrm{d}s \\ &= \max_{s \leq y} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| (1 - e^{-y}).\end{aligned}$$

For a symbol x,

$$e^{-\lambda_x} \left(h^{\mathrm{L}}(\lambda_x) - (1 - e^{-\lambda_x t}) \right) = e^{-\lambda_x} \left(g(\lambda_x t) - (1 - e^{-\lambda_x t}) \right).$$

Hence,

$$\begin{aligned}|e^{-\lambda_x} \left(h^{\mathrm{L}}(\lambda_x) - 1 - e^{-\lambda_x t} \right)| &\leq (1 - e^{-\lambda_x t}) \max_{0 \leq y \leq \infty} e^{-y} \max_{0 \leq s \leq yt} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| \\ &\leq (1 - e^{-\lambda_x t}) \max_{0 \leq s \leq \infty} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| e^{-s/t}.\end{aligned}$$

The lemma follows by summing over all the symbols and substituting $\sum_x 1 - e^{-\lambda_x t} \leq \sum_x 1 - e^{-\lambda_x (t+1)} = \mathbb{E}[\Phi_+] + \mathbb{E}[U]$. □

The above two lemmas yield our main result.

Theorem 5.11. For any random variable L over $\mathbb{Z}_+$ and $t \geq 1$,

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \leq \mathbb{E}[\Phi_+] \cdot \mathbb{E}^2[t^L] + \mathbb{E}[U] + (\mathbb{E}[\Phi_+] + \mathbb{E}[U])^2 \xi_L(t)^2.$$

We have therefore reduced the problem of computing mean-squared loss, to that of computing expectation of certain function of the random variable. We now apply the above theorem for Binomial and Poisson smoothings. Notice that the above bound is distribution dependent and can be used to obtain stronger results for certain distributions. However, in the rest of the chapter, we concentrate on obtaining minimax guarantees.

5.5.3 Poisson smoothing

Corollary 5.12. For $t \geq 1$, $L \sim \mathrm{poi}(r)$ with $r = \frac{1}{2t} \log\left(\frac{n(t+1)^2}{t-1}\right)$,

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \leq \frac{c_t}{n^{1/t}},$$

where $0 \leq c_t \leq 3$ and $\lim_{t \to \infty} c_t = 1$.

Proof. For $L \sim \text{poi}(r)$,

$$\mathbb{E}[t^L] = e^{-r} \sum_{\ell=0}^{\infty} \frac{(rt)^\ell}{\ell!} = e^{r(t-1)}. \tag{5.11}$$

Furthermore,

$$\mathbb{E}\left[\frac{(-s)^L}{L!}\right] = e^{-r} \sum_{j=0}^{\infty} \frac{(-sr)^j}{(j!)^2} = e^{-r} J_0(2\sqrt{sr}),$$

where J_0 is the Bessel function of first order which takes values in $[-1, 1]$ cf. [73, 9.1.60]. Therefore

$$\xi_L(t) \leq e^{-r}. \tag{5.12}$$

Equations (5.11) and (5.12) together with Theorem 5.11 yields

$$\mathbb{E}[(U^{\text{L}} - U)^2] \leq \mathbb{E}[\Phi_+] \cdot e^{2r(t-1)} + \mathbb{E}[U] + (\mathbb{E}[\Phi_+] + \mathbb{E}[U])^2 \cdot e^{-2r}.$$

Since $\mathbb{E}[\Phi_+] \leq n$ and $\mathbb{E}[U] \leq nt$,

$$\mathbb{E}[(U^{\text{L}} - U)^2] \leq ne^{2r(t-1)} + nt + (n + nt)^2 e^{-2r}.$$

Choosing $r = \frac{1}{2t} \log \frac{n(t+1)^2}{t-1}$ yields

$$\mathcal{E}_{n,t}(U^{\text{L}}) \leq \frac{1}{(nt)^{1/t}} \cdot \left(\frac{t(t-1)}{(t+1)^2}\right)^{\frac{1-t}{t}} + \frac{1}{nt},$$

and the lemma with $c_t \stackrel{\text{def}}{=} \frac{1}{t^{1/t}} \cdot \left(\frac{t(t-1)}{(t+1)^2}\right)^{\frac{1-t}{t}} + \frac{1}{t}$. □

5.5.4 Binomial smoothing

We now prove the results when $L \sim \text{Bin}(v, q)$. Our analysis holds for all $q \in [0, 2/(2+t)]$ and in this range, the performance of the estimator improves as q increases, and hence the NMSE bounds are strongest for $q = 2/(2+t)$. Therefore, we consider binomial smoothing for two cases: the Efron-Thisted suggested value $q = 1/(1+t)$ and the optimized value $q = 2/(2+t)$.

Corollary 5.13. For $t \geq 1$ and $L \sim \text{Bin}(v, q)$, if $v = \left\lceil \frac{1}{2} \log_2 \frac{nt^2}{t-1} \right\rceil$ and $q = \frac{1}{t+1}$, then

$$\mathcal{E}_{n,t}(U^{\text{L}}) \leq \frac{c_t}{n^{\log_2(1+1/t)}},$$

where c_t satisfies $0 \le c_t \le 6$ and $\lim_{t\to\infty} c_t = 1$; if $v = \left\lceil \frac{1}{2}\log_3 \frac{nt^2}{t-1} \right\rceil$ and $q = \frac{2}{t+2}$, then

$$\mathcal{E}_{n,t}(U^{\text{L}}) \le \frac{c'_t}{(nt)^{\log_3(1+2/t)}},$$

where c'_t satisfies $0 \le c'_t \le 6$ and $\lim_{t\to\infty} c'_t = 1$.

Proof. If $L \sim \text{Bin}(v, q)$,

$$\mathbb{E}[t^L] = \sum_{\ell=0}^{v} \binom{v}{\ell} (tq)^\ell (1-q)^{v-\ell} = (1 + q(t-1))^v.$$

Furthermore,

$$\mathbb{E}\left[\frac{(-s)^L}{L!}\right] = \sum_{j=0}^{v} \frac{(-s)^j}{j!}\binom{v}{j}(q)^j(1-q)^{v-j} = (1-q)^v L_v\left(\frac{qs}{1-q}\right),$$

where

$$L_v(y) = \sum_{j=0}^{v} \frac{(-y)^j}{j!}\binom{v}{j} \tag{5.13}$$

is the Laguerre polynomial of degree v. If $\frac{tq}{2(1-q)} \le 1$, for any $s \ge 0$,

$$e^{-\frac{s}{t}}\left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| \le (1-q)^v e^{-\frac{s}{t}} e^{\frac{qs}{2(1-q)}} \le (1-q)^v,$$

where the second inequality follows from the fact cf. [73, 22.14.12] that for all $y \ge 0$ and all $v \ge 0$,

$$|L_v(y)| \le e^{y/2}. \tag{5.14}$$

Hence for $q \le 2/(t+2)$,

$$\mathbb{E}[(U^{\text{L}} - U)^2] \le \mathbb{E}[\Phi_+] \cdot (1 + q(t-1))^{2v} + \mathbb{E}[U] + (\mathbb{E}[\Phi_+] + \mathbb{E}[U])^2 \cdot (1-q)^{2v}.$$

Since $\mathbb{E}[U] \le nt$ and $\mathbb{E}[\Phi_+] \le n$,

$$\mathbb{E}[(U^{\text{L}} - U)^2] \le n \cdot (1 + q(t-1))^{2v} + nt + (nt+n)^2 \cdot (1-q)^{2v}. \tag{5.15}$$

Substituting the Efron-Thisted suggested $q = \frac{1}{t+1}$ results in

$$\mathcal{E}_{n,t}(U^{\text{L}}) \le \left(\frac{2^{2v}}{nt^2} + \frac{(t+1)^2}{t^2}\right)\left(\frac{t}{t+1}\right)^{2v} + \frac{1}{nt}.$$

Choosing $v = \left\lceil \frac{1}{2} \log_2 \frac{nt^2}{t-1} \right\rceil$ yields the first result with

$$c_t \stackrel{\text{def}}{=} \left(\frac{4}{t-1} + \left(\frac{t+1}{t} \right)^2 \right) \cdot \left(\frac{t-1}{t^2} \right)^{\log_2(1+1/t)} + \frac{1}{t}.$$

For the second result, substituting $q = \frac{2}{t+2}$ in (5.15) results in

$$\mathcal{E}_{n,t}(U^{\text{L}}) \leq \left(\frac{3^{2v}}{nt^2} + \frac{(t+1)^2}{t^2} \right) \left(\frac{t}{t+2} \right)^{2v} + \frac{1}{nt}.$$

Choosing $v = \left\lceil \frac{1}{2} \log_3 \frac{nt^2}{t-1} \right\rceil$ yields the result with

$$c'_t \stackrel{\text{def}}{=} \left(\frac{9}{t-1} + \frac{(t+1)^2}{t^2} \right) \cdot \left(\frac{t-1}{t^2} \right)^{\log_3(1+2/t)} + \frac{1}{t}. \qquad \square$$

In terms of the exponent, the result is strongest for $L \sim \text{Bin}(v, 2/(t+2))$. Hence, we state the following asymptotic result, which is a direct consequence of Corollary 5.13:

Corollary 5.14. For $L \sim \text{Bin}(v, q)$, $q = \frac{2}{t+2}$,$v = \lceil \log_3(\frac{nt^2}{t-1}) \rceil$, and any fixed δ, the maximum t till which U^{L} incurs a NMSE of δ is

$$\lim_{n\to\infty} \frac{\max\{t : \mathcal{E}_{n,t}(U^{\text{L}}) < \delta\}}{\log n} \geq \frac{2}{\log 3 \cdot \log \frac{1}{\delta}}.$$

Proof. By Corollary 5.13, if $t \to \infty$, then

$$\mathcal{E}_{n,t}(U^{\text{L}}) \leq (1 + o(1)) n^{-\frac{2+o(1)}{t \log 3}}.$$

where $o(1) = o_t(1)$ is uniform in n. Consequently, if $t = (\alpha + o(1)) \log n$ and $n \to \infty$, then

$$\limsup_{n\to\infty} \mathcal{E}_{n,t}(U^{\text{L}}) \leq e^{-\frac{2}{\alpha \log 3}}.$$

Thus for any fixed δ, the maximum t till which U^{L} incurs a NMSE of δ is

$$\lim_{n\to\infty} \frac{\max\{t : \mathcal{E}_{n,t}(U^{\text{L}}) < \delta\}}{\log n} \geq \frac{2}{\log 3 \cdot \log \frac{1}{\delta}}. \qquad \square$$

Corollaries 5.12 and 5.13 imply Theorem 5.1 for the Poisson model.

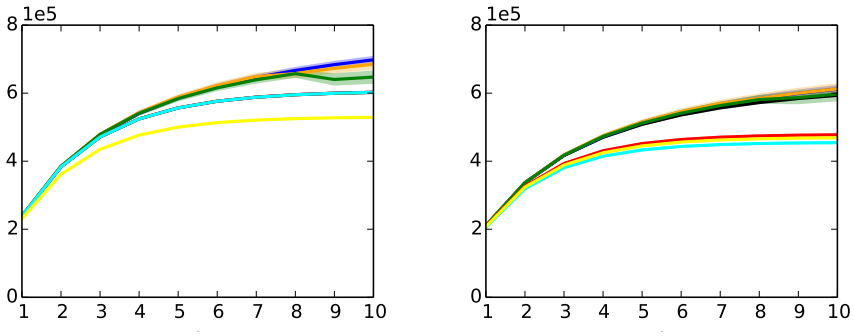
5.6 Experiments

We demonstrate the efficacy of our estimators by comparing their performance with that of several state-of-the-art support-size estimators currently used by ecologists: Chao-Lee estimator [58, 59], Abundance Coverage Estimator (ACE) [83], and the jackknife estimator [84], combined with the Shen-Chao-Lin unseen-species estimator [85]. We consider various natural synthetic distributions and established datasets. Starting with the former, Figure 5.5 shows the *species discovery curve*, the prediction of U as a function of t of several predictors for various distributions.

The true value is shown in black, and the other estimators are color coded, with the solid line representing their mean estimate, and the shaded area corresponding to one standard deviation. Note that the Chao-Lee and ACE estimators are designed specifically for uniform distributions, hence in Figure 5.5(a) they coincide with the true value, but for all other distributions, our proposed smoothed Good-Toulmin estimators outperform the existing ones.

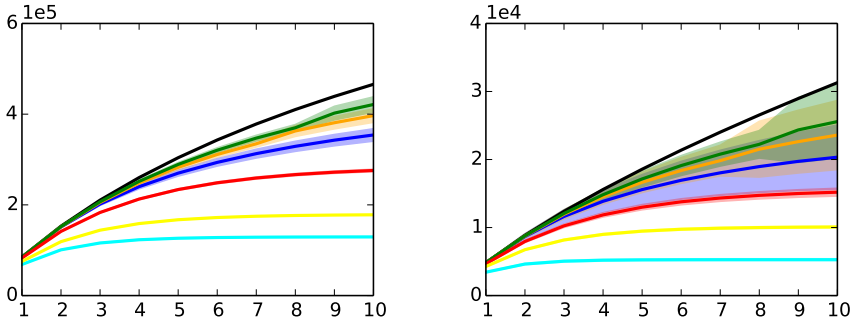
Of the proposed estimators, the binomial-smoothing estimator with parameter $q = \frac{2}{2+t}$ has a stronger theoretical guarantee and performs slightly better than the others. Hence when considering real data we plot only its performance and compare it with the other state-of-the art estimators. We test the estimators on three real datasets taken from various scientific applications where the samples size n ranges from few hundreds to a million. For all these date sets, our estimator outperforms the existing procedures.

Figure 5.6(a) shows the first real-data experiment, predicting vocabulary size based on partial text. Shakespeare's play *Hamlet* consists of $n_{\text{total}} = 31999$ words, of which 4804 are distinct. We randomly select n of the n_{total} words without replacement, predict the number of unseen words in $n_{\text{total}} - n$ new ones, and add it to those observed. The results shown are averaged over 100 trials. Observe that the new estimator outperforms existing ones and that as little as 20% of the data already yields an accurate estimate of the total number of distinct words. Figure 5.6(b) repeats the experiment but instead of random sampling, uses the first n consecutive words, with similar conclusions.



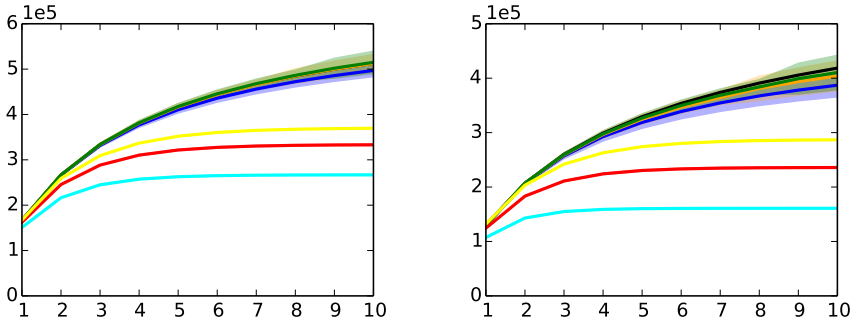
(a) Uniform

(b) Two steps: $\frac{1}{2k} \times \frac{k}{2} \cup \frac{3}{2k} \times \frac{k}{2}$



(c) Zipf-1: $p_i \propto \frac{1}{i}$

(d) Zipf-1.5: $p_i \propto \frac{1}{i^{1.5}}$

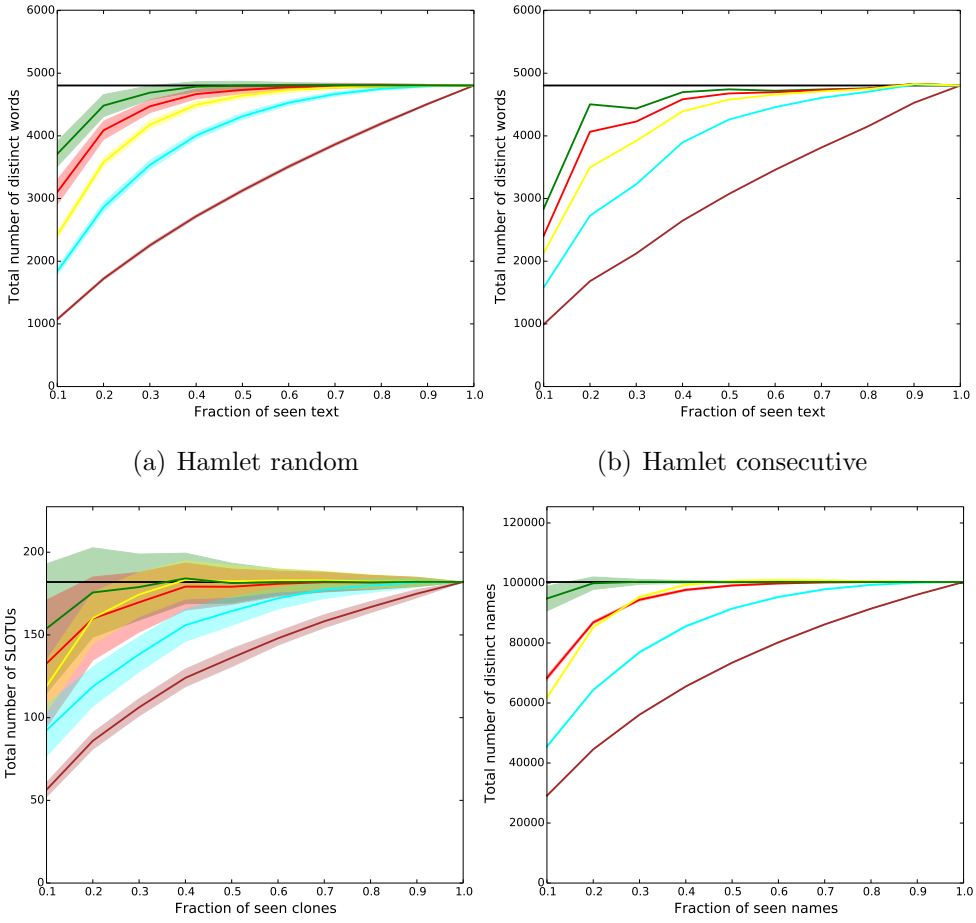


(e) Dirichlet-1 prior

(f) Dirichlet-1/2 prior

True value	Previous	Proposed
——	Chao-Lee ——	Poisson smoothing ——
	ACE ——	Binomial smoothing $q = 1/(t+1)$ ——
	Jackknife ——	Binomial smoothing $q = 2/(t+2)$ ——

Figure 5.5: Comparisons of the estimated number of unseen species as a function of t. All experiments have distribution support size 10^6, $n = 5 \cdot 10^5$, and are averaged over 100 iterations.



(a) Hamlet random

(b) Hamlet consecutive

(c) SLOTUs

(d) Last names

True value	Previous	Proposed
——	Chao-Lee ——	Binomial smoothing $q = 2/(t+2)$ ——
	ACE ——	
	Jackknife ——	
	Empirical ——	

Figure 5.6: Estimates for number of: (a) distinct words in Hamlet with random sampling (b) distinct words in Hamlet with consecutive sampling (c) SLOTUs on human skin (d) last names.

Figure 5.6(c) estimates the number of bacterial species on the human skin. [69] considered forearm skin biota of six subjects. They identified $n_{\text{total}} = 1221$ clones consisting of 182 different species-level operational taxonomic units (SLOTUs). As before, we select n out of the n_{total} clones without replacement and predict the number of distinct SLOTUs found. Again the estimates are more accurate than those of existing estimators and are reasonably accurate already with 20% of the data.

Finally, Figure 5.6(d) considers the 2000 United States Census [86], which lists all U.S. last names corresponding to at least 100 individuals. With these many repetitions, even just a small fraction of the data will cover all names, hence we first subsampled the data $n_{\text{total}} = 10^6$ and obtained a list of 100328 distinct last names. As before we estimate for this number using n randomly chosen names, again with similar conclusions.

Acknowledgement

Chapter 5 is adapted from Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu, "Estimating the number of unseen species: A bird in the hand is worth $\log n$ in the bush", Manuscript, 2015 [87].

Chapter 6

Extensions and lower bounds

In the previous chapters, we proved the performance of our estimators under the Poisson model. In this chapter we extend them to the multinomial model (fixed sample size), the Bernoulli-product model, and the hypergeometric model (sampling without replacement) [60], for which upper bounds of NMSE for general smoothing distributions that are analogous to Theorem 5.11 are presented in Theorem 6.3, 6.5 and 6.11, respectively. Using these results, we obtain the NMSE for Poisson and Binomial smoothings similar to Corollaries 5.12 and 5.13. We remark that up to multiplicative constants, the NMSE under multinomial and Bernoulli-product model are similar to those of Poisson model; however, the NMSE under hypergeometric model is slightly larger. Finally, we also prove lower bounds on the performance any estimator for the multinomial and Poisson models in Section 6.4.

6.1 The multinomial model

The multinomial model corresponds to the setting described in Section 5.1, where upon observing n i.i.d. samples, the objective is to estimate the expected number of new symbols $U(X^n, X_{n+1}^{n+m})$ that would be observed if we took m more samples. We can write the expected number of new symbols as

$$U(X^n, X_{n+1}^{n+m}) = \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_x'>0}.$$

As before we abbreviate

$$U \stackrel{\text{def}}{=} U(X^n, X_{n+1}^{n+m})$$

and similarly $U^{\text{E}} \stackrel{\text{def}}{=} U^{\text{E}}(X^n, t)$ for any estimator E. The difficulty in handling multinomial distributions is that, unlike the Poisson model, the number of occurrences of symbols are correlated; in particular, they sum up to n. This dependence renders the analysis cumbersome. In the multinomial setting each symbol is distributed according to $\text{Bin}(n, p_x)$ and hence

$$\mathbb{E}[\mathbb{1}_{N_x=i}] = \binom{n}{i} p_x^i (1-p_x)^{n-i}.$$

As an immediate consequence,

$$\mathbb{E}[\Phi_i] = \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=i}\right] = \sum_x \binom{n}{i} p_x^i (1-p_x)^{n-i}.$$

We now bound the bias and variance of an arbitrary linear estimator U^{h}. We first show that the bias $\mathbb{E}[U^{\text{h}} - U]$ under the multinomial model is close to that under the Poisson model, which is $\sum_x e^{-\lambda_x}(h(\lambda_x) - (1 - e^{-t\lambda_x}))$ as given in (5.6).

Lemma 6.1. The bias of $U^{\text{h}} = \sum_{i=1}^{\infty} \Phi_i h_i$ satisfies

$$\left|\mathbb{E}[U^{\text{h}} - U] - \sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right)\right| \le 2 \sup_i |h_i| + 2.$$

Proof. First we recall a result on Poisson approximation: For $X \sim \text{Bin}(n, p)$ and $Y \sim \text{poi}(np)$,

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \le 2p \sup_i |f(i)|, \tag{6.1}$$

which follows from the total variation bound $d_{\text{TV}}(\text{Bin}(n,p), \text{poi}(np)) \le p$ [88, Theorem 1] and the fact that $d_{\text{TV}}(\mu, \nu) = \frac{1}{2}\sup_{\|f\|_\infty \le 1} \int f d\mu - \int f d\nu$. In particular, taking $f(x) = \mathbb{1}_{x=0}$ gives

$$0 \le e^{-np} - (1-p)^n \le 2p.$$

Note that the linear estimator can be expressed as $U^{\text{h}} = \sum_x h_{N_x}$. Under the multinomial model,

$$\mathbb{E}[U^{\text{h}} - U] = \sum_x \mathbb{E}_{N_x \sim \text{Bin}(n,p_x)}[h_{N_x}] - \sum_x (1-p_x)^n (1 - (1-p_x)^m).$$

Under the Poisson model,

$$\sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right) = \sum_x \mathbb{E}_{N_x \sim \mathrm{poi}(np_x)}[h_{N_x}] - \sum_x e^{-np_x}(1 - e^{-mp_x}).$$

Then

$$\left|\sum_x \mathbb{E}_{N_x \sim \mathrm{Bin}(n,p_x)}[h_{N_x}] - \sum_x \mathbb{E}_{N_x \sim \mathrm{poi}(np_x)}[h_{N_x}]\right| \overset{(6.1)}{\leq} 2 \sup_i |h_i| \sum_x p_x = 2 \sup_i |h_i|.$$

Furthermore,

$$\begin{aligned} &\sum_x (1-p_x)^n(1-(1-p_x)^m) - \sum_x e^{-np_x}(1-e^{-mp_x}) \\ \leq &\sum_x e^{-np_x}(e^{-mp_x} - (1-p_x)^m) \overset{(6.7)}{\leq} \sum_x e^{-np_x} 2p_x \leq 2. \end{aligned}$$

Similarly, $\sum_x (1-p_x)^n(1-(1-p_x)^m) - \sum_x e^{-np_x}(1-e^{-mp_x}) \geq -2$. Assembling the above proves the lemma. □

The next result bounds the variance.

Lemma 6.2. For any linear estimator U^{h},

$$\mathrm{Var}(U^{\mathrm{h}} - U) \leq 8n \max\left\{\sup_{i \geq 1} h_i^2, 1\right\} + 8m.$$

Proof. Recognizing that $U^{\mathrm{h}} - U$ is a function of $n+m$ independent random variables, namely, $X_1, \ldots, X_{n+m}$ drawn i.i.d. from p, we apply Steele's variance inequality [89] to bound its variance. Similar to (6.1),

$$U^{\mathrm{h}} - U = \sum_x h_{N_x} + \mathbb{1}_{N_x=0}\mathbb{1}_{N_{x'}>0}$$

Changing the value of any one of the first n samples changes the multiplicities of two symbols, and hence the value of $U^{\mathrm{h}} - U$ can change by at most $4\max(\max_{i\geq 1}|h_i|, 1)$. Similarly, changing any one of the last m samples changes the value of $U^{\mathrm{h}} - U$ by at most four. Applying Steele's inequality gives the lemma. □

Lemmas 6.1 and 6.2 are analogous to Lemma 5.6. Together with (5.9) and Lemma 5.10, we obtain the main result for the multinomial model.

Theorem 6.3. For $t \geq 1$ and any random variable L over $\mathbb{Z}_+$,

$$\mathbb{E}[(U^{\text{L}} - U)^2] \leq 8n\,\mathbb{E}^2[t^L] + 8m + \left((n(t+1)\xi_L(t) + 2\mathbb{E}[t^L] + 2\right)^2.$$

Similar to Corollaries 5.12 and 5.13, one can compute the NMSE for Binomial and Poisson smoothings. We remark that up to multiplicative constants the results are identical to those for the Poisson model.

6.2 The Bernoulli-product model

Consider the following species assemblage model. There are k distinct species and each one can be found in one of n independent sampling units. Thus every species can be present in multiple sampling units simultaneously and each sampling unit can capture multiple species. For example species x can be found in sampling units $1, 3$ and 5 and species y can be found in units $2, 3$, and 4. Given the data collected from n sampling units, the objective is to estimate the expected number of new species that would be observed if we placed m more units.

The aforementioned problem is typically modeled as by the *Bernoulli-product model*. Since, in this model each sample only has presence-absence data, it is often referred to as incidence model [61]. For notational simplicity, we use the same notation as the other three models. In Bernoulli-product model, for a symbol x, N_x denotes the number of sampling units in which x appears and Φ_i denotes the number of symbols that appeared in i sampling units. Given a set of distinct symbols (potentially infinite), each symbol x is observed in each sampling unit independently with probability p_x and the observations from each sampling unit are independent of each other. To distinguish from the multinomial and Poisson sampling models where each sample can be only one symbol, we refer to samples here as sampling units. Given the results of n sampling units, the goal is to estimate the expected number of new symbols that would appear in the next m sampling units. Let $p_S = \sum_x p_x$. Note that p_S is also the expected number of symbols that we observe for each sampling unit and need not sum to 1. For example, in the species application, probability of catching bumble bee can be 0.5 and honey bee be 0.7.

This model is significantly different from the multinomial model in two ways. Firstly, here given n sampling units the number of occurrences of symbols are independent of each other. Secondly, $p_{\scriptscriptstyle S} \stackrel{\text{def}}{=} \sum_x p_x$ need not be 1. In the Bernoulli-product model, the probability observing each symbol at a particular sample is p_x and hence in n samples, the number of occurrences is distributed $\text{Bin}(n, p_x)$. Therefore the probability that x is be observed in i sampling units is

$$\mathbb{E}[\mathbb{1}_{N_x=i}] = \binom{n}{i} p_x^i (1-p_x)^{n-i},$$

and an immediate consequence on the number of distinct symbols that appear i sampling units is

$$\mathbb{E}[\Phi_i] = \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=i}\right] = \sum_x \binom{n}{i} p_x^i (1-p_x)^{n-i}.$$

Furthermore, the expected total number of symbols is $np_{\scriptscriptstyle S}$ and hence

$$\sum_{i=1}^{n} \mathbb{E}[\Phi_i] i = np_{\scriptscriptstyle S}.$$

Under the Bernoulli-product model the objective is to estimate the number of new symbols that we observe in m more sampling units and is

$$U(X^n, X_{n+1}^{n+m}) = \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N_x'>0}.$$

As before, we abbreviate

$$U \stackrel{\text{def}}{=} U(X^n, X_{n+1}^{n+m})$$

and similarly $U^{\text{E}} \stackrel{\text{def}}{=} U^{\text{E}}(X^n, t)$ for any estimator E. Since the probabilities need not add up to 1, we redefine our definition of $\mathcal{E}_{n,t}(U^{\text{E}})$ as

$$\mathcal{E}_{n,t}(U^{\text{E}}) \stackrel{\text{def}}{=} \max \mathbb{E}_p \left(\frac{U - U^{\text{E}}}{ntp_{\scriptscriptstyle S}} \right)^2.$$

Under this model, the SGT estimator satisfy similar results to that of Corollaries 5.12 and 5.13, up to multiplicative constants. The main ingredient is to bound the bias and variance (like Lemma 5.6). We note that since the marginal of N_x is $\text{Bin}(n, p_x)$ under both the multinomial and the Bernoulli-product model, the bias bound follows entirely analogously as in Lemma 6.1. The proof of variance bound is very similar to that of Lemma 5.6 and hence is omitted.

Lemma 6.4. The bias of the linear estimator U^{h} is

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right)\right| \le 2p_S\left(\sup_i |h_i| + 1\right),$$

and the variance

$$\mathrm{Var}(U^{\mathrm{h}} - U) \le np_S \cdot \left(t + \sup_{i\ge 1} h_i^2\right).$$

The above lemma together with (5.9) and Lemma 5.10 yields the main result for the Bernoulli-product model.

Theorem 6.5. For any random variable L over $\mathbb{Z}_+$ and $t \ge 1$,

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le np_S \cdot \left(t + \mathbb{E}^2[t^L]\right) + (n(t+1)p_S\xi_L(t) + 2p_S(\mathbb{E}[t^L] + 1))^2.$$

Similar to Corollaries 5.12 and 5.13, one can compute the normalized mean squared loss for Binomial and Poisson smoothings. We remark that up to multiplicative constants the results would be similar to that for the Poisson model.

6.3 The hypergeometric model

The hypergeometric model considers the population estimation problem with samples drawn without replacement. Given n samples drawn uniformly at random, without replacement from a set $\{y_1, \ldots, y_R\}$ of R symbols, the objective is to estimate the number of new symbols that would be observed if we had access to m more random samples without replacement, where $n + m \le R$. Unlike the Poisson, multinomial, and Bernoulli-product models we have considered so far, where the samples are independently and identically distributed, in the hypergeometric model the samples are *dependent* hence a modified analysis is needed.

Let $r_x \stackrel{\text{def}}{=} \sum_{i=1}^R \mathbb{1}_{y_i = x}$ be the number of occurrences of symbol x in the R symbols, which satisfies $\sum_x r_x = R$. Denote by N_x the number of times x appears in the n samples drawn without replacements, which is distributed according to the hypergeometric distribution $\mathrm{Hyp}(R, r_x, n)$ with the following probability mass function:*

$$\Pr(N_x = i) = \frac{\binom{r_x}{i}\binom{R - r_x}{n - i}}{\binom{R}{n}}.$$

*We adopt the convention that $\binom{n}{k} = 0$ for all $k < 0$ and $k > n$ throughout.

We also denote the joint distribution of $\{N_x\}$, which is multivariate hypergeometric, by $\mathrm{Hyp}(\{r_x\}, n)$. Consequently,

$$\mathbb{E}[\Phi_i] = \sum_x \Pr(N_x = i) = \sum_x \frac{\binom{r_x}{i}\binom{R-r_x}{n-i}}{\binom{R}{n}}.$$

Furthermore, conditioned on $N_x = 0$, $N_x{}'$ is distributed as $\mathrm{Hyp}(R-n, r_x, m)$ and hence

$$\mathbb{E}[U] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}] \cdot \mathbb{E}[\mathbb{1}_{N_x{}'>0} | \mathbb{1}_{N_x=0}] = \sum_x \frac{\binom{R-r_x}{n}}{\binom{R}{n}} \cdot \left(1 - \frac{\binom{R-n-r_x}{m}}{\binom{R-n}{m}}\right). \tag{6.2}$$

As before, we abbreviate

$$U \stackrel{\text{def}}{=} U(X^n, X_{n+1}^{n+m})$$

which we want to estimate and similarly for any estimator $U^{\mathrm{E}} \stackrel{\text{def}}{=} U^{\mathrm{E}}(X^n, t)$. We now bound the variance and bias of a linear estimator U^{h} under the hypergeometric model.

Lemma 6.6. For any linear estimator U^{h},

$$\mathrm{Var}(U^{\mathrm{h}} - U) \le 12n \sup_i h_i^2 + 6n + 3m.$$

Proof. We first note that for a random variable Y that lies in the interval $[a, b]$,

$$\mathrm{Var}(Y) \le \frac{(a-b)^2}{4}.$$

For notational convenience define $h_0 = 0$. Then $U^{\mathrm{h}} = \sum_x h_{N_x}$. Let $Z = \sum \mathbb{1}_{N_x=0}$ and $Z' = \sum \mathbb{1}_{N_x=N_x'=0}$ denote the number of unobserved symbols in the first n samples and the total $n+m$ samples, respectively. Then $U = Z -' Z$. Since the collection of random variables $\mathbb{1}_{N_x=0}$ indexed by x are negatively correlated, we have

$$\mathrm{Var}\big(Z\big) \le \sum_x \mathrm{Var}(\mathbb{1}_{N_x=0}\big) = \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}(1 - \mathbb{1}_{N_x=0})] \le \sum_x \mathbb{E}\left[\mathbb{1}_{N_x>0}\right] \le n.$$

Analogously, $\mathrm{Var}(Z') \le n+m$ and hence

$$\begin{aligned}
\mathrm{Var}(U^{\mathrm{h}} - U) &= \mathrm{Var}(U^{\mathrm{h}} - Z + Z') \\
&\le 3\mathrm{Var}(U^{\mathrm{h}}) + 3\mathrm{Var}(Z') + 3\mathrm{Var}(Z) \\
&\le 3\mathrm{Var}(U^{\mathrm{h}}) + 6n + 3m.
\end{aligned}$$

Thus it remains to show

$$\mathrm{Var}(U^{\mathrm{h}}) \le 4n \sup_i h_i^2. \tag{6.3}$$

By induction on n, we show that for any $n \in \mathbb{N}$, any set of nonnegative integers $\{r_x\}$ and any function $(x,k) \mapsto f(x,k)$ with $k \in \mathbb{Z}_+$ satisfying $f(x,0) = 0$,

$$\mathrm{Var}\left(\sum_x f(x, N_x)\right) \le 4n\|f\|_\infty^2, \tag{6.4}$$

where $\{N_x\} \sim \mathrm{Hyp}(\{r_x\}, n)$ and $\|f\|_\infty = \sup_{x,k} |f(x,k)|$. Then the desired Equation (6.3) follows from (6.4) with $f(x,k) = h_k$.

We first prove (6.4) for $n = 1$, in which case exactly one of N_x's is one and the rest are zero. Hence, $|\sum_x f(x, N_x)| \le \|f\|_\infty$ and $\mathrm{Var}(\sum_x f(x, N_x)) \le \|f\|_\infty^2$.

Next assume the induction hypothesis holds for $n-1$. Let X_1 denote the first sample and let $\tilde{N}_x$ denote the number of occurrences of symbol x in samples $X_2, \ldots, X_n$. Then $N_x = \tilde{N}_x + \mathbb{1}_{X_1 = x}$. Furthermore, conditioned on $X_1 = y$, $\{\tilde{N}_x\} \sim \mathrm{Hyp}(\{\tilde{r}_x\}, n-1)$, where $\tilde{r}_x = r_x - \mathbb{1}_{x=y}$. By the law of total variance, we have

$$\mathrm{Var}\left(\sum_x f(x, N_x)\right) = \mathbb{E}\left[V(X_1)\right] + \mathrm{Var}\left(g(X_1)\right). \tag{6.5}$$

where

$$V(y) \stackrel{\mathrm{def}}{=} \mathrm{Var}\left(\sum_x f(x, N_x) \middle| X_1 = y\right), \quad g(y) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\sum_x f(x, N_x) \middle| X_1 = y\right]$$

For the first term in (6.5), note that

$$V(y) = \mathrm{Var}\left(\sum_x f(x, \tilde{N}_x + \mathbb{1}_{x=y}) \middle| X_1 = y\right) = \mathrm{Var}\left(\sum_x f_y(x, \tilde{N}_x) \middle| X_1 = y\right).$$

where we defined $f_y(x,k) \stackrel{\mathrm{def}}{=} f(x, k + \mathbb{1}_{x=y})$. Hence, by the induction hypothesis, $V(y) \le 4(n-1)\|f_y\|_\infty^2 \le 4(n-1)\|f\|_\infty^2$ and $\mathbb{E}\left[V(X_1)\right] \le 4(n-1)||f||_\infty^2$.

For the second term in (6.5), observe that for any $y \neq z$

$$g(y) = \mathbb{E}[f(y, \tilde{N}_y + 1) | X_1 = y] + \mathbb{E}[f(z, \tilde{N}_z) | X_1 = y] + \mathbb{E}\left[\sum_{x \neq y,z} f(x, \tilde{N}_x) \middle| X_1 = y\right],$$

and

$$g(z) = \mathbb{E}[f(z, \tilde{N}_z + 1)|X_1 = z] + \mathbb{E}[f(y, \tilde{N}_y)|X_1 = z] + \mathbb{E}\left[\sum_{x \neq y,z} f(x, \tilde{N}_x)\middle| X_1 = z\right],$$

Observe that $\{N_x\}_{x\neq y,z}$ have the same joint distribution conditioned on either $X_1 = y$ or $X_1 = z$ and hence $\mathbb{E}[\sum_{x\neq y,z} f(x, \tilde{N}_x)|X_1 = y] = \mathbb{E}[\sum_{x\neq y,z} f(x, \tilde{N}_x)|X_1 = z]$. Therefore $|g(y) - g(z)| \leq 4\|f\|_\infty$ for any $y \neq z$. This implies that the function g takes values in an interval of length at most $4\|f\|_\infty$. Therefore $\mathrm{Var}(g(X_1)) \leq \frac{1}{4}(4\|f\|_\infty)^2 = 4\|f\|_\infty^2$. This completes the proof of (6.4) and hence the lemma. □

Let

$$B(h, r_x) \stackrel{\text{def}}{=} \sum_{i=1}^{r_x} \binom{r_x}{i} \left(\frac{n}{R}\right)^i \left(1 - \frac{n}{R}\right)^{r_x - i} h_i - \left(1 - \frac{n}{R}\right)^{r_x} \left(1 - \left(1 - \frac{m}{R-n}\right)^{r_x}\right).$$

To bound the bias, we first prove an auxiliary result.

Lemma 6.7. For any linear estimator U^{h},

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x B(h, r_x)\right| \leq 4 \max\left(\sup_i |h_i|, 1\right) + \frac{2R}{R-n}.$$

Proof. Recall that $N_x \sim \mathrm{Hyp}(R, r_x, n)$. Let $\tilde{N}_x$ be a random variable distributed as $\mathrm{Bin}(r_x, n/R)$. Since $\mathrm{Hyp}(R, r_x, n)$ coincides with $\mathrm{Hyp}(R, n, r_x)$, we have

$$d_{\mathrm{TV}}(\mathrm{Bin}(r_x, n/R), \mathrm{Hyp}(R, r_x, n)) = d_{\mathrm{TV}}(\mathrm{Bin}(r_x, n/R), \mathrm{Hyp}(R, n, r_x)) \leq \frac{2r_x}{R},$$

where the last inequality follows from [90, Theorem 4]. Since

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sup_{\|f\|_\infty \leq 1} \int f d\mu - \int f d\nu = \sup_E \mu(E) - \nu(E),$$

we have

$$\left|\mathbb{E}[f(N_x)] - \mathbb{E}[f(\tilde{N}_x)]\right| \leq \frac{4r_x}{R} \sup_i |f(i)|, \tag{6.6}$$

and

$$\begin{aligned}\left|\frac{\binom{R-n-r_x}{m}}{\binom{R-n}{m}} - \left(1 - \frac{m}{R-n}\right)^{r_x}\right| &\leq d_{\mathrm{TV}}(\mathrm{Bin}(r_x, m/(R-n)), \mathrm{Hyp}(R-n, m, r_x)) \\ &\leq \frac{2r_x}{R-n}. \end{aligned} \tag{6.7}$$

Define $f_x(i) = h_i - \mathbb{1}_{i=0}\left(1 - \left(1 - \frac{m}{R-n}\right)^{r_x}\right)$. In view of (6.2) and the fact that $\sum r_x = R$, we have

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x \mathbb{E}[f_x(N_x)]\right| \le \frac{2R}{R-n}.$$

Applying (6.6) yields

$$\sum_x \left|\mathbb{E}[f_x(\tilde{N}_x)] - \mathbb{E}\left[f_x(N_x)\right]\right| \le 4 \sup_i |f_x(i)| \le 4 \max\left(\sup_i |h_i|, 1\right).$$

The above equation together with (6.7) results in the lemma since $B(h, r_x) = \mathbb{E}[f_x(\tilde{N}_x)]$. □

Note that to upper bound the bias, we need to bound $\sum_x B(h, r_x)$. It is easy to verify for the GT coefficients $h_i^{\mathrm{GT}} = -(-t)^i$ with $t = m/n$, $B(h^{\mathrm{GT}}, r_x) = 0$. Therefore, if we choose $h = h^{\mathrm{L}}$ based on the tail of random variable L with $h_i^{\mathrm{L}} = h_i^{\mathrm{GT}} \Pr(L \ge i)$ as defined in (5.8), we have

$$\begin{aligned} B(h^{\mathrm{L}}, r_x) &= \sum_{i=1}^{r_x} \binom{r_x}{i} \left(\frac{n}{R}\right)^i \left(1 - \frac{n}{R}\right)^{r_x - i} (-t)^i \Pr(L < i) \\ &= \left(1 - \frac{n}{R}\right)^{r_x} \sum_{i=1}^{r_x} \binom{r_x}{i} \left(-\frac{m}{R-n}\right)^i \Pr(L < i). \end{aligned} \tag{6.8}$$

Similar to Lemma 5.9, our strategy is to find an integral presentation of the bias. This is done in the following lemma.

Lemma 6.8. For any $y \ge 0$ and any $k \in \mathbb{N}$,

$$\sum_{i=1}^{k} \binom{k}{i} (-y)^i \Pr(L < i) = -k(1-y)^k \int_0^y \mathbb{E}\left[\binom{k-1}{L}(-s)^L\right](1-s)^{-k-1} ds. \tag{6.9}$$

Remark 6.9. For the special case of $y = 1$, (6.9) is understood in the limiting sense: Letting $\delta = 1 - y$ and $\beta = \frac{1-s}{\delta}$, we can rewrite the right-hand side as

$$-k \int_1^{1/\delta} \mathbb{E}\left[\binom{k-1}{L}(\beta\delta - 1)^L\right] k\beta^{-k-1} d\beta.$$

For all $|\delta| \le 1$ and hence $0 \le 1 - \beta\delta \le 2$, we have

$$\left|\mathbb{E}\left[\binom{k-1}{L}(\beta\delta - 1)^L\right]\right| = \left|\mathbb{E}\left[\binom{k-1}{L}(\beta\delta - 1)^L \mathbb{1}_{L<k}\right]\right| \le 4^k.$$

By dominated convergence theorem, as $\delta \to 0$, the right-hand side converges to $-\mathbb{E}\left[\binom{k-1}{L}(-1)^L\right]$ and coincides with the left-hand side, which can be easily obtained by applying $\binom{k}{i} = \binom{k-1}{i} + \binom{k-1}{i-1}$.

Proof. Denote the left-hand side of (6.9) by $F(y)$. Using $i\binom{k}{i} = k\binom{k-1}{i-1}$, we have

$$\begin{aligned} & F'(y) \\ = & \sum_{i=1}^{k} \binom{k}{i}(-i)(-y)^{i-1} \Pr(L < i) = -k \sum_{i=1}^{k} \binom{k-1}{i-1}(-y)^{i-1} \Pr(L < i) \\ = & -k \sum_{i=1}^{k} \binom{k-1}{i-1}(-y)^{i-1} \Pr(L < i-1) - k \sum_{i=1}^{k} \binom{k-1}{i-1}(-y)^{i-1} \Pr(L = i-1). \end{aligned} \tag{6.10}$$

The second term is simply $-k\mathbb{E}\left[\binom{k-1}{L}(-y)^L\right] \stackrel{\text{def}}{=} G(y)$. For the first term, since $L \geq 0$ almost surely and $\binom{k}{i} = \binom{k-1}{i} + \binom{k-1}{i-1}$, we have

$$\begin{aligned} & k \sum_{i=1}^{k} \binom{k-1}{i-1}(-y)^{i-1} \Pr(L < i-1) \\ = & k \sum_{i=1}^{k} \binom{k-1}{i}(-y)^{i} \Pr(L < i) \\ = & k \sum_{i=1}^{k} \binom{k}{i}(-y)^{i} \Pr(L < i) - k \sum_{i=1}^{k} \binom{k-1}{i-1}(-y)^{i} \Pr(L < i) \\ = & kF(y) - yF'(y). \end{aligned} \tag{6.11}$$

Combining (6.10) and (6.11) yields the following ordinary differential equation:

$$F'(y)(1-y) + kF(y) = G(y), \quad F(0) = 0,$$

whose solution is readily obtained as $F(y) = (1-y)^k \int_0^y (1-s)^{-k-1} G(s) ds$, *i.e.*, the desired Equation (6.9). □

Combining Lemma 6.7–6.8 yields the following bias bound:

Lemma 6.10. For any random variable L over $\mathbb{Z}_+$ and $t = m/n \geq 1$,

$$|\mathbb{E}[U^{\mathrm{L}} - U]| \leq nt \cdot \max_{0 \leq s \leq 1} \left| \mathbb{E}\left[\binom{r_x - 1}{L}(-s)^L\right] \right| + 4\mathbb{E}[t^L] + \frac{2R}{R-n}.$$

Proof. Recall the coefficient bound (5.9) that $\sup_i |h_i| \le \mathbb{E}[t^L]$. By Lemma 6.7 and the assumption that $t \ge 1$,

$$\left| \mathbb{E}[U^{\mathrm{h}} - U] - \sum_x B(h^{\mathrm{L}}, r_x) \right| \le 4\mathbb{E}[t^L] + \frac{2R}{R-n}.$$

Thus it suffices to bound $\sum_x B(h^{\mathrm{L}}, r_x)$. For every x, using (6.8) and applying Lemma 6.8 with $y = \frac{m}{R-n}$ and $k = r_x$, we obtain

$$B(h^{\mathrm{L}}, r_x) = -\left(1 - \frac{n+m}{R}\right)^{r_x} \int_0^{\frac{m}{R-n}} \mathbb{E}\left[\binom{r_x - 1}{L}(-s)^L\right] r_x (1-s)^{-r_x-1} ds.$$

Since $0 \le \frac{m}{R-n} \le 1$, letting $K = \max_{0 \le s \le 1} \left|\mathbb{E}\left[\binom{r_x-1}{L}(-s)^L\right]\right|$, we have

$$\begin{aligned} |B(h^{\mathrm{L}}, r_x)| &\le \left(1 - \frac{n+m}{R}\right)^{r_x} K \int_0^{\frac{m}{R-n}} r_x (1-s)^{-r_x-1} ds. \\ &= K\left(\left(1 - \frac{n}{R}\right)^{r_x} - \left(1 - \frac{n+m}{R}\right)^{r_x}\right) \le K\left(1 - \frac{n}{R}\right)^{r_x - 1} \frac{m r_x}{R}, \end{aligned}$$

where the last inequality follows from the convexity of $x \mapsto (1-x)^{r_x}$. Summing over all symbols x results in the lemma. □

Combining Lemma 6.10 and Lemma 6.6 gives the following NMSE bound:

Theorem 6.11. Under the assumption of Lemma 6.10, $\mathbb{E}[(U^{\mathrm{L}} - U)^2]$ is at most

$$12(n+1)\mathbb{E}^2[t^L] + 6n + 3m + \frac{12R^2}{(R-n)^2} + 3m^2 \max_{1 \ge \alpha > 0} \left| \mathbb{E}\left[\binom{r_x - 1}{L}(-\alpha)^L\right] \right|^2.$$

As before, we can choose various smoothing distribution and obtain upper bounds on the mean squared error.

Corollary 6.12. If $L \sim \mathrm{poi}(r)$ and $R - n \ge m \ge n$, then

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le 12(n+1)e^{2r(t-1)} + 3m^2 e^{-r} + 9m + 48.$$

Furthermore, if $r = \frac{1}{2t-1} \cdot \log(nt^2)$,

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le \frac{27}{(nt^2)^{\frac{1}{2t-1}}} + \frac{9nt + 48}{(nt)^2}.$$

Proof. For $L \sim \mathrm{poi}(r)$, $\mathbb{E}[t^L] = e^{r(t-1)}$ and

$$\max_{0\le\alpha\le 1}\left|\mathbb{E}\left[\binom{r_x-1}{L}(-\alpha)^L\right]\right| = e^{-r}\max_{0\le\alpha\le 1}|L_{r_x-1}(\alpha r)| \le e^{-r/2},$$

where L_{r_x-1} is the Laguerre polynomial of degree r_x-1 defined in (5.13) and the last equality follows the bound (5.14). Furthermore, $R/(R-n) = 1+n/(R-n) \le 1+n/m \le 2$ and $n \le m$, and hence the first part of the lemma. The second part follows by substituting the value of r. □

6.4 Lower bounds

Under the multinomial model (i.i.d. sampling), we lower bound the risk $\mathcal{E}_{n,t}(U^{\mathrm{E}})$ for any estimator U^{E} using the support size estimation lower bound in [82]. Since the lower bound in [82] also holds for the Poisson model, so does our lower bound.

Recall that for a discrete distribution p, $\mathrm{Supp}(p) = \sum_x \mathbb{1}_{p_x>0}$ denotes its support size. It is shown that given n i.i.d. samples drawn from a distribution p whose minimum non-zero mass $p^+_{\min}$ is at least $1/k$, the minimax mean-square error for estimating $\mathrm{Supp}(p)$ satisfies

$$\min_{\hat{\mathrm{Supp}}}\max_{p:p^+_{\min}\ge 1/k}\mathbb{E}[(\hat{\mathrm{Supp}}-\mathrm{Supp}(p))^2] \ge c'k^2\cdot\exp\left(-c\max\left(\sqrt{\frac{n\log k}{k}},\frac{n}{k}\right)\right). \tag{6.12}$$

where c, c' are universal positive constants with $c > 1$. We prove Theorem 5.2 under the multinomial model with c being the universal constant from (6.12).

Suppose we have an estimator $\hat{U}$ for U that can accurately predict the number of new symbols arising in the next m samples, we can then produce an estimator for the support size by adding the number of symbols observed, Φ_+, in the current n samples, namely,

$$\hat{\mathrm{Supp}} = \hat{U} + \Phi_+. \tag{6.13}$$

Note that $U = \sum_x \mathbb{1}_{N_x=0}\mathbb{1}_{N'_x>0}$. When $m = \infty$, U is the total number of unseen symbols and we have $\mathrm{Supp}(p) = U + \Phi_+$. Consequently, if $\hat{U}$ can foresee too far

into the future (*i.e.,* for too large an m), then (6.13) will constitute a support size estimator that is too good to be true.

Combining Theorem 5.2 with the positive result (Corollary 5.12 or 5.13) yields the following characterization of the minimax risk:

Corollary 6.13. For all $t \geq c$, we have

$$\inf_{U^{\mathrm{E}}} \mathcal{E}_{n,t}(U^{\mathrm{E}}) = \exp\left(-\Theta\left(\max\left\{\frac{\log n}{t}, 1\right\}\right)\right)$$

Consequently, as $n \to \infty$, the minimax risk $\inf_{U^{\mathrm{E}}} \mathcal{E}_{n,t}(U^{\mathrm{E}}) \to 0$ if and only if $t = o(\log n)$.

Proof of Theorem 5.2. Recall that $m = nt$. Let $\hat{U}$ be an arbitrary estimator for U. For the support size estimator $\hat{\mathrm{Supp}} = \hat{U} + \Phi_+$ defined in (6.13), it must obey the lower bound (6.12). Hence there exists some p satisfying $p^+_{\min} \geq 1/k$, such that

$$\mathbb{E}[(\mathrm{Supp}(p) - \hat{\mathrm{Supp}})^2] \geq c' k^2 \cdot \exp\left(-c \max\left(\sqrt{\frac{n \log k}{k}}, \frac{n}{k}\right)\right). \tag{6.14}$$

Let $S = \mathrm{Supp}(p)$ denote the support size, which is at most k. Let $\tilde{U} \stackrel{\mathrm{def}}{=} \mathbb{E}_{X_{n+1}^{n+m}}[U]$ be the expectation of U over the unseen samples X_{n+1}^{n+m} conditioned on the available samples X_1^n. Then $\tilde{U} = \sum_x \mathbb{1}_{N_x=0}\left(1 - (1-p_x)^{nt}\right)$. Since the estimator $\hat{U}$ is independent of X_{n+1}^{n+m}, by convexity,

$$\mathbb{E}_{X_1^{n+m}}[(U - \hat{U})^2] \geq \mathbb{E}_{X_1^n}[(\mathbb{E}_{X_{n+1}^{n+m}}[U - \hat{U}])^2] = \mathbb{E}[(\tilde{U} - \hat{U})^2]. \tag{6.15}$$

Notice that with probability one,

$$|S - \tilde{U} - \Phi_+| \leq S e^{-nt/k} \leq k e^{-nt/k}, \tag{6.16}$$

which follows from

$$\tilde{U} + \Phi_+ = \sum_{x: p_x > 0} \mathbb{1}_{N_x=0}\left(1 - (1-p_x)^{nt}\right) + \mathbb{1}_{N_x>0} \leq S,$$

and, on the other hand,

$$\begin{aligned}
&\tilde{U} + \Phi_+ \\
&= \sum_{x: p_x \geq 1/k} \mathbb{1}_{N_x=0}\left(1 - (1-p_x)^{nt}\right) + \mathbb{1}_{N_x>0} \\
&\geq \sum_x \mathbb{1}_{N_x=0}\left(1 - (1-1/k)^{nt}\right) + \mathbb{1}_{N_x>0} \geq S(1 - (1-1/k)^{nt}) \geq S(1 - e^{-nt/k}).
\end{aligned}$$

Expanding the left hand side of (6.14),

$$\begin{aligned}
&\mathbb{E}[(S - \hat{\mathrm{Supp}})^2] \\
&= \mathbb{E}\left[\left(S - \tilde{U} - \Phi_+ + \tilde{U} - \hat{U}\right)^2\right] \leq 2\mathbb{E}[(S - \tilde{U} - \Phi_+)^2] + 2\mathbb{E}[(\tilde{U} - \hat{U}))^2] \\
&\overset{(6.16)}{\leq} 2k^2 e^{-2nt/k} + 2\mathbb{E}[(\tilde{U} - \hat{U}))^2] \overset{(6.15)}{\leq} 2k^2 e^{-2nt/k} + 2\mathbb{E}[(U - \hat{U}))^2]
\end{aligned}$$

Let

$$k = \min\left\{\frac{nt^2}{c^2 \log \frac{nt^2}{c^2}}, \frac{nt}{\log \frac{4}{c'}}\right\},$$

which ensures that

$$c'k^2 \cdot \exp\left(-c \max\left\{\sqrt{\frac{n \log k}{k}}, \frac{n}{k}\right\}\right) \geq 4k^2 e^{-2nt/k}. \tag{6.17}$$

Then

$$\mathbb{E}[(U - \hat{U})^2] \geq k^2 e^{-2nt/k},$$

establishes the following lower bound with $\alpha \overset{\text{def}}{=} \frac{c'^2}{4 \log^2(4/c')}$ and $\beta \overset{\text{def}}{=} c^2$:

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \geq \min\left\{\alpha, \frac{4t^2}{\beta^2 \log^2 \frac{nt^2}{\beta}} \left(\frac{\beta}{nt^2}\right)^{2\beta/t}\right\}.$$

To verify (6.17), since $t \geq c$ by assumption, we have $\exp(\frac{2tn}{k} - \frac{cn}{k}) \geq \exp(\frac{nt}{k}) \geq \frac{4}{c'}$. Similarly, since $k \log k \leq \frac{nt^2}{c^2}$ by definition, we have $\frac{2nt}{k} \geq 2c'\sqrt{\frac{n \log k}{k}}$ and hence $\exp\left(\frac{2tn}{k} - c\sqrt{\frac{n \log k}{k}}\right) \geq \exp(\frac{nt}{k}) \geq \frac{4}{c'}$, completing the proof of (6.17).

Thus we have shown that there exist universal positive constants α, β such that

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \geq \min\left\{\alpha, \frac{4t^2}{\beta^2 \log^2 \frac{nt^2}{\beta}} \left(\frac{\beta}{nt^2}\right)^{2\beta/t}\right\}.$$

Let $y = \left(\frac{nt^2}{\beta}\right)^{2\beta/t}$, then

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \geq \min\left\{\alpha, 16\frac{1}{y \log^2 y}\right\}.$$

Since $y > 1$, $y^3 \geq y \log^2 y$ and hence for some constants $c_1, c_2 > 0$,

$$\begin{aligned}\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \geq \min\left\{\alpha, 16\frac{1}{y^3}\right\} &\geq \min\left\{\alpha, \left(\frac{\beta}{nt^2}\right)^{6\beta/t}\right\} \\ &\geq c_1 \min\left\{1, \left(\frac{1}{n}\right)^{c_2/t}\right\} \\ &\geq \frac{c_1}{n^{c_2/t}}.\end{aligned}$$

□

Acknowledgement

Chapter 6 is adapted from Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu, "Estimating the number of unseen species: A bird in the hand is worth $\log n$ in the bush", Manuscript, 2015 [87].

Part III

How efficiently can one learn in high dimensions?

Chapter 7

Learning Gaussian mixtures

7.1 Introduction

7.1.1 Background

Meaningful information often resides in high-dimensional spaces: voice signals are expressed in many frequency bands, credit ratings are influenced by multiple parameters, and document topics are manifested in the prevalence of numerous words. Some applications, such as topic modeling and genomic analysis consider data in over 1000 dimensions, [4, 5].

Typically, information can be generated by different types of sources: voice is spoken by men or women, credit parameters correspond to wealthy or poor individuals, and documents address topics such as sports or politics. In such cases the overall data follow a mixture distribution [6, 7, 8].

Mixtures of high-dimensional distributions are therefore central to the understanding and processing of many natural phenomena. Methods for recovering the mixture components from the data have consequently been extensively studied by statisticians, engineers, and computer scientists.

Initially, heuristic methods such as expectation-maximization were developed [91, 92]. Over the past decade, rigorous algorithms were derived to recover mixtures of d-dimensional spherical Gaussians [93, 94, 95, 96, 97] and general Gaussians [98, 99, 100, 101, 102, 103]. Many of these algorithms consider mix-

tures where the ℓ_1 distance between the mixture components is $2 - o_d(1)$, namely approaches the maximum of 2 as d increases. They identify the distribution components in time and samples that grow polynomially in d. Recently, [100, 101, 102] showed that the parameters of any k-component d-dimensional Gaussian mixture can be recovered in time and samples that grow as a high-degree polynomial in d and exponentially in k.

A different approach that avoids the large component-distance requirement and the high time and sample complexity, considers a slightly relaxed notion of approximation, sometimes called *PAC learning* [104], or *proper learning*, that does not approximate each mixture component, but instead derives a mixture distribution that is close to the original one. Specifically, given a distance bound $\epsilon > 0$, error probability $\delta > 0$, and samples from the underlying mixture $\mathbf{f}$, where we use boldface letters for d-dimensional objects, PAC learning seeks a mixture estimate $\hat{\mathbf{f}}$ with at most k components such that $D(\mathbf{f}, \hat{\mathbf{f}}) \leq \epsilon$ with probability $\geq 1 - \delta$, where $D(\cdot, \cdot)$ is some given distance measure, for example ℓ_1 distance or KL divergence.

An important and extensively studied special case of Gaussian mixtures is mixture of *spherical-Gaussians* [93, 94, 95, 96, 97], where for each component the d coordinates are distributed independently with the same variance, though possibly with different means. Note that different components can have different variances. Due to their simple structure, spherical-Gaussian mixtures are easier to analyze and under a minimum-separation assumption have provably-practical algorithms for clustering and parameter estimation. We consider spherical-Gaussian mixtures as they are important on their own and form a natural first step towards learning general Gaussian mixtures.

7.1.2 Sample complexity

Reducing the number of samples required for learning is of great practical significance. For example, in topic modeling every sample is a whole document, in credit analysis every sample is a person's credit history, and in genetics, every sample is a human DNA. Hence samples can be very scarce and obtaining them can be very costly. By contrast, current CPUs run at several Giga Hertz, hence

samples are typically much more scarce of a resource than time.

For one-dimensional distributions, the need for sample-efficient algorithms has been broadly recognized. The sample complexity of many problems is known quite accurately, often to within a constant factor. For example, for discrete distributions over $\{1,\ldots,s\}$, an approach was proposed in [105] and its modifications were used in [76] to estimate the probability multiset using $\Theta(s/\log s)$ samples. Learning one-dimensional m-modal distributions over $\{1,\ldots,s\}$ requires $\Theta(m\log(s/m)/\epsilon^3)$ samples [106]. Similarly, one-dimensional mixtures of k structured distributions (logconcave, monotone hazard rate, and unimodal) over $\{1,\ldots,s\}$ can be learned with $\mathcal{O}(k/\epsilon^4)$, $\mathcal{O}(k\log(s/\epsilon)/\epsilon^4)$, and $\mathcal{O}(k\log(s)/\epsilon^4)$ samples, respectively, and these bounds are tight up to a factor of ϵ [107].

Unlike the one-dimensional case, in high dimensions, sample complexity bounds are quite weak. For example, to learn a mixture of $k=2$ spherical Gaussians, existing estimators use $\mathcal{O}(d^{12})$ samples, and this number increases exponentially with k [108]. We close this gap by constructing estimators with near-linear sample complexity.

7.1.3 Previous and new results

Our main contribution is PAC learning d-dimensional spherical Gaussian mixtures with near-linear samples. In the process of deriving these results we also prove results for learning one-dimensional Gaussians.

d-dimensional Gaussian mixtures

Several papers considered PAC learning of discrete- and Gaussian-product mixtures. [109] considered mixtures of two d-dimensional Bernoulli products where all probabilities are bounded away from 0. They showed that this class of mixtures is PAC learnable in $\widetilde{\mathcal{O}}(d^2/\epsilon^4)$ time and samples, where the $\widetilde{\mathcal{O}}$ notation hides logarithmic factors. [110] eliminated the probability constraints and generalized the results from binary to arbitrary discrete alphabets and from 2 to k mixture components, showing that these mixtures are PAC learnable in $\widetilde{\mathcal{O}}\big((d/\epsilon)^{2k^2(k+1)}\big)$ time. Although they did not explicitly mention sample complexity, their algo-

rithm uses $\widetilde{\mathcal{O}}\big((d/\epsilon)^{4(k+1)}\big)$ samples. [108] generalized these results to Gaussian products and showed that mixtures of k Gaussians, where the difference between the means is bounded by B times the standard deviation, are PAC learnable in $\widetilde{\mathcal{O}}\big((dB/\epsilon)^{2k^2(k+1)}\big)$ time, and can be shown to use $\widetilde{\mathcal{O}}\big((dB/\epsilon)^{4(k+1)}\big)$ samples. These algorithms consider the KL divergence between the distribution and its estimate, but it can be shown that the ℓ_1 distance would result in similar complexities. It can also be shown that these algorithms or their simple modifications have similar time and sample complexities for spherical Gaussians as well.

Our main contribution for this problem is to provide an algorithm that PAC learns mixtures of spherical-Gaussians in ℓ_1 distance with number of samples nearly-linear, and running time polynomial in the dimension d. Specifically, in Theorem 7.12 we show that mixtures of k spherical-Gaussian distributions can be learned using

$$n = \mathcal{O}\left(\frac{dk^9}{\epsilon^4}\log^2\frac{d}{\delta}\right) = \mathcal{O}_{k,\epsilon}\left(d\log^2\frac{d}{\delta}\right)$$

samples and in time

$$\mathcal{O}\Big(n^2 d\log n + d\Big(\frac{k^7}{\epsilon^3}\log^2\frac{d}{\delta}\Big)^{\frac{k^2}{2}}\Big) = \widetilde{\mathcal{O}}_{k,\epsilon}(d^3).$$

Recall that for similar problems, previous algorithms used $\widetilde{\mathcal{O}}\big((d/\epsilon)^{4(k+1)}\big)$ samples. Furthermore, recent algorithms typically construct the covariance matrix [97, 108], hence require $\geq nd^2$ time. In that sense, for small k, the time complexity we derive is comparable to the best such algorithms one can hope for. Additionally, the exponential dependence on k in the time complexity is $d(\frac{k^7}{\epsilon^3}\log^2\frac{d}{\delta})^{k^2/2}$, significantly lower than the $d^{\mathcal{O}(k^3)}$ dependence in previous results.

Conversely, Theorem 7.2 shows that any algorithm for PAC learning a mixture of k spherical Gaussians requires $\Omega(dk/\epsilon^2)$ samples, hence our algorithms are nearly sample optimal in the dimension. In addition, their time complexity significantly improves on previously known ones.

One-dimensional Gaussian mixtures

We also construct a simple estimator that learns k-component one-dimensional Gaussian mixtures using $\widetilde{\mathcal{O}}(k\epsilon^{-2})$ samples and in $\widetilde{\mathcal{O}}((k/\epsilon)^{3k+1})$

time. We note that independently and concurrently with this work [111] showed that mixtures of two one-dimensional Gaussians can be learnt with $\widetilde{\mathcal{O}}(\epsilon^{-2})$ samples and in time $\mathcal{O}(\epsilon^{-5})$. Combining with some of the techniques in this thesis, they extend their algorithm to mixtures of k Gaussians, and reduce the exponent to $3k-1$.

7.1.4 The approach and technical contributions

Let $d(\mathbf{f}, \mathcal{F})$ be the smallest ℓ_1 distance between a distribution $\mathbf{f}$ and any distribution in a collection $\mathcal{F}$. The popular SCHEFFE estimator [112] takes a surprisingly small $\mathcal{O}(\log|\mathcal{F}|)$ independent samples from an unknown distribution $\mathbf{f}$ and time $\mathcal{O}(|\mathcal{F}|^2)$ to find a distribution in $\mathcal{F}$ whose distance from $\mathbf{f}$ is at most a constant factor larger than $d(\mathbf{f}, \mathcal{F})$. Recently [113] modified the algorithm to lower the time complexity of the Scheffe algorithm from $\mathcal{O}(|\mathcal{F}|^2)$ time to $\mathcal{O}(|\mathcal{F}|)$. We use this modified version thus reducing the time complexity of our algorithms.

Given the above, our goal is to construct a small class of distributions such that one of them is ϵ-close to the underlying distribution.

Consider for example mixtures of k components in one dimension with means and variances bounded by B. Take the collection of all mixtures derived by quantizing the means and variances of all components to ϵ_m accuracy, and quantizing the weights to ϵ_w accuracy. It can be shown that if $\epsilon_m, \epsilon_w \leq \epsilon/k^2$ then one of these candidate mixtures would be $\mathcal{O}(\epsilon)$-close to any mixture, and hence to the underlying one. There are at most $(B/\epsilon_m)^{2k} \cdot (1/\epsilon_w)^k = (B/\epsilon)^{\widetilde{\mathcal{O}}(k)}$ candidates and running SCHEFFE on these mixtures would lead to an estimate. However, this approach requires a bound on the means and variances. We remove this requirement on the bound, by selecting the quantizations based on samples and we describe it in Section 7.3.

In d dimensions, consider spherical Gaussians with the same variance and means bounded by B. Again, take the collection of all distributions derived by quantizing the means of all components in all coordinates to ϵ_m accuracy, and quantizing the weights to ϵ_w accuracy. It can be shown that for d-dimensional Gaussian to get distance ϵ from the underlying distribution, it suffices to take

$\epsilon_m, \epsilon_w \leq \epsilon^2/\text{poly}(dk)$. There are at most $(B/\epsilon_m)^{dk} \cdot (1/\epsilon_w)^k = 2^{\widetilde{\mathcal{O}}_\epsilon(dk)}$ possible combinations of the k mean vectors and weights. Hence SCHEFFE implies an exponential-time algorithm with sample complexity $\widetilde{\mathcal{O}}(dk)$. To reduce the dependence on d, one can approximate the span of the k mean vectors. This reduces the problem from d to k dimensions, allowing us to consider a distribution collection of size $2^{\mathcal{O}(k^2)}$, with SCHEFFE sample complexity of just $\mathcal{O}(k^2)$. [110, 108] constructs the sample correlation matrix and uses k of its columns to approximate the span of mean vectors. This approach requires the k columns of the sample correlation matrix to be very close to the actual correlation matrix, requiring a lot more samples.

We derive a spectral algorithm that approximates the span of the k mean vectors using the top k eigenvectors of the sample covariance matrix. Since we use the entire covariance matrix instead of just k columns, a weaker concentration suffices and the sample complexity can be reduced.

Using recent tools from non-asymptotic random matrix theory [114, 115, 116] we show that the span of the means can be approximated with $\widetilde{\mathcal{O}}(d)$ samples. This result allows us to address most "reasonable" distributions, but still there are some "corner cases" that need to be analyzed separately. To address them, we modify some known clustering algorithms such as single-linkage, and spectral projections. While the basic algorithms were known before, our contribution here, which takes a fair bit of effort and space, is to show that judicious modifications of the algorithms and rigorous statistical analysis yield polynomial time algorithms with near-linear sample complexity. We provide a simple and practical spectral algorithm that estimates all such mixtures in $\mathcal{O}_{k,\epsilon}(d \log^2 d)$ samples.

The rest of the chapter is organized as follows. In Section 7.2, we introduce notations and state a lower bound. In Section 7.3 we show a simple learning algorithm for one-dimensional Gaussian mixtures. In Section 7.4, we motivate and present the algorithm for d-dimensional Gaussian mixtures. We then provide guarantees for the proposed algorithm in Sections 7.5, 7.6, 7.7. Finally, we prove lower bounds on the sample complexity in Section 7.8.

7.2 Preliminaries

7.2.1 Notation

For arbitrary product distributions $\mathbf{p}_1, \ldots, \mathbf{p}_k$ over a d dimensional space let $p_{j,i}$ be the distribution of $\mathbf{p}_j$ over coordinate i, and let $\hat{\boldsymbol{\mu}}_{j,i}$ and $\sigma_{j,i}$ be the mean and variance of $p_{j,i}$ respectively. Let $\mathbf{f} = (w_1, \ldots, w_k, \mathbf{p}_1, \ldots, \mathbf{p}_k)$ be the mixture of these distributions with mixing weights $w_1, \ldots, w_k$. We denote estimates of a quantity $\mathbf{x}$ by $\hat{\mathbf{x}}$. It can be empirical mean or a more complex estimate. $||\cdot||$ denotes the spectral norm of a matrix and $||\cdot||_2$ is the ℓ_2 norm of a vector. We use $D(\cdot, \cdot)$ to denote the ℓ_1 distance between two distributions.

7.2.2 Selection from a pool of distributions

Many algorithms for learning mixtures over the domain $\mathcal{X}$ first obtain a *small* collection $\mathcal{F}$ of mixtures and then perform Maximum Likelihood test using the samples to output a distribution [110, 109]. Our algorithm also obtains a set of distributions containing at least one that is close to the underlying in ℓ_1 distance. The estimation problem now reduces to the following. Given a class $\mathcal{F}$ of distributions and samples from an unknown distribution $\mathbf{f}$, find a distribution in $\mathcal{F}$ that is *close* to $\mathbf{f}$. Let $D(\mathbf{f}, \mathcal{F}) \stackrel{\text{def}}{=} \min_{\mathbf{f}_i \in \mathcal{F}} D(\mathbf{f}, \mathbf{f}_i)$.

The well-known Scheffe's method [112] uses $\mathcal{O}(\epsilon^{-2} \log |\mathcal{F}|)$ samples from the underlying distribution $\mathbf{f}$, and in time $\mathcal{O}(\epsilon^{-2}|\mathcal{F}|^2 T \log |\mathcal{F}|)$ outputs a distribution in $\mathcal{F}$ with ℓ_1 distance of at most $9.1 \cdot \max(D(\mathbf{f}, \mathcal{F}), \epsilon)$ from $\mathbf{f}$, where T is the time required to compute the probability of an $x \in \mathcal{X}$ by a distribution in $\mathcal{F}$. A naive application of this algorithm requires time quadratic in the number of distributions in $\mathcal{F}$. Recently [113] proposed a variant of Scheffe's method called MODIFIED SCHEFFE that works in near linear time, albeit requiring slightly more samples. More precisely,

Lemma 7.1 ([113])**.** Let $\epsilon > 0$. For some constant c, given $\frac{c}{\epsilon^2} \log\left(\frac{|\mathcal{F}|}{\delta}\right)$ independent samples from a distribution $\mathbf{f}$, with probability $\geq 1 - \delta$, the output $\hat{\mathbf{f}}$ of MODIFIED SCHEFFE $D(\hat{\mathbf{f}}, \mathbf{f}) \leq 1000 \max(\epsilon, D(\mathbf{f}, \mathcal{F}))$. Furthermore, the algorithm runs in time

$\mathcal{O}\big(\frac{|\mathcal{F}|T\log(|\mathcal{F}|/\delta)}{\epsilon^2}\big)$.

For our problem of estimating k component mixtures in d-dimensions, $T = \mathcal{O}(dk)$ and $|\mathcal{F}| = \widetilde{\mathcal{O}}_{k,\epsilon}(d^2)$.

7.2.3 Lower bound

Using Fano's inequality, we show an information theoretic lower bound of $\Omega(dk/\epsilon^2)$ samples to learn k-component d-dimensional spherical Gaussian mixtures for any algorithm. More precisely,

Theorem 7.2 (Section 7.8)**.** Any algorithm that learns all k-component d-dimensional spherical Gaussian mixtures up to ℓ_1 distance ϵ with probability $\geq 1/2$ requires at least $\Omega(\frac{dk}{\epsilon^2})$ samples.

7.3 One-dimensional mixtures

Over the past decade estimation of one dimensional distributions has gained significant attention [11, 76, 106, 107, 111, 117]. We provide a simple estimator for learning one dimensional Gaussian mixtures using the Modified Scheffe estimator. Formally, given samples from f, a mixture of Gaussian distributions $p_i \stackrel{\text{def}}{=} N(\mu_i, \sigma_i^2)$ with weights $w_1, w_2, \ldots w_k$, our goal is to find a mixture $\hat{f} = (\hat{w}_1, \hat{w}_2, \ldots \hat{w}_k, \hat{p}_1, \hat{p}_2, \ldots \hat{p}_k)$ such that $D(f, \hat{f}) \leq \epsilon$. We make no assumption on the weights, means or the variances of the components. While we do not use the one dimensional algorithm in the d-dimensional setting, it provides insight to the usage of the Modified Scheffe estimator and may be of independent interest. As stated in before, our quantizations are based on samples and is an immediate consequence of the following observation for samples from a Gaussian distribution.

Lemma 7.3. Given n independent samples $x_1, \ldots, x_n$ from $N(\mu, \sigma^2)$, with probability $\geq 1 - \delta$ there are two samples x_j, x_k such that $|x_j - \mu| \leq \sigma\frac{7\log 2/\delta}{2n}$ and $|x_j - x_k - \sigma| \leq 2\sigma\frac{7\log 2/\delta}{2n}$.

Proof. The density of $N(\mu, \sigma^2)$ is $\geq (7\sigma)^{-1}$ in the interval $[\mu - \sqrt{2}\sigma, \mu + \sqrt{2}\sigma]$. Therefore, the probability that a sample occurs in the interval $\mu - \epsilon\sigma, \mu + \epsilon\sigma$ is

$\geq 2\epsilon/7$. Hence, the probability that none of the n samples occurs in $[\mu-\epsilon\sigma, \mu+\epsilon\sigma]$ is $\leq (1-2\epsilon/7)^n \leq e^{-2n\epsilon/7}$. If $\epsilon \geq \frac{7\log 2/\delta}{2n}$, then the probability that none of the samples occur in the interval is $\leq \delta/2$. A similar argument shows that there is a sample within interval, $[\mu+\sigma-\epsilon\sigma, \mu+\sigma+\epsilon\sigma]$, proving the lemma. □

The above lemma states that given samples from a Gaussian distribution, there would be a sample close to the mean and there would be two samples that are about a standard deviation apart. Hence, if we consider the set of all Gaussians $N(x_j, (x_j - x_k)^2) : 1 \leq j, k \leq n$, then that set would contain a Gaussian close to the underlying one. The same holds for mixtures and for a Gaussian mixture and we can create the set of candidate mixtures as follows.

Lemma 7.4. Given $n \geq \frac{120k\log(4k/\delta)}{\epsilon}$ samples from a mixture f of k Gaussians. Let $S = \{N(x_j, (x_j - x_k)^2) : 1 \leq j, k \leq n\}$ and $W = \{0, \frac{\epsilon}{2k}, \frac{2\epsilon}{2k} \ldots, 1\}$ be a set of weights. Let

$$\mathcal{F} \stackrel{\text{def}}{=} \{(\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_k, \hat{p}_1, \hat{p}_2, \ldots \hat{p}_k) : \\ \hat{p}_i \in S, \forall 1 \leq i \leq k-1, \hat{w}_i \in W, \hat{w}_k = 1 - (\hat{w}_1 + \ldots \hat{w}_{k-1}) \geq 0\}$$

be a set of $n^{2k}(2k/\epsilon)^{k-1} \leq n^{3k-1}$ candidate distributions. There exists $\hat{f} \in \mathcal{F}$ such that $D(f, \hat{f}) \leq \epsilon$.

Proof. Let $f = (w_1, w_2, \ldots w_k, p_1, p_2, \ldots p_k)$. For
$\hat{f} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_{k-1}, 1 - \sum_{i=1}^{k-1} \hat{w}_i, \hat{p}_1, \hat{p}_2, \ldots \hat{p}_k)$, by the triangle inequality,

$$D(f, \hat{f}) \leq \sum_{i=1}^{k-1} 2|\hat{w}_i - w_i| + \sum_{i=1}^{k} w_i D(p_i, \hat{p}_i).$$

We show that there is a distribution in $\hat{f} \in \mathcal{F}$ such that the sum above is bounded by ϵ. Since we quantize the grids as multiples of $\epsilon/2k$, we consider distributions in $\mathcal{F}$ such that each $|\hat{w}_i - w_i| \leq \epsilon/4k$, and therefore $\sum_i |\hat{w}_i - w_i| \leq \frac{\epsilon}{2}$.

We now show that for each p_i there is a $\hat{p}_i$ such that $w_i D(p_i, \hat{p}_i) \leq \frac{\epsilon}{2k}$, thus proving that $D(f, \hat{f}) \leq \epsilon$. If $w_i \leq \frac{\epsilon}{4k}$, then $w_i D(p_i, \hat{p}_i) \leq \frac{\epsilon}{2k}$. Otherwise, let $w_i' > \frac{\epsilon}{4k}$ be the fraction of samples from p_i. By Lemma 7.3 and 7.15, with

probability $\geq 1 - \delta/2k$,

$$\begin{aligned} D(p_i, \hat{p}_i)^2 &\leq 2\frac{(\mu_i - \mu_i')^2}{\sigma_i^2} + 16\frac{(\sigma_i - \sigma_i')^2}{\sigma_i^2} \\ &\leq \frac{25\log^2 \frac{4k}{\delta}}{(nw_i')^2} + \frac{800\log^2 \frac{4k}{\delta}}{(nw_i')^2} \\ &\leq \frac{825\log^2 \frac{4k}{\delta}}{(nw_i')^2}. \end{aligned}$$

Therefore,

$$w_i D(p_i, \hat{p}_i) \leq \frac{30 w_i \log \frac{4k}{\delta}}{nw_i'}.$$

Since $w_i > \epsilon/4k$, with probability $\geq 1 - \delta/2k$, $w_i \leq 2w_i'$. By the union bound with probability $\geq 1 - \delta/k$, $w_i D(p_i, \hat{p}_i) \leq \frac{60 \log \frac{4k}{\delta}}{n}$. Hence if $n \geq \frac{120k \log \frac{4k}{\delta}}{\epsilon}$, the above quantity is less than $\epsilon/2k$. The total error probability is $\leq \delta$ by the union bound. □

Running the Modified Scheffe algorithm on the above set of candidates $\mathcal{F}$ yields a mixture that is close to the underlying one. By Lemma 7.1 and the above lemma we obtain

Corollary 7.5. Let $n \geq c \cdot \frac{k}{\epsilon^2} \log \frac{k}{\epsilon\delta}$ for some constant c. There is an algorithm that runs in time $\mathcal{O}\left(\left(\frac{k \log(k/\epsilon\delta)}{\epsilon}\right)^{3k-1} \frac{k^2 \log(k/\epsilon\delta)}{\epsilon^2}\right)$, and returns a mixture $\hat{f}$ such that $D(f, \hat{f}) \leq 1000\epsilon$ with probability $\geq 1 - 2\delta$.

Proof. Use $n' \stackrel{\text{def}}{=} \frac{120k \log \frac{4k}{\delta}}{\epsilon}$ samples to generate a set of at most n'^{3k-1} candidate distributions as stated in Lemma 7.4. With probability $\geq 1-\delta$, one of the candidate distributions is ϵ-close to the underlying one. Run Modified Scheffe on this set of candidate distributions to obtain a 1000ϵ-close estimate of f with probability $\geq 1 - \delta$ (Lemma 7.1). The run time is dominated by the run time of Modified Scheffe which is $\mathcal{O}\left(\frac{|\mathcal{F}|T \log \frac{|\mathcal{F}|}{\delta}}{\epsilon^2}\right)$, where $|\mathcal{F}| = n'^{3k-1}$ and $T = k$. The total error probability is $\leq 2\delta$ by the union bound. □

Remark 7.6. [111] considered the one dimensional Gaussian mixture problem for two component mixtures. While the process of identifying the candidate means is same for both the results, the process of identifying the variances and proof techniques are different.

7.4 d-dimensional mixtures

Algorithm LEARN k-SPHERE learns mixtures of k spherical Gaussians using near-linear samples. For clarity and simplicity of proofs, we first prove the result when all components have the same variance σ^2, *i.e.*, $\mathbf{p}_i = N(\boldsymbol{\mu}_i, \sigma^2 \mathbb{I}_d)$ for $1 \leq i \leq k$. A modification of this algorithm works for components with different variances. The core ideas are same and we discuss the changes in Section 7.4.3. The algorithm starts out by estimating σ^2 and we discuss this step later. We estimate the means in three steps, a coarse single-linkage clustering, recursive spectral clustering and search over span of means. We now discuss the necessity of these steps.

7.4.1 Estimating the span of means

A simple modification of the one dimensional algorithm can be used to learn mixtures in d dimensions, however, the number of candidate mixtures would be exponential in d, the number of dimensions. As stated in before, given the span of the mean vectors $\boldsymbol{\mu}_i$, we can grid the k dimensional span to the required accuracy ϵ_g and use MODIFIED SCHEFFE, to obtain a polynomial time algorithm. One of the natural and well-used methods to estimate the span of mean vectors is using the correlation matrix [97]. Consider the correlation-type matrix,

$$S = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}(i)\mathbf{X}(i)^t - \sigma^2 \mathbb{I}_d.$$

For a sample $\mathbf{X}$ from a particular component j, $\mathbb{E}[\mathbf{X}\mathbf{X}^t] = \sigma^2 \mathbb{I}_d + \boldsymbol{\mu}_j \boldsymbol{\mu}_j{}^t$, and the expected fraction of samples from $\mathbf{p}_j$ is w_j. Hence

$$\mathbb{E}[S] = \sum_{j=1}^{k} w_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j{}^t.$$

Therefore, as $n \to \infty$, S converges to $\sum_{j=1}^{k} w_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j{}^t$, and its top k eigenvectors span the means.

While the above intuition is well understood, the number of samples necessary for convergence is not well studied. We wish $\widetilde{\mathcal{O}}(d)$ samples to be sufficient for the convergence irrespective of the values of the means. However this is not true

when the means are far apart. In the following example we demonstrate that the convergence of averages can depend on their separation.

Example 7.7. Consider the special case, $d = 1$, $k = 2$, $\sigma^2 = 1$, $w_1 = w_2 = 1/2$, and mean differences $|\mu_1 - \mu_2| = L \gg 1$. Given this prior information, one can estimate the average of the mixture, that yields $(\mu_1 + \mu_2)/2$. Solving equations obtained by $\mu_1 + \mu_2$ and $\mu_1 - \mu_2 = L$ yields μ_1 and μ_2. The variance of the mixture is $1 + L^2/4 > L^2/4$. With additional Chernoff type bounds, one can show that given n samples the error in estimating the average is

$$|\mu_1 + \mu_2 - \hat{\mu}_1 - \hat{\mu}_2| \approx \Theta\left(L/\sqrt{n}\right).$$

Hence, estimating the means to high precision requires $n \geq L^2$, *i.e.,* the higher separation, the more samples are necessary if we use the sample mean.

A similar phenomenon happens in the convergence of the correlation matrices, where the variances of quantities of interest increase with separation. In other words, for the span to be accurate the number of samples necessary increases with the separation. To overcome this, a natural idea is to cluster the Gaussians such that the component means in the same cluster are close and then estimate the span of means, and apply SCHEFFE on the span within each cluster.

For clustering, we use another spectral algorithm. Even though spectral clustering algorithms are studied in [97, 99], they assume that the weights are strictly bounded away from 0, which does not hold here. We use a simple recursive clustering algorithm that takes a cluster C with average $\overline{\boldsymbol{\mu}}(C)$. If there is a component in the cluster such that $\sqrt{w_i}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2$ is $\Omega(\log(n/\delta)\sigma)$, then the algorithm divides the cluster into two nonempty clusters without any mis-clustering. For technical reasons similar to the above example, we first use a coarse clustering algorithm that ensures that the mean separation of any two components within each cluster is $\widetilde{\mathcal{O}}(d^{1/4}\sigma)$.

Our algorithm thus comprises of (i) variance estimation (ii) a coarse clustering ensuring that means are within $\widetilde{\mathcal{O}}(d^{1/4}\sigma)$ of each other in each cluster (iii) a recursive spectral clustering that reduces the mean separation to $\mathcal{O}(\sqrt{k^3 \log(n/\delta)}\sigma)$

(iv) estimating the span of mean within each cluster, and (v) quantizing the means and running MODIFIED SCHEFFE on the resulting candidate mixtures.

Algorithm LEARN k-SPHERE

Input: n samples $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)$ from $\mathbf{f}$ and ϵ.

1. **Sample variance:** $\hat{\sigma}^2 = \min_{a \neq b : a,b \in [k+1]} ||\mathbf{x}(a) - \mathbf{x}(b)||_2^2 / 2d$.

2. **Coarse single-linkage clustering:** Start with each sample as a cluster,

   - While $\exists$ two clusters with squared-distance $\leq 2d\hat{\sigma}^2 + 23\hat{\sigma}^2\sqrt{d \log(n^2/\delta)}$, merge them.

3. **Recursive spectral-clustering:** While there is a cluster C with $|C| \geq n\epsilon/5k$ and spectral norm of its sample covariance matrix $\geq 12k^2\hat{\sigma}^2 \log n^3/\delta$,

   - Use $n\epsilon/8k^2$ of the samples to find the largest eigenvector and discard these samples.
   - Project the remaining samples on the largest eigenvector.
   - Perform single-linkage in the projected space (as before) till the distance between clusters is $> 3\hat{\sigma}\sqrt{\log(n^2k/\delta)}$ creating new clusters.

4. **Exhaustive search**: Let $\epsilon_g = \epsilon/(16k^{3/2})$, $L = 200\sqrt{k^4\epsilon^{-1} \log \frac{n^2}{\delta}}$, $L' = \frac{32k\sqrt{\log n^2/\delta}}{\epsilon}$, and $G = \{-L, \ldots, -\epsilon_g, 0, \epsilon_g, 2\epsilon_g, \ldots L\}$. Let $W = \{0, \epsilon/(4k), 2\epsilon/(4k), \ldots 1\}$ and $\Sigma \stackrel{\text{def}}{=} \{\sigma^2 : \sigma^2 = \hat{\sigma}^2(1 + i\epsilon/d\sqrt{128dk^2}), \forall -L' < i \leq L'\}$.

   - For each cluster C find its top $k-1$ eigenvectors $\mathbf{u}_1, \ldots \mathbf{u}_{k-1}$. Let $\text{Span}(C) = \{\hat{\overline{\boldsymbol{\mu}}}(C) + \sum_{i=1}^{k-1} g_i\hat{\sigma}\mathbf{u}_i : g_i \in G\}$.
   - Let $\text{Span} = \cup_{C:|C| \geq \frac{n\epsilon}{5k}} \text{Span}(C)$.
   - For all $w_i' \in W$, $\sigma'^2 \in \Sigma$, $\hat{\boldsymbol{\mu}}_i \in \text{Span}$,
     add $\{(w_1', \ldots, w_{k-1}', 1 - \sum_{i=1}^{k-1} w_i', N(\hat{\boldsymbol{\mu}}_1, \sigma'^2), \ldots, N(\hat{\boldsymbol{\mu}}_k, \sigma'^2)\}$ to $\mathcal{F}$.

5. Run MODIFIED SCHEFFE on $\mathcal{F}$ and output the resulting distribution.

7.4.2 Sketch of correctness

We now describe the steps stating the performance of each step of Algorithm LEARN k-SPHERE. To simplify the bounds and expressions, we assume that $d > 1000$ and $\delta \geq \min(2n^2e^{-d/10}, 1/3)$. For smaller values of δ, we run the algorithm with error $1/3$ and repeat it $\mathcal{O}(\log \frac{1}{\delta})$ times to choose a set of candidate mixtures $\mathcal{F}_\delta$. By the Chernoff-bound with error $\leq \delta$, $\mathcal{F}_\delta$ contains a mixture ϵ-close to $\mathbf{f}$. Finally, we run MODIFIED SCHEFFE on $\mathcal{F}_\delta$ to obtain a mixture that is close to $\mathbf{f}$. By the union bound and Lemma 7.1, the error of the new algorithm is $\leq 2\delta$.

Variance estimation: Let $\hat{\sigma}$ be the variance estimate from step 1. If $\mathbf{X}(1)$ and $\mathbf{X}(2)$ are two samples from the components i and j respectively, then $\mathbf{X}(1) - \mathbf{X}(2)$ is distributed $N(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j, 2\sigma^2\mathbb{I}_d)$. Hence for large d, $||\mathbf{X}(1) - \mathbf{X}(2)||_2^2$ concentrates around $2d\sigma^2 + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2$. By the pigeon-hole principle, given $k+1$ samples, two of them are from the same component. Therefore, the minimum pairwise distance between $k+1$ samples is close to $2d\sigma^2$. This is made precise in the next lemma which states that $\hat{\sigma}^2$ is a good estimate of the variance.

Lemma 7.8 (Section 7.6.1)**.** Given n samples from the k-component mixture, with probability $1 - 2\delta$, $|\hat{\sigma}^2 - \sigma^2| \leq 2.5\sigma^2\sqrt{\log(n^2/\delta)/d}$.

Coarse single-linkage clustering: The second step is a single-linkage routine that clusters mixture components with *far* means. Single-linkage is a simple clustering scheme that starts out with each data point as a cluster, and at each step merges the two nearest clusters to form a larger cluster. The algorithm stops when the distance between clusters is larger than a pre-specified threshold.

Suppose the samples are generated by a one-dimensional mixture of k components that are *far*, then *with high probability*, when the algorithm generates k clusters all the samples within a cluster are generated by a single component. More precisely, if $\forall i, j \in [k]$, $|\mu_i - \mu_j| = \Omega(\sigma \log n)$, then all the n samples concentrate around their respective means and the separation between any two samples from different components would be larger than the largest separation between any two samples from the same component. Hence for a suitable value of threshold, single-linkage correctly identifies the clusters. For d-dimensional Gaussian mixtures a similar property holds, with minimum separation $\Omega((d \log \frac{n}{\delta})^{1/4}\sigma)$. More precisely,

Lemma 7.9 (Section 7.6.2)**.** After Step 2 of LEARN k-SPHERE, with probability $\geq 1-2\delta$, all samples from each component will be in the same cluster and the maximum distance between two components within each cluster is $\leq 10k\sigma\left(d\log\frac{n^2}{\delta}\right)^{1/4}$.

Recursive spectral-clustering: The clusters formed at the beginning of this step consist of components with mean separation $\mathcal{O}(\sigma d^{1/4}\log\frac{n}{\delta})$. We now recursively zoom into the clusters formed and show that it is possible to cluster the components with much smaller mean separation. Note that since the matrix is symmetric, the largest magnitude of the eigenvalue is the same as the spectral norm. We first find the largest eigenvector of

$$S(C) \stackrel{\text{def}}{=} \frac{1}{|C|}\Big(\sum_{\mathbf{x}\in C}(\mathbf{x}-\hat{\overline{\boldsymbol{\mu}}}(C))(\mathbf{x}-\hat{\overline{\boldsymbol{\mu}}}(C))^t\Big) - \hat{\sigma}^2\mathbb{I}_d,$$

which is the sample covariance matrix with its diagonal term reduced by $\hat{\sigma}^2$. We then project our samples to this vector and if there are two components with means far apart, then using single-linkage we divide the cluster into two. The following lemma shows that this step performs accurate clustering of components with well separated means.

Lemma 7.10 (Section 7.6.3)**.** Let $n \geq c\cdot\frac{dk^4}{\epsilon}\log\frac{n^3}{\delta}$. After recursive clustering, with probability $\geq 1-4\delta$, the samples are divided into clusters such that for each component i within a cluster C, $\sqrt{w_i}\,||\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}}(C)||_2 \leq 25\sigma\sqrt{k^3\log(n^3/\delta)}$. Furthermore, all the samples from one component remain in a single cluster.

Exhaustive search and Scheffe: After step 3, all clusters have a small weighted radius $\sqrt{w_i}\,||\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}}(C)||_2 \leq 25\sigma\sqrt{k^3\log\frac{n^3}{\delta}}$. It can be shown that the eigenvectors give an accurate estimate of the span of $\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}}(C)$ within each cluster. More precisely,

Lemma 7.11 (Section 7.6.4)**.** Let $n \geq c\cdot\frac{dk^9}{\epsilon^4}\log^2\frac{d}{\delta}$ for some constant c. After step 3, with probability $\geq 1-7\delta$, if $|C| \geq n\epsilon/5k$, then the projection of $[\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)]/\,||\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}}(C)||_2$ on the space orthogonal to the span of top $k-1$ eigenvectors has magnitude $\leq \frac{\epsilon\sigma}{8\sqrt{2}k\sqrt{w_i}||\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}}(C)||_2}$.

We now have accurate estimates of the spans of the cluster means and each cluster has components with close means. It is now possible to grid the set of possibilities in each cluster to obtain a set of distributions such that one of them is close to the underlying. There is a trade-off between a dense grid to obtain a good estimation and the computation time required. The final step takes the sparsest grid possible to ensure an error $\leq \epsilon$. This is quantized below.

Theorem 7.12 (Section 7.6.5)**.** Let $n \geq c \cdot \frac{dk^9}{\epsilon^4} \log^2 \frac{d}{\delta}$ for some constant c. Then Algorithm Learn k-Sphere, with probability $\geq 1 - 9\delta$, outputs a distribution $\hat{\mathbf{f}}$ such that $D(\hat{\mathbf{f}}, \mathbf{f}) \leq 1000\epsilon$. Furthermore, the algorithm runs in time

$$\mathcal{O}\Big(n^2 d \log n + d\Big(\frac{k^7}{\epsilon^3} \log^2 \frac{d}{\delta}\Big)^{\frac{k^2}{2}}\Big).$$

Note that the run time is calculated based on an efficient implementation of single-linkage clustering and the exponential term is not optimized.

7.4.3 Mixtures with unequal variances

We generalize the results to mixtures with components having different variances. Let $\mathbf{p}_i = N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \mathbb{I}_d)$ be the ith component. The key differences between Learn k-Sphere and the algorithm for learning mixtures with unequal variances are:

1. In Learn k-Sphere, we first estimated the component variance σ and divided the samples into clusters such that within each cluster the means are separated by $\widetilde{\mathcal{O}}(d^{1/4}\sigma)$. We modify this step such that the samples are clustered such that within each cluster the components not only have mean separation $\mathcal{O}(d^{1/4}\sigma)$, but variances are also a factor at most $1 + \widetilde{\mathcal{O}}\big(1/\sqrt{d}\big)$ apart.

2. Once the variances in each cluster are within a multiplicative factor of $1 + \widetilde{\mathcal{O}}\big(1/\sqrt{d}\big)$ of each other, it can be shown that the performance of the recursive spectral clustering step does not change more than constants.

3. After obtaining clusters with *similar* means and variances, the exhaustive search algorithm follows, though instead of having a single σ' for all clusters, we can have a different σ' for each cluster, which is estimated using the average pair wise distance between samples in the cluster.

The changes in the recursive clustering step and the exhaustive search step are easy to see and we omit them. The coarse clustering step requires additional tools and we describe them in Section 7.7.

7.5 Preliminaries

7.5.1 Bounds on ℓ_1 distance

For two d dimensional product distributions $\mathbf{p}_1$ and $\mathbf{p}_2$, if we bound the ℓ_1 distance on each coordinate by ϵ, then by triangle inequality $D(\mathbf{p}_1, \mathbf{p}_2) \leq d\epsilon$. However this bound is often weak. One way to obtain a stronger bound is to relate ℓ_1 distance to Bhattacharyya parameter, which is defined as follows: Bhattacharyya parameter $B(p_1, p_2)$ between two distributions p_1 and p_2 is

$$B(p_1, p_2) = \int_{x \in \mathcal{X}} \sqrt{p_1(x) p_2(x)} dx.$$

The ℓ_1 distance between p_1 and p_2 can be bounded in terms of $B(p_1, p_2)$ as follows.

Lemma 7.13. For distributions p_1 and p_2,

$$D(p_1, p_2)^2 \leq 8(1 - B(p_1, p_2)).$$

Proof. Since $\int_{x \in \mathcal{X}} p_1(x) dx = \int_{x \in \mathcal{X}} p_2(x) dx = 1$,

$$\int_{x \in \mathcal{X}} \left(\sqrt{p_1(x)} - \sqrt{p_2(x)} \right)^2 dx = 2(1 - B(p_1, p_2)).$$

Moreover since $(a + b)^2 \leq 2a^2 + 2b^2$,

$$\int_{x \in \mathcal{X}} \left(\sqrt{p_1(x)} + \sqrt{p_2(x)} \right)^2 dx \leq 4$$

These bounds along with the following Cauchy-Schwarz inequality yields the lemma.

$$\int_{x\in\mathcal{X}} \left(\sqrt{p_1(x)}+\sqrt{p_2(x)}\right)^2 dx \cdot \int_{x\in\mathcal{X}} \left(\sqrt{p_1(x)}-\sqrt{p_2(x)}\right)^2 dx$$
$$\geq \left(\int_{x\in\mathcal{X}} |p_1(x)-p_2(x)|dx\right)^2 = D(p_1,p_2)^2. \qquad \square$$

By the definition of Bhattacharyya distance, it is multiplicative for product distributions, namely for two product distributions $\mathbf{p}_1$ and $\mathbf{p}_2$, $B(\mathbf{p}_1,\mathbf{p}_2) = \prod_{i=1}^{d} B(p_{1,i},p_{2,i})$. We use this with the previous lemma to bound the ℓ_1 distance of Gaussian mixtures.

We first bound Bhattacharyya parameter for two one-dimensional Gaussian distributions.

Lemma 7.14. The Bhattacharyya parameter for two one dimensional Gaussian distributions $p_1 = N(\mu_1,\sigma_1^2)$ and $p_2 = N(\mu_2,\sigma_2^2)$ is

$$B(p_1,p_2) \geq 1 - \frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)} - \frac{(\sigma_1^2-\sigma_2^2)^2}{(\sigma_1^2+\sigma_2^2)^2}.$$

Proof. For Gaussian distributions a straight-forward computation shows that $B(p_1,p_2) = ye^{-x}$, where $x = \frac{(\mu_1-\mu_2)^2)}{4(\sigma_1^2+\sigma_2^2)}$ and $y = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}$. Observe that

$$y = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} = \sqrt{1-\frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}} \geq 1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2} \geq 1 - \frac{(\sigma_1^2-\sigma_2^2)^2}{(\sigma_1^2+\sigma_2^2)^2}.$$

Hence,

$$B(p_1,p_2) = ye^{-x} \geq y(1-x) \geq (1-x)\left(1-\frac{(\sigma_1^2-\sigma_2^2)^2}{(\sigma_1^2+\sigma_2^2)^2}\right) \geq 1 - x - \frac{(\sigma_1^2-\sigma_2^2)^2}{(\sigma_1^2+\sigma_2^2)^2}.$$

Substituting the value of x results in the lemma. $\square$

Therefore,

$$\begin{aligned} B(\mathbf{p}_1,\mathbf{p}_2) &= \prod_{i=1}^{d} B(p_{1,i},p_{2,i}) \\ &\geq \prod_{i=1}^{d} \left[1 - \frac{(\hat{\boldsymbol{\mu}}_{1,i}-\hat{\boldsymbol{\mu}}_{2,i})^2}{4(\sigma_{1,i}^2+\sigma_{2,i}^2)} - \frac{(\sigma_{1,i}^2-\sigma_{2,i}^2)^2}{(\sigma_{1,i}^2+\sigma_{2,i}^2)^2}\right] \\ &\geq 1 - \sum_{i=1}^{d} \left[\frac{(\hat{\boldsymbol{\mu}}_{1,i}-\hat{\boldsymbol{\mu}}_{2,i})^2}{4(\sigma_{1,i}^2+\sigma_{2,i}^2)} + \frac{(\sigma_{1,i}^2-\sigma_{2,i}^2)^2}{(\sigma_{1,i}^2+\sigma_{2,i}^2)^2}\right], \end{aligned}$$

where the last step uses $\prod(1-x_i) \geq 1 - \sum_i x_i$ for $x_i \in (0,1)$.

Using this with Lemma 7.13,

Lemma 7.15. For any two Gaussian product distributions $\mathbf{p}_1$ and $\mathbf{p}_2$,

$$D(\mathbf{p}_1, \mathbf{p}_2)^2 \leq \sum_{i=1}^{d} 2\frac{(\hat{\boldsymbol{\mu}}_{1,i} - \hat{\boldsymbol{\mu}}_{2,i})^2}{\sigma_{1,i}^2 + \sigma_{2,i}^2} + 8\frac{(\sigma_{1,i}^2 - \sigma_{2,i}^2)^2}{(\sigma_{1,i}^2 + \sigma_{2,i}^2)^2}. \qquad \square$$

7.5.2 Matrix eigenvalues

We now state few simple lemmas on the eigenvalues of perturbed matrices.

Lemma 7.16. Let $\lambda_1^A \geq \lambda^A \geq \ldots \lambda_d^A \geq 0$ and $\lambda_1^B \geq \lambda^B \geq \ldots \lambda_d^B \geq 0$ be the eigenvalues of two symmetric matrices A and B respectively. If $||A - B|| \leq \epsilon$, then $\forall\, i$, $|\lambda_i^A - \lambda_i^B| \leq \epsilon$.

Proof. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_d$ be a set of eigenvectors of A that corresponds to $\lambda_1^A, \lambda_2^A, \ldots \lambda_d^A$. Similarly let $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_d$ be eigenvectors of B Consider the first eigenvalue of B,

$$\lambda_1^B = ||B|| = ||A + (B - A)|| \geq ||A|| - ||B - A|| \geq \lambda_1^A - \epsilon.$$

Now consider an $i > 1$. If $\lambda_i^B < \lambda_i^A - \epsilon$, then by definition of eigenvalues

$$\max_{\mathbf{v}:\forall j \leq i-1, \mathbf{v}\cdot\mathbf{v}_j = 0} ||B\mathbf{v}||_2 < \lambda_i^A - \epsilon.$$

Now consider a unit vector $\sum_{j=1}^{i} \alpha_j \mathbf{u}_j$ in the span of $\mathbf{u}_1, \ldots \mathbf{u}_i$, that is orthogonal to $\mathbf{v}_1, \ldots \mathbf{v}_{i-1}$. For this vector,

$$\begin{aligned}
\left|\left| B\sum_{j=1}^{i} \alpha_j \mathbf{u}_j \right|\right|_2 &\geq \left|\left| A\sum_{j=1}^{i} \alpha_j \mathbf{u}_j \right|\right|_2 - \left|\left| (A-B)\sum_{j=1}^{i} \alpha_j \mathbf{u}_j \right|\right|_2 \\
&\geq \sqrt{\sum_{j=1}^{i} \alpha_j^2 (\lambda_j^A)^2} - \epsilon \\
&\geq \lambda_i^A - \epsilon,
\end{aligned}$$

a contradiction. Hence, $\forall i \leq d$, $\lambda_i^B \geq \lambda_i^A - \epsilon$. The proof in the other direction is similar and omitted. $\square$

Lemma 7.17. Let $A = \sum_{i=1}^{k} \eta_i^2 \mathbf{u}_i \mathbf{u}_i^t$ be a positive semidefinite symmetric matrix for $k \leq d$. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_k$ span a $k-1$ dimensional space. Let $B = A + R$, where $||R|| \leq \epsilon$. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{k-1}$ be the top $k-1$ eigenvectors of B. Then the projection of $\mathbf{u}_i$ in space orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{k-1}$ is $\leq \frac{2\sqrt{\epsilon}}{\eta_i}$.

Proof. Let λ_i^B be the i^{th} largest eigenvalue of B. Observe that $B + \epsilon \mathbb{I}_d$ is a positive semidefinite matrix as for any vector $\mathbf{v}$, $\mathbf{v}^t(A + R + \epsilon \mathbb{I}_d)\mathbf{v} \geq 0$. Furthermore $||A + R + \epsilon \mathbb{I}_d - A|| \leq 2\epsilon$. Since eigenvalues of $B + \epsilon \mathbb{I}_d$ is $\lambda^B + \epsilon$, by Lemma 7.16, for all $i \leq d$, $|\lambda_i^A - \lambda_i^B - \epsilon| \leq 2\epsilon$. Therefore, $|\lambda_i^B|$ for $i \geq k$ is $\leq 3\epsilon$.

Let $\mathbf{u}_i = \sum_{j=1}^{k-1} \alpha_{i,j} \mathbf{v}_j + \sqrt{1 - \sum_{j=1}^{k-1} \alpha_{i,j}^2}\, \mathbf{u}'$, for a vector $\mathbf{u}'$ orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{k-1}$. We compute $\mathbf{u}'^t A \mathbf{u}'$ in two ways. Since $A = B - R$,

$$|\mathbf{u}'^t (B - R)\mathbf{u}'| \leq |\mathbf{u}'^t B \mathbf{u}'| + |\mathbf{u}'^t R \mathbf{u}'| \leq ||B\mathbf{u}'||_2 + ||R||\,.$$

Since $\mathbf{u}'$ is orthogonal to first k eigenvectors, we have $||B\mathbf{u}'||_2 \leq 3\epsilon$ and hence $|\mathbf{u}'^(B - R)\mathbf{u}'| \leq 4\epsilon$.

$$\mathbf{u}'^t A \mathbf{u}' \geq \eta_i^2 \big(1 - \sum_{j=1}^{k-1} \alpha_{i,j}^2\big).$$

We have shown that the above quantity is $\leq 4\epsilon$. Therefore $\left(1 - \sum_{j=1}^{k-1} \alpha_{i,j}^2\right)^{1/2} \leq 2\sqrt{\epsilon}/\eta_i$. □

7.6 Proofs for Learn k-Sphere

We first state a simple concentration result that helps us in other proofs.

Lemma 7.18. Given n samples from a set of Gaussian distributions, with probability $\geq 1 - 2\delta$, for every pair of samples $\mathbf{X} \sim N(\boldsymbol{\mu}_1, \sigma^2 \mathbb{I}_d)$ and $\mathbf{Y} \sim N(\boldsymbol{\mu}_2, \sigma^2 \mathbb{I}_d)$, $||\mathbf{X} - \mathbf{Y}||_2^2$ is at most

$$2d\sigma^2 + 4\sigma^2 \sqrt{d \log \frac{n^2}{\delta}} + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 + 4\sigma\, ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 \sqrt{\log \frac{n^2}{\delta}} + 4\sigma^2 \log \frac{n^2}{\delta}. \quad (7.1)$$

and

$$||\mathbf{X} - \mathbf{Y}||_2^2 \geq 2d\sigma^2 - 4\sigma^2 \sqrt{d \log \frac{n^2}{\delta}} + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 - 4\sigma\, ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 \sqrt{\log \frac{n^2}{\delta}}. \quad (7.2)$$

Proof. We prove the lower bound, the proof for the upper bound is similar and omitted. Since $\mathbf{X}$ and $\mathbf{Y}$ are Gaussians, $\mathbf{X}-\mathbf{Y}$ is distributed as $N(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2, 2\sigma^2)$. Rewriting $||\mathbf{X}-\mathbf{Y}||_2$

$$||\mathbf{X}-\mathbf{Y}||_2^2 = ||\mathbf{X}-\mathbf{Y}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||_2^2+||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2^2+2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\cdot(\mathbf{X}-\mathbf{Y}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)).$$

Let $\mathbf{Z} = \mathbf{X}-\mathbf{Y}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, then $\mathbf{Z} \sim N(\mathbf{0}, 2\sigma^2\mathbb{I}_d)$. Therefore by Lemma A.5, with probability $1-\delta/n^2$,

$$||\mathbf{Z}||_2^2 \geq 2d\sigma^2 - 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}}.$$

Furthermore $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\cdot\mathbf{Z}$ is sum of Gaussians and hence a Gaussian distribution. It has mean 0 and variance $2\sigma^2\,||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2^2$. Therefore, by Lemma A.4 with probability $1-\delta/n^2$,

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\cdot\mathbf{Z} \geq -2\sigma\,||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2\sqrt{\log\frac{n^2}{\delta}}.$$

By the union bound with probability $1-2\delta/n^2$,

$$||\mathbf{X}-\mathbf{Y}||_2^2 \geq 2d\sigma^2 - 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} + ||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2^2 - 4\sigma\,||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2\sqrt{\log\frac{n^2}{\delta}}.$$

There are $\binom{n}{2}$ pairs and the lemma follows by the union bound. □

7.6.1 Proof of Lemma 7.8

We show that if Equations (7.1) and (7.2) are satisfied, then the lemma holds. The error probability is that of Lemma 7.18 and is $\leq 2\delta$. Since the minimum is over $k+1$ indices, at least two samples are from the same component. Applying Equations (7.1) and (7.2) for these two samples

$$2d\hat{\sigma}^2 \leq 2d\sigma^2 + 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} + 4\sigma^2\log\frac{n^2}{\delta}.$$

Similarly by Equations (7.1) and (7.2) for any two samples $\mathbf{X}(a), \mathbf{X}(b)$ in $[k+1]$,

$$\begin{aligned}||\mathbf{X}(a)-\mathbf{X}(b)||_2^2 &\geq 2d\sigma^2 - 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} + \left|\left|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right|\right|_2^2 - 4\sigma\left|\left|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right|\right|_2\sqrt{\log\frac{n^2}{\delta}} \\ &\geq 2d\sigma^2 - 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} - 4\sigma^2\log\frac{n^2}{\delta},\end{aligned}$$

where the last inequality follows from the fact that $\alpha^2-4\alpha\beta \geq -4\beta^2$. The result follows from the assumption that $d > 20\log n^2/\delta$.

7.6.2 Proof of Lemma 7.9

We show that if Equations (7.1) and (7.2) are satisfied, then the lemma holds. The error probability is that of Lemma 7.18 and is $\leq 2\delta$. Since Equations (7.1) and (7.2) are satisfied, by the proof of Lemma 7.8,

$$|\hat{\sigma}^2 - \sigma^2| \leq 2.5\sigma^2\sqrt{\frac{\log(n^2/\delta)}{d}}.$$

If two samples $X(a)$ and $X(b)$ are from the same component, by Lemma 7.18,

$$\begin{aligned}||\mathbf{X}(a) - \mathbf{X}(b)||_2^2 &\leq 2d\sigma^2 + 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} + 4\sigma^2\log\frac{n^2}{\delta} \\ &\leq 2d\sigma^2 + 5\sigma^2\sqrt{d\log\frac{n^2}{\delta}}.\end{aligned}$$

By Lemma 7.8, the above quantity is less than $2d\hat{\sigma}^2 + 23\hat{\sigma}^2\sqrt{d\log\frac{n^2}{\delta}}$. Hence all the samples from the same component are in a single cluster.

Suppose there are two samples from different components in a cluster, then by Equations (7.1) and (7.2),

$$\begin{aligned}&2d\hat{\sigma}^2 + 23\hat{\sigma}^2\sqrt{d\log\frac{n^2}{\delta}} \\ &\geq 2d\sigma^2 - 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2 - 4\sigma\left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2\sqrt{\log\frac{n^2}{\delta}}.\end{aligned}$$

Relating $\hat{\sigma}^2$ and σ^2 using Lemma 7.8,

$$\begin{aligned}&2d\sigma^2 + 40\sigma^2\sqrt{d\log\frac{n^2}{\delta}} \\ &\geq 2d\sigma^2 - 4\sigma^2\sqrt{d\log\frac{n^2}{\delta}} + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2 - 4\sigma\left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2\sqrt{\log\frac{n^2}{\delta}}.\end{aligned}$$

Hence $\left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2 \leq 10\sigma\left(d\log\frac{n^2}{\delta}\right)^{1/4}$. There are at most k components; therefore, any two components within the same cluster are at a distance $\leq 10k\sigma\left(d\log\frac{n^2}{\delta}\right)^{1/4}$.

7.6.3 Proof of Lemma 7.10

The proof is involved and we show it in steps. We first show few concentration bounds which we use later to argue that the samples are clusterable when the

sample covariance matrix has a large eigenvalue. Let $\hat{w}_i$ be the fraction of samples from component i. Let $\hat{\boldsymbol{\mu}}_i$ be the empirical average of samples from $\mathbf{p}_i$. Let $\hat{\overline{\boldsymbol{\mu}}}(C)$ be the empirical average of samples in cluster C. If C is the entire set of samples we use $\hat{\overline{\boldsymbol{\mu}}}$ instead of $\hat{\overline{\boldsymbol{\mu}}}(C)$. We first show a concentration inequality that we use in rest of the calculations.

Lemma 7.19. Given n samples from a k-component Gaussian mixture with probability $\geq 1 - 2\delta$, for every component i

$$||\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i||_2^2 \leq \left(d + 3\sqrt{d \log \frac{2k}{\delta}}\right) \frac{\sigma^2}{n\hat{w}_i} \text{ and } |\hat{w}_i - w_i| \leq \sqrt{\frac{2w_i \log \frac{2k}{\delta}}{n}} + \frac{2}{3}\frac{\log \frac{2k}{\delta}}{n}. \tag{7.3}$$

Proof. Since $\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i$ is distributed $N(0, \sigma^2 \mathbb{I}_d / n\hat{w}_i)$, by Lemma A.5 with probability $\geq 1 - \delta/k$,

$$||\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i||_2^2 \leq \left(d + 2\sqrt{d \log \frac{2k}{\delta}} + 2 \log \frac{2k}{\delta}\right) \frac{\sigma^2}{n\hat{w}_i} \leq \left(d + 3\sqrt{d \log \frac{2k}{\delta}}\right) \frac{\sigma^2}{n\hat{w}_i}.$$

The second inequality uses the fact that $d \geq 20 \log n^2/\delta$. For bounding the weights, observe that by Lemma A.3 with probability $\geq 1 - \delta/k$,

$$|\hat{w}_i - w_i| \leq \sqrt{\frac{2w_i \log 2k/\delta}{n}} + \frac{2}{3}\frac{\log 2k/\delta}{n}.$$

By the union bound the error probability is $\leq 2k\delta/2k = \delta$. □

A simple application of triangle inequality yields the following lemma.

Lemma 7.20. Given n samples from a k-component Gaussian mixture if Equation (7.3) holds, then

$$\left|\left|\sum_{i=1}^{k} \hat{w}_i (\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^t\right|\right| \leq \left(d + 3\sqrt{d \log \frac{2k}{\delta}}\right) \frac{k\sigma^2}{n}.$$

Lemma 7.21. Given n samples from a k-component Gaussian mixture, if Equation (7.3) holds and the maximum distance between two components is at most $10k\sigma\left(d \log \frac{n^2}{\delta}\right)^{1/4}$, then $\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}})\right|\right|_2 \leq c\sigma\sqrt{\frac{dk \log \frac{n^2}{\delta}}{n}}$, for a constant c.

Proof. Observe that

$$\begin{aligned}\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}} &= \sum_{i=1}^{k} \hat{w}_i \hat{\boldsymbol{\mu}}_i - w_i \boldsymbol{\mu}_i \\ &= \sum_{i=1}^{k} \hat{w}_i (\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + (\hat{w}_i - w_i)\boldsymbol{\mu}_i \\ &= \sum_{i=1}^{k} \hat{w}_i (\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + (\hat{w}_i - w_i)(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}). \end{aligned} \tag{7.4}$$

Hence by Equation (7.3) and the fact that the maximum distance between two components is at most $10k\sigma\big(d \log \frac{n^2}{\delta}\big)^{1/4}$,

$$\begin{aligned}\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2 \leq \sum_{i=1}^{k} \hat{w}_i \sqrt{\left(d + 3\sqrt{d \log \frac{2k}{\delta}}\right)} \frac{\sigma}{\sqrt{n\hat{w}_i}} \\ + \left(\sqrt{\frac{2w_i \log 2k/\delta}{n}} + \frac{2}{3}\frac{\log 2k/\delta}{n}\right) 10k \left(d \log \frac{n^2}{\delta}\right)^{1/4} \sigma.\end{aligned}$$

For $n \geq d \geq \max(k^4, 20 \log n^2/\delta, 1000)$, we get the above term is $\leq c\sqrt{\frac{kd \log n^2/\delta}{n}}\sigma$, for some constant c. □

We now make a simple observation on covariance matrices.

Lemma 7.22. Given n samples from a k-component mixture,

$$\begin{aligned}&\left|\left|\sum_{i=1}^{k} \hat{w}_i (\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t - \sum_{i=1}^{k} \hat{w}_i (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\right|\right| \\ &\leq 2\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2^2 + \sum_{i=1}^{k} 2\hat{w}_i \left|\left|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right|\right|_2^2 \\ &+ 2\left(\sqrt{k}\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2 + \sum_{i=1}^{k} \sqrt{\hat{w}_i}\left|\left|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right|\right|_2\right) \max_j \sqrt{\hat{w}_j}\left|\left|\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}\right|\right|_2.\end{aligned}$$

Proof. Observe that for any two vectors $\mathbf{u}$ and $\mathbf{v}$,

$$\mathbf{u}\mathbf{u}^t - \mathbf{v}\mathbf{v}^t = \mathbf{u}(\mathbf{u}^t - \mathbf{v}^t) + (\mathbf{u} - \mathbf{v})\mathbf{v}^t = (\mathbf{u} - \mathbf{v})(\mathbf{u} - \mathbf{v})^t + \mathbf{v}(\mathbf{u} - \mathbf{v})^t + (\mathbf{u} - \mathbf{v})\mathbf{v}^t.$$

Hence by triangle inequality,

$$\left|\left|\mathbf{u}\mathbf{u}^t - \mathbf{v}\mathbf{v}^t\right|\right| \leq \left|\left|\mathbf{u} - \mathbf{v}\right|\right|_2^2 + 2\left|\left|\mathbf{v}\right|\right|_2 \left|\left|\mathbf{u} - \mathbf{v}\right|\right|_2.$$

Applying the above observation to $\mathbf{u} = \hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}}$ and $\mathbf{v} = \boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}$, we get

$$\begin{aligned}
&\sum_{i=1}^{k} \hat{w}_i \left|\left|(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t - (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\right|\right| \\
&\leq \sum_{i=1}^{k} \left(\hat{w}_i \left|\left|\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}} - \boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right|\right|_2^2 + 2\sqrt{\hat{w}_i}\left|\left|\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right|\right|_2 \sqrt{\hat{w}_i}\left|\left|\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}} - \boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right|\right|_2 \right) \\
&\leq \sum_{i=1}^{k} \left(2\hat{w}_i \left|\left|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right|\right|_2^2 + 2\hat{w}_i \left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2^2 \right) \\
&+ \sum_{i=1}^{k} \left(2 \max_j \sqrt{\hat{w}_j}\left|\left|\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}\right|\right|_2 \left(\sqrt{\hat{w}_i}\left|\left|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right|\right|_2 + \sqrt{\hat{w}_i}\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2 \right) \right) \\
&\leq 2\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2^2 + \sum_{i=1}^{k} 2\hat{w}_i \left|\left|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right|\right|_2^2 \\
&+ 2\left(\sqrt{k}\left|\left|\hat{\overline{\boldsymbol{\mu}}} - \overline{\boldsymbol{\mu}}\right|\right|_2 + \sum_{i=1}^{k} \sqrt{\hat{w}_i}\left|\left|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right|\right|_2 \right) \max_j \sqrt{\hat{w}_j}\left|\left|\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}\right|\right|_2 .
\end{aligned}$$

The lemma follows from triangle inequality. □

The following lemma immediately follows from Lemmas 7.21 and 7.22.

Lemma 7.23. Given n samples from a k-component Gaussian mixture, if Equation (7.3) and the maximum distance between two components is at most $10k\sigma\left(d\log\frac{n^2}{\delta}\right)^{1/4}$, then

$$\begin{aligned}
&\left|\left|\sum_{i=1}^{k} \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t - \sum_{i=1}^{k} \hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\right|\right| \\
&\leq \frac{c\sigma^2 dk^2 \log\frac{n^2}{\delta}}{n} + c\sigma\sqrt{\frac{dk^2\log\frac{n^2}{\delta}}{n}} \max_i \sqrt{\hat{w}_i}\left|\left|\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right|\right|_2 ,
\end{aligned}$$

for a constant c.

Lemma 7.24. For a set of samples $\mathbf{X}(1), \ldots \mathbf{X}(n)$ from a k-component mixture,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{(\mathbf{X}(i) - \hat{\overline{\boldsymbol{\mu}}})(\mathbf{X}(i) - \hat{\overline{\boldsymbol{\mu}}})^t}{n} &= \sum_{i=1}^{k} \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t - \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^t \\
&+ \sum_{j|\mathbf{X}(j)\sim p_i} \frac{(\mathbf{X}(j) - \boldsymbol{\mu}_i)(\mathbf{X}(j) - \boldsymbol{\mu}_i)^t}{n}.
\end{aligned}$$

where $\hat{w}_i$ and $\hat{\boldsymbol{\mu}}_i$ are the empirical weights and averages of components i and $\hat{\overline{\boldsymbol{\mu}}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$.

Proof. The given expression can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}(i)-\hat{\overline{\boldsymbol{\mu}}})(\mathbf{X}(i)-\hat{\overline{\boldsymbol{\mu}}})^t = \sum_{i=1}^{k}\hat{w}_i\sum_{j|\mathbf{X}(j)\sim p_i}\frac{1}{n\hat{w}_i}\mathbf{X}(j)-\hat{\overline{\boldsymbol{\mu}}})(\mathbf{X}(j)-\hat{\overline{\boldsymbol{\mu}}})^t.$$

First observe that for any set of points x_i and their average $\hat{x}$ and any value a,

$$\sum_i (x_i-a)^2 = \sum_i (x_i-\hat{x})^2 + (\hat{x}-a)^2.$$

Hence for samples from a component i,

$$\begin{aligned}
&\sum_{j|\mathbf{X}(j)\sim p_i}\frac{1}{n\hat{w}_i}(\mathbf{X}(j)-\hat{\overline{\boldsymbol{\mu}}})(\mathbf{X}(j)-\hat{\overline{\boldsymbol{\mu}}})^t\\
&=\sum_{j|\mathbf{X}(j)\sim p_i}\frac{1}{n\hat{w}_i}(\hat{\boldsymbol{\mu}}_i-\hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i-\hat{\overline{\boldsymbol{\mu}}})^t+\sum_{j|\mathbf{X}(j)\sim p_i}\frac{1}{n\hat{w}_i}(\mathbf{X}(j)-\hat{\boldsymbol{\mu}}_i)(\mathbf{X}(j)-\hat{\boldsymbol{\mu}}_i)^t\\
&=(\hat{\boldsymbol{\mu}}_i-\hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i-\hat{\overline{\boldsymbol{\mu}}})^t+\sum_{j|\mathbf{X}(j)\sim p_i}\frac{1}{n\hat{w}_i}(\mathbf{X}(j)-\hat{\boldsymbol{\mu}}_i)(\mathbf{X}(j)-\hat{\boldsymbol{\mu}}_i)^t\\
&=(\hat{\boldsymbol{\mu}}_i-\hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i-\hat{\overline{\boldsymbol{\mu}}})^t\\
&+\sum_{j|\mathbf{X}(j)\sim p_i}\frac{1}{n\hat{w}_i}(\mathbf{X}(j)-\boldsymbol{\mu}_i)(\mathbf{X}(j)-\boldsymbol{\mu}_i)^t-(\hat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i)(\hat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i)^t.
\end{aligned}$$

Summing over all components results in the lemma. □

We now bound the error in estimating the eigenvalue of the covariance matrix.

Lemma 7.25. Given $\mathbf{X}(1),\ldots\mathbf{X}(n)$, n samples from a k-component Gaussian mixture, if Equations (7.1), (7.2), and (7.3) hold, then with probability $\geq 1-2\delta$,

$$\begin{aligned}
&\left|\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}(i)-\hat{\overline{\boldsymbol{\mu}}})(\mathbf{X}(i)-\hat{\overline{\boldsymbol{\mu}}})^t-\hat{\sigma}^2\mathbb{I}_d-\sum_{i=1}^{k}\hat{w}_i(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})^t\right|\right|\\
&\leq c(n)\stackrel{\text{def}}{=} c\sigma^2\sqrt{\frac{d\log\frac{n^2}{\delta}}{n}}+c\sigma^2\frac{dk^2\log\frac{n^2}{\delta}}{n}+c\sigma\sqrt{\frac{dk^2\log\frac{n^2}{\delta}}{n}}\max_i\sqrt{\hat{w}_i}\,||\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}}||_2\,,
\end{aligned}\tag{7.5}$$

for a constant c.

Proof. Since Equations (7.1), (7.2), and (7.3) hold, conditions in Lemmas 7.21 and 7.23 are satisfied. By Lemma 7.23,

$$\left|\left|\sum_{i=1}^{k} \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t - \sum_{i=1}^{k} \hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\right|\right|$$
$$= \mathcal{O}\left(\sigma^2 \frac{dk^2 \log \frac{n^2}{\delta}}{n} + \sigma\sqrt{\frac{dk^2 \log \frac{n^2}{\delta}}{n}} \max_i \sqrt{\hat{w}_i}\, ||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}||_2\right).$$

Hence it remains to show,

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}(i) - \hat{\overline{\boldsymbol{\mu}}})(\mathbf{X}(i) - \hat{\overline{\boldsymbol{\mu}}})^t - \sum_{i=1}^{k} \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t\right|\right| = \mathcal{O}\left(\sqrt{\frac{kd \log \frac{5k^2}{\delta}}{n}}\sigma^2\right).$$

By Lemma 7.24, the covariance matrix can be rewritten as

$$\sum_{i=1}^{k} \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}})^t - \hat{w}_i(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^t$$
$$+ \sum_{i=1}^{k} \sum_{j|\mathbf{X}(j)\sim p_i} \frac{1}{n}(\mathbf{X}(j) - \boldsymbol{\mu}_i)(\mathbf{X}(j) - \boldsymbol{\mu}_i)^t - \hat{\sigma}^2 \mathbb{I}_d. \tag{7.6}$$

We now bound the norms of second and third terms in the above equation. Consider the third term, $\sum_{i=1}^{k} \sum_{j|\mathbf{X}(j)\sim p_i} \frac{1}{n}(\mathbf{X}(j) - \boldsymbol{\mu}_i)(\mathbf{X}(j) - \boldsymbol{\mu}_i)^t$. Conditioned on the fact that $\mathbf{X}(j) \sim p_i$, $\mathbf{X}(j) - \boldsymbol{\mu}_i$ is distributed $N(0, \sigma^2 \mathbb{I}_d)$, therefore by Lemma A.7 and Lemma 7.8 ,with probability $\geq 1 - 2\delta$,

$$\left|\left|\sum_{i=1}^{k} \sum_{j|\mathbf{X}(j)\sim p_i} \frac{1}{n}(\mathbf{X}(j) - \boldsymbol{\mu}_i)(\mathbf{X}(j) - \boldsymbol{\mu}_i)^t - \hat{\sigma}^2 \mathbb{I}_d\right|\right| \leq c'\sqrt{\frac{d \log \frac{2d}{\delta}}{n}}\sigma^2 + 2.5\sigma^2\sqrt{\frac{\log \frac{n^2}{\delta}}{d}}.$$

The second term in Equation (7.6) is bounded by Lemma 7.20. Hence together with the fact that $d \geq 20 \log n^2/\delta$ we get that with probability $\geq 1 - 2\delta$, the second and third terms are bounded by $\mathcal{O}\left(\sigma^2\sqrt{\frac{dk}{n} \log \frac{n^2}{\delta}}\right)$. □

Lemma 7.26. Let $\mathbf{u}$ be the largest eigenvector of the sample covariance matrix and $n \geq c \cdot dk^2 \log \frac{n^2}{\delta}$. If $\max_i \sqrt{\hat{w}_i}\, ||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}||_2 = \alpha\sigma$ and Equation (7.5) holds, then there exists i such that $|\mathbf{u} \cdot (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})| \geq \sigma(\alpha - 1 - 1/\alpha)/\sqrt{k}$.

Proof. Observe that $\left|\left|\sum_j w_j \mathbf{v}_j \mathbf{v}_j^t\right|\right| \geq \left|\left|\sum_j w_j \mathbf{v}_j \mathbf{v}_j^t \frac{\mathbf{v}_i}{||\mathbf{v}_i||}\right|\right|_2 \geq w_i ||\mathbf{v}_i||_2^2$. Therefore

$$\left|\left|\sum_{i=1}^k \hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\right|\right| \geq \left|\left|\sum_{j=1}^k \hat{w}_j(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}})^t(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})/\left|\left|\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right|\right|\right|\right|_2$$
$$\geq \alpha^2\sigma^2.$$

Hence by Lemma 7.25 and the triangle inequality, the largest eigenvalue of the sample-covariance matrix is $\geq \alpha^2\sigma^2 - c(n)$. Similarly by applying Lemma 7.25 again we get, $\left|\left|\sum_{i=1}^k \hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\mathbf{u}\right|\right|_2 \geq \alpha^2\sigma^2 - 2c(n)$. By triangle inequality and Cauchy-Schwartz inequality,

$$\begin{aligned}
\left|\left|\sum_{i=1}^k \hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\mathbf{u}\right|\right|_2 &\leq \sum_{i=1}^k \left|\left|\hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^t\mathbf{u}\right|\right|_2 \\
&\leq \sum_{i=1}^k \hat{w}_i \left|\left|(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})\right|\right|_2 \max_j |(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}) \cdot \mathbf{u}| \\
&\leq \sqrt{\sum_{i=1}^k \hat{w}_i \left|\left|(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})\right|\right|_2^2} \max_j |(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}) \cdot \mathbf{u}| \\
&\leq \sqrt{k}\alpha\sigma \max_j |(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}) \cdot \mathbf{u}|.
\end{aligned}$$

Hence $\sqrt{k}\alpha\sigma \max_i |(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}) \cdot \mathbf{u}| \geq \alpha^2\sigma^2 - 2c(n)$. The lemma follows by substituting the bound on n in $c(n)$. □

We now make a simple observation on Gaussian mixtures.

Lemma 7.27. The samples from a subset of components A of the Gaussian mixture are distributed according to a Gaussian mixture of components A with weights being $w_i' = w_i/(\sum_{j\in A} w_j)$.

We now prove Lemma 7.10.

Proof of Lemma 7.10. Observe that we run the recursive clustering at most n times. At every step, the underlying distribution within a cluster is a Gaussian mixture. Let Equations (7.1), (7.2) hold with probability $1 - 2\delta$. Let Equations (7.3) (7.5) all hold with probability $\geq 1 - \delta'$, where $\delta' = \delta/2n$ at each of n

steps. By the union bound the total error is $\leq 2\delta + \delta' \cdot 2n \leq 3\delta$. Since Equations (7.1), (7.2) holds, the conditions of Lemmas 7.8 and 7.9 hold. Furthermore it can be shown that discarding at most $n\epsilon/4k$ samples at each step does not affect the calculations.

We first show that if $\sqrt{w_i}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \geq 25\sqrt{k^3 \log(n^3/\delta)}\sigma$, then the algorithm gets into the loop. Let w_i' be the weight of the component within the cluster and $n' \geq n\epsilon/5k$ be the number of samples in the cluster. Let $\alpha = 25\sqrt{k^3 \log(n^3/\delta)}$. By Fact 7.27, the components in cluster C have weight $w_i' \geq w_i$. Hence $\sqrt{w_i'}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \geq \alpha\sigma$. Since $\sqrt{w_i'}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \geq \alpha\sigma$, and by Lemma 7.9 $||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)|| \leq 10k\sigma(d \log n^2/\delta)^{1/4}$, we have

$$w_i' \geq \alpha^2/(100k^2\sqrt{d \log n^2/\delta}).$$

Hence by lemma 7.19, $w_i' \geq w_i/2$ and $\sqrt{\hat{w}_i'}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \geq \alpha\sigma/\sqrt{2}$. Hence by Lemma 7.25 and triangle inequality the largest eigenvalue of $S(C)$ is at least

$$\alpha^2\sigma^2/2 - c(n') \geq \alpha^2\sigma^2/4 \geq \alpha^2\hat{\sigma}^2/8 \geq 12\hat{\sigma}^2k^3 \log n^2/\delta' = 12\hat{\sigma}^2k^3 \log n^3/\delta.$$

Therefore the algorithm gets into the loop.

If $n' \geq n\epsilon/8k^2 \geq c{\cdot}dk^2 \log \frac{n^3}{\delta}$, then by Lemma 7.26, there exists a component i such that

$$|\mathbf{u} \cdot (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))| \geq \sigma(\alpha/\sqrt{2} - 1 - \sqrt{2}/\alpha)/\sqrt{k},$$

where $\mathbf{u}$ is the top eigenvector of the first $n\epsilon/4k^2$ samples.

Observe that $\sum_{i \in C} w_i \mathbf{u} \cdot (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)) = 0$ and

$$\max_i |\mathbf{u} \cdot (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))| \geq \sigma(\alpha/\sqrt{2} - 1 - \sqrt{2}/\alpha)/\sqrt{k}.$$

Let $\boldsymbol{\mu}_i$ be sorted according to their values of $\mathbf{u} \cdot (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))$, then

$$\begin{aligned}
\max_i |\mathbf{u} \cdot (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+1})| &\geq \sigma \frac{\alpha/\sqrt{2} - 1 - \sqrt{2}/\alpha}{k^{3/2}} \\
&\geq 12\sigma\sqrt{\log \frac{n^3}{\delta}} \\
&\geq 9\hat{\sigma}\sqrt{\log \frac{n^3}{\delta}},
\end{aligned}$$

where the last inequality follows from Lemma 7.8 and the fact that $d \geq 20 \log n^2/\delta$. For a sample from component $\mathbf{p}_i$, similar to the proof of Lemma 7.9, by Lemma A.4, with probability $\geq 1 - \delta/n^2k$,

$$||u \cdot (\mathbf{X}(i) - \boldsymbol{\mu}_i)|| \leq \sigma\sqrt{2\log(n^2k/\delta)}_2 \leq 2\hat{\sigma}\sqrt{\log(n^2k/\delta)},$$

where the second inequality follows from Lemma 7.8. Since there are two components that are far apart by $\geq 9\hat{\sigma}\sqrt{\log \frac{n^2}{\delta}}\hat{\sigma}$ and the maximum distance between a sample and its mean is $\leq 2\hat{\sigma}\sqrt{\log(n^2k/\delta)}$ and the algorithm divides into at-least two non-empty clusters such that no two samples from the same distribution are clustered into two clusters.

For the second part observe that by the above concentration on $\mathbf{u}$, no two samples from the same component are clustered differently irrespective of the mean separation. Note that we are using the fact that each sample is clustered at most $2k$ times to get the bound on the error probability. The total error probability by the union bound is $\leq 4\delta$. □

7.6.4 Proof of Lemma 7.11

We show that if the conclusions in Lemmas 7.10 and 7.19 holds, then the lemma is satisfied. We also assume that the conclusions in Lemma 7.25 holds for all the clusters with error probability $\delta' = \delta/k$. By the union bound the total error probability is $\leq 7\delta$.

By Lemma 7.10 all the components within each cluster satisfy

$$\sqrt{w_i}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \leq 25\sigma\sqrt{k^3\log(n^3/\delta)}.$$

Let $n \geq c \cdot dk^9\epsilon^{-4}\log^2 d/\delta$. For notational convenience let

$$S(C) = \frac{1}{|C|}\sum_{i=1}^{|C|}(\mathbf{X}(i) - \overline{\boldsymbol{\mu}}(C))(\mathbf{X}(i) - \overline{\boldsymbol{\mu}}(C))^t - \hat{\sigma}^2\mathbb{I}_d.$$

Therefore by Lemma 7.25 for large enough c,

$$\left|\left|S(C) - \frac{n}{|C|}\sum_{i\in C}\hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))^t\right|\right| \leq \frac{\epsilon^2\sigma^2}{1000k^2}\frac{n}{|C|}.$$

Let $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{k-1}$ be the top eigenvectors of $\frac{1}{|C|}\sum_{i\in C} w_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))^t$. Let

$$\eta_i = \sqrt{\hat{w}_i'}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 = \sqrt{\hat{w}_i}\sqrt{\frac{n}{|C|}}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2\,,$$

and $\boldsymbol{\Delta}_i = \frac{\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))}{||(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))||_2}$. Therefore,

$$\sum_{i\in C}\frac{n}{|C|}\sum_{i\in C}\hat{w}_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C))^t = \sum_{i\in C}\eta_i^2\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^t.$$

Hence by Lemma 7.17, the projection of $\boldsymbol{\Delta}_i$ on the space orthogonal to top $k-1$ eigenvectors of $S(C)$ is at most

$$\sqrt{\frac{\epsilon^2\sigma^2}{1000k^2}\frac{n}{|C|}}\frac{1}{\eta_i} \leq \frac{\epsilon\sigma}{16\sqrt{\hat{w}_i}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2\,k} \leq \frac{\epsilon\sigma}{8\sqrt{2}\sqrt{w_i}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2\,k}.$$

The last inequality follows from the bound on $\hat{w}_i$ in Lemma 7.19.

7.6.5 Proof of Theorem 7.12

We show that the theorem holds if the conclusions in Lemmas 7.11 and 7.21 holds with error probability $\delta' = \delta/k$. Since in the proof of Lemma 7.11, the probability that Lemma 7.10 holds is included, Lemma 7.10 also holds with the same probability. Since there are at most k clusters, by the union bound the total error probability is $\leq 9\delta$.

For every component i, we show that there is a choice of mean vector and weight in the search step such that $w_i D(\mathbf{p}_i, \hat{\mathbf{p}}_i) \leq \epsilon/2k$ and $|w_i - \hat{w}_i| \leq \epsilon/4k$. That would imply that there is a $\hat{\mathbf{f}}$ during the search such that

$$D(\mathbf{f}, \hat{\mathbf{f}}) \leq \sum_C \sum_{i\in C} w_i D(\mathbf{p}_i, \hat{\mathbf{p}}_i) + 2\sum_{i=1}^{k-1}|w_i - \hat{w}_i| \leq \frac{\epsilon}{2k} + \frac{\epsilon}{2k} = \epsilon.$$

Since the weights are gridded by $\epsilon/4k$, there exists a $\hat{w}_i$ such that $|w_i - \hat{w}_i| \leq \epsilon/4k$. We now show that there exists a choice of mean vector such that $w_i D(\mathbf{p}_i, \hat{\mathbf{p}}_i) \leq \epsilon/2k$. Note that if a component has weight $\leq \epsilon/4k$, the above inequality follows immediately. Therefore we only look at those components with $w_i \geq \epsilon/4k$, by Lemma 7.19, for such components $\hat{w}_i \geq \epsilon/5k$ and therefore we only look at clusters

such that $|C| \geq n\epsilon/5k$. By Lemmas 7.15 and for any i,

$$D(\mathbf{p}_i, \hat{\mathbf{p}}_i)^2 \leq 2\sum_{j=1}^{d} \frac{(\hat{\boldsymbol{\mu}}_{i,j} - \hat{\boldsymbol{\mu}}_{i,j})^2}{\sigma^2} + 8d\frac{(\sigma^2 - \hat{\sigma}^2)^2}{\sigma^4}.$$

Note that since we are discarding at most $n\epsilon/8k^2$ random samples at each step. A total number of $\leq n\epsilon/8k$ random samples are discarded. It can be shown that this does not affect our calculations and we ignore it in this proof. By Lemma 7.8, the first estimate of σ^2 satisfies $|\hat{\sigma}^2 - \sigma^2| \leq 2.5\sigma^2\sqrt{\frac{\log n^2/\delta}{d}}$. Hence while searching over values of $\hat{\sigma}^2$, there exist one such that $|\sigma'^2 - \sigma^2| \leq \epsilon\sigma^2/\sqrt{64dk^2}$. Hence,

$$D(\mathbf{p}_i, \hat{\mathbf{p}}_i)^2 \leq 2\frac{||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2^2}{\sigma^2} + \frac{\epsilon^2}{8k^2}.$$

Therefore if we show that there is a mean vector $\hat{\boldsymbol{\mu}}_i$ during the search such that $||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2 \leq \epsilon\sigma/\sqrt{16k^2\hat{w}_i}$, that would prove the Lemma. By triangle inequality,

$$||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2 \leq \left|\left|\overline{\boldsymbol{\mu}}(C) - \hat{\overline{\boldsymbol{\mu}}}(C)\right|\right|_2 + \left|\left|\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C) - (\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}}(C))\right|\right|_2.$$

By Lemma 7.21 for large enough n,

$$\left|\left|\overline{\boldsymbol{\mu}}(C) - \hat{\overline{\boldsymbol{\mu}}}(C)\right|\right|_2 \leq c\sigma\sqrt{\frac{dk\log^2 n^2/\delta}{|C|}} \leq \frac{\epsilon\sigma}{8k\sqrt{w_i}}.$$

The second inequality follows from the bound on n and the fact that $|C| \geq n\hat{w}_i$. Since $w_i \geq \epsilon/4k$, by Lemma 7.19, $\hat{w}_i \geq w_i/2$, we have

$$||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2 \leq \left|\left|\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C) - (\hat{\boldsymbol{\mu}}_i - \hat{\overline{\boldsymbol{\mu}}}(C))\right|\right|_2 + \frac{\epsilon\sigma}{8k\sqrt{w_i}}.$$

Let $\mathbf{u}_1 \ldots \mathbf{u}_{k-1}$ are the top eigenvectors the sample covariance matrix of cluster C. We now prove that during the search, there is a vector of the form $\sum_{j=1}^{k-1} g_j\epsilon_g\hat{\sigma}\mathbf{u}_j$ such that

$$\left|\left|\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C) - \sum_{j=1}^{k-1} g_j\epsilon_g\hat{\sigma}\mathbf{u}_j\right|\right|_2 \leq \frac{\epsilon\sigma}{8k\sqrt{w_i}},$$

during the search, thus proving the lemma. Let $\eta_i = \sqrt{w_i}\,||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2$. By Lemma 7.11, there are set of coefficients α_i such that

$$\frac{\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)}{||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2} = \sum_{j=1}^{k-1} \alpha_j\mathbf{u}_j + \sqrt{1 - ||\alpha||^2}\mathbf{u}',$$

where $\mathbf{u}'$ is perpendicular to $\mathbf{u}_1 \dots \mathbf{u}_{k-1}$ and $\sqrt{1 - ||\alpha||^2} \leq \epsilon\sigma/(8\sqrt{2}\eta_i k)$. Hence, we have

$$\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C) = \sum_{j=1}^{k-1} ||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \, \alpha_j \mathbf{u}_j + ||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \sqrt{1 - ||\alpha||_2^2}\mathbf{u}',$$

Since $w_i \geq \epsilon/4k$ and by Lemma 7.10, $\eta_i \leq 25\sqrt{k^3}\sigma \log(n^3/\delta)$, and $||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2 \leq 100\sqrt{k^4\epsilon^{-1}}\sigma \log(n^3/\delta)$. Therefore $\exists g_j$ such that $|g_j\hat{\sigma} - \alpha_j| \leq \epsilon_g\hat{\sigma}$ on each eigenvector. Hence,

$$\begin{aligned} w_i \left|\left| \boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C) - \sum_{i=1}^{k-1} g_j \epsilon_g \hat{\sigma} \mathbf{u}_j \right|\right|_2^2 &\leq w_i k \epsilon_g^2 \hat{\sigma}^2 + w_i \, ||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2^2 \, (1 - ||\alpha||^2) \\ &\leq k\epsilon_g^2\hat{\sigma}^2 + \eta_i^2 \frac{\epsilon^2\sigma^2}{128\eta_i^2 k^2} \\ &\leq \frac{\epsilon^2\sigma^2}{128k^2} + \frac{\epsilon^2\sigma^2}{128k^2} \leq \frac{\epsilon^2\sigma^2}{64k^2}. \end{aligned}$$

The last inequality follows by Lemma 7.8 and the fact that $\epsilon_g \leq \epsilon/16k^{3/2}$, and hence the theorem. The run time can be easily computed by retracing the steps of the algorithm and using an efficient implementation of single-linkage.

7.7 Proofs for mixtures with unequal variances

In this section, we outline the analysis for the case when the components have different variances.

The main difference would be the coarse clustering algorithm which we describe now. The algorithm repeatedly finds components with smallest variances and clusters samples such that within each cluster the variances differ by a factor of $1 + \widetilde{\mathcal{O}}\big(1/\sqrt{d}\big)$ and the means are close-by. However, two subtleties arise.

Randomized thresholding: Suppose we fix a threshold for clustering in step 3 of the coarse clustering algorithm, then there might be a component whose average distance from $\mathbf{x}(a)$ or $\mathbf{x}(b)$ is exactly the threshold and due to randomness in samples, few samples can lie in one cluster and few can lie on the other. We overcome this, by choosing a random threshold, thus making it unlikely that there is a component with average distance at the threshold.

Components with single sample: If two samples are from the same component i, then their squared-distance concentrates around $2d\boldsymbol{\sigma}_i^2$. We can use this fact to estimate the variance. However if there is only one sample from a component, we cannot estimate its variance and moreover it can affect the calculations of other components. Hence in Step 4, we find such components and discard the corresponding samples.

Generalized coarse clustering: Let $\alpha = 4\sqrt{\log(n^2/\delta)/d}$. Initialize C to the set of all samples. Repeat the following k times.

1. Find threshold $\mathtt{t} = \min_{a \neq b, a,b \in C} ||\mathbf{x}(a) - \mathbf{x}(b)||_2$. Let a and b be the indices that achieve this minimum.
2. Let r be a uniform random variable between 10 and $4000k^2$.
3. Find the set of samples C_1 that are at a distance $\leq \mathtt{t}\sqrt{(1+\alpha \mathrm{r})}$ from either $\mathbf{x}(a)$ or $\mathbf{x}(b)$.
4. If the $\max_{c,d \in C_1} ||\mathbf{x}(c) - \mathbf{x}(d)||_2^2 > \mathtt{t}\sqrt{(1+50\alpha \mathrm{r})}$, discard $\mathbf{x}(a)$, $\mathbf{x}(b)$ and the samples that achieve the maximum, else declare C_1 as a new cluster and remove samples in C_1 from C.

The rest of the analysis is similar to the case with equal variances. We now outline analysis for **Generalized coarse clustering**. We first show an auxiliary concentration inequality that helps us prove the rest of the results.

Lemma 7.28. Given n samples from a set of Gaussian distributions, with probability $\geq 1 - 2\delta$, for every pair of samples $\mathbf{X} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2 \mathbb{I}_d)$ and $\mathbf{Y} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2 \mathbb{I}_d)$,

$$1 - 4\sqrt{\frac{\log \frac{n^2}{\delta}}{d}} \leq \frac{||\mathbf{X} - \mathbf{Y}||_2^2}{d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2} \leq 1 + 4\sqrt{\frac{\log \frac{n^2}{\delta}}{d}}. \tag{7.7}$$

Proof. Since $\mathbf{X}$ and $\mathbf{Y}$ are Gaussians, $\mathbf{X} - \mathbf{Y}$ is distributed $N(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, (\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2)\mathbb{I}_d)$.

Therefore substituting $t = \log \frac{n^2}{\delta}$ in Lemma A.6, with probability $1 - 4\delta/n^2$,

$$||\mathbf{X} - \mathbf{Y}||_2^2 \geq d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) - 2(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2)\sqrt{d \log \frac{n^2}{\delta}} + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 - 2\sqrt{\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2}\,||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 \sqrt{\log \frac{n^2}{\delta}}.$$

and

$$||\mathbf{X} - \mathbf{Y}||_2^2 \leq d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) + 2(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2)\sqrt{d \log \frac{n^2}{\delta}} + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 + 2\sqrt{\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2}\,||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 \sqrt{\log \frac{n^2}{\delta}} + 2(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) \log \frac{n^2}{\delta}.$$

There are $\binom{n}{2}$ pairs and the error probability follows by the union bound. Dividing the bounds by $d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2$ and using the arithmetic-geometric mean inequality we get

$$1 - 3\sqrt{\frac{\log \frac{n^2}{\delta}}{d}} \leq \frac{||\mathbf{X} - \mathbf{Y}||_2^2}{d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2} \leq 1 + 3\sqrt{\frac{\log \frac{n^2}{\delta}}{d}} + 2\frac{\log \frac{n^2}{\delta}}{d}.$$

Using $d \geq 20 \log \frac{n^2}{\delta}$ proves the lemma. □

We now show a few properties of Coarse clustering. In particular, we show that

- There is no mis-clustering.
- After k steps of iteration, all the samples would be clustered.
- The means and variances of all components within any cluster are close to each other.

Let $\alpha \stackrel{\text{def}}{=} 4\sqrt{\frac{\log \frac{n^2}{\delta}}{d}}$. For the rest of the proof we assume that $d \geq 4000 \log(n^2/\delta)$, thus $\alpha \leq 1/10$. We first show that the probability of mis-clustering is $\leq 1/100$.

Lemma 7.29. If Equation (7.7) holds, then after coarse clustering algorithm, with probability $\geq 99/100$, all the samples from each component will be in the same cluster.

Proof. Without loss of generality, let $\mathbf{x}(a)$ be from component 1 and $\mathbf{x}(b)$ be from component 2. If for all components i and $j \in \{1, 2\}$ if

$$\left(d(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_i^2) + \left|\left|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right|\right|_2^2\right)(1+\alpha) < \mathtt{t}^2(1+\alpha\mathrm{r})$$

or

$$\left(d(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_i^2) + \left|\left|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right|\right|_2^2\right)(1-\alpha) > \mathtt{t}^2(1+\alpha\mathrm{r}),$$

then by Equation (7.7) the pairwise distances concentrate and all the samples would be clustered without any error. Hence the error probability is the probability there exists i, j such that $\mathtt{t}^2(1+\alpha\mathrm{r})$ belongs to the set

$$\left[\left(d(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_i^2) + \left|\left|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right|\right|_2^2\right)(1-\alpha), \left(d(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_i^2) + \left|\left|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right|\right|_2^2\right)(1+\alpha)\right].$$

For a given i, j, this probability is at most

$$\frac{2}{4000\mathtt{t}^2k^2 - 10}\left(d(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_i^2) + \left|\left|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right|\right|_2^2\right)$$
$$\times \mathbb{1}\left(\left(d(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_i^2) + \left|\left|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right|\right|_2^2\right)(1-\alpha) \le \mathtt{t}^2(1+4000k^2\alpha)\right).$$

Since $d \ge c \cdot k^4 \log \frac{n^2}{\delta}$ for a large enough constant c, we have $1 + 4000k^2\alpha \le 2$. Hence, the above probability is $\le \frac{4}{3990(1-\alpha)k^2} \le \frac{1}{997(1-\alpha)k^2}$. Since $\alpha \le 1/10$, this is $\le \frac{1}{200k^2}$. By the union bound over all possible components i, j, the error probability is $\le \frac{1}{100k}$. Since we run the algorithm k times, by the union bound the total error probability is $\le \frac{1}{100}$. □

Lemma 7.30. If Equation (7.7) holds and there is no mis-clustering, and a cluster is created at any of the k steps , then for each pair of components i, j in that cluster with $\hat{w}_i, \hat{w}_j \ge 2/n$, $2d\sigma_i^2 \in [\mathtt{t}^2(1-\alpha), \mathtt{t}^2(1+56\alpha\mathrm{r})]$ and $\left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2 \le c \cdot k^2\mathtt{t}^2\alpha$ for some constant c. Furthermore, for every other component l, $||\boldsymbol{\mu}_i - \boldsymbol{\mu}_l||_2^2 + \sigma_i^2 \le c \cdot \mathtt{t}^2$.

Proof. The square of the maximum separation between any two samples in a cluster is $\le \mathtt{t}^2(1+50\alpha\mathrm{r})$ and the points are clustered correctly. Let i be a component such that $\hat{w}_i \ge 2/n$. Let $\mathbf{x}(g)$ and $\mathbf{x}(h)$ be two samples from component i, then

$$\begin{aligned} 2d\boldsymbol{\sigma}_i^2(1-\alpha) &\le ||\mathbf{x}(g) - \mathbf{x}(h)||_2^2 \\ &\le \mathtt{t}^2(1+50\alpha\mathrm{r}), \end{aligned}$$

where the first inequality follows from Equation (7.7). Hence, $2d\boldsymbol{\sigma}_i^2 \leq \mathtt{t}^2(1 + 50\alpha\mathrm{r})/(1-\alpha) \leq \mathtt{t}^2(1+56\alpha\mathrm{r})$. Furthermore, since $\mathbf{x}(g)$ and $\mathbf{x}(h)$ has pairwise distance $\geq \mathtt{t}$, by Equation (7.7),

$$2d\boldsymbol{\sigma}_i^2(1+\alpha) \geq ||\mathbf{x}(g) - \mathbf{x}(h)||_2^2 \geq \mathtt{t}^2,$$

and hence $2d\boldsymbol{\sigma}_i^2 \geq \mathtt{t}^2(1-\alpha)$.

For two samples $\mathbf{x}(g)$ and $\mathbf{x}(h)$ generated by components i and j, we have,

$$\begin{aligned}
\mathtt{t}^2(1+50\alpha\mathrm{r}) &\overset{(a)}{\geq} ||\mathbf{x}(g) - \mathbf{x}(h)||_2^2 \\
&\overset{(b)}{\geq} \left(d(\boldsymbol{\sigma}_i^2 + \boldsymbol{\sigma}_j^2) + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2\right)(1-\alpha) \\
&\overset{(c)}{\geq} \mathtt{t}^2(1-2\alpha) + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2(1-\alpha),
\end{aligned}$$

where (a) follows from the fact that the maximum separation between two samples is $\leq \mathtt{t}^2(1+50\alpha\mathrm{r})$, Equation (7.7) implies (b), and (c) follows from first part of the lemma. Hence, we have $\left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2 \leq \mathtt{t}^2(50\alpha\mathrm{r} + 2\alpha)/(1-\alpha) \leq \mathtt{t}^2(3\cdot 10^6\alpha k^2)$.

Let $\mathbf{x}(g)$ and $\mathbf{x}(h)$ be from components i and l respectively. Similar to the first two parts of the lemma we have, maximum separation between any two samples is

$$\begin{aligned}
\mathtt{t}^2(1+50\alpha\mathrm{r}) &\geq ||\mathbf{x}(g) - \mathbf{x}(h)||_2^2 \\
&\geq \left(d(\boldsymbol{\sigma}_i^2 + \boldsymbol{\sigma}_l^2) + ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_l||_2^2\right)(1-\alpha) \\
&\geq \left(d\boldsymbol{\sigma}_l^2 + ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_l||_2^2\right)(1-\alpha).
\end{aligned}$$

Hence $d\boldsymbol{\sigma}_l^2 + ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_l||_2^2 \leq \mathtt{t}^2(1+50\alpha\mathrm{r})/(1-\alpha) \leq c\cdot\mathtt{t}^2$, for some constant c. The last part follows from the assumption that $d = \Omega(k^4 \log \frac{n^2}{\delta})$. □

Lemma 7.31. If Equation (7.7) holds and there is no mis-clustering, at end of the generalized coarse clustering $|C| = 0$.

Proof. We show that if C is non-empty, at each iteration the number of components in C decreases by at least one. Since there is no mis-clustering, if we create a cluster at a particular iteration, it would contain all the samples from at least one component and hence the number of components in C reduces by one. We

now show that if we discard four samples, at least one of them would be a unique sample from its component ($\hat{w}_i = 1/n$) and hence discarding it would reduce the number of components by one.

Let $\mathbf{x}(a), \mathbf{x}(b)$ be the two samples that attain the minimum and without loss of generality let the corresponding components be 1 and 2. Let $\mathbf{x}(c), \mathbf{x}(d)$ be the two samples that achieve the maximum and i, j be their corresponding components. We now show that if $\min(\hat{w}_1, \hat{w}_2, \hat{w}_i, \hat{w}_j) \geq 2/n$, then the samples would not be discarded thus proving our claim. By Equation (7.7),

$$d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2) + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 \leq \frac{||\mathbf{x}(a) - \mathbf{x}(b)||_2^2}{1-\alpha} \leq \mathtt{t}^2(1+3\alpha),$$

and since two samples from component 1 or 2 did not achieve the minimum, $2d\boldsymbol{\sigma}_2^2 \geq \mathtt{t}^2(1-\alpha)$ and $2d\boldsymbol{\sigma}_1^2 \geq \mathtt{t}^2(1-\alpha)$. Rearranging and substituting in the three equations we get, $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 \leq 4\mathtt{t}^2\alpha$, $2d\boldsymbol{\sigma}_1^2 \leq \mathtt{t}^2(1+7\alpha)$ and $2d\boldsymbol{\sigma}_2^2 \leq \mathtt{t}^2(1+7\alpha)$. Without loss of generality, let $\mathbf{x}(c)$ be included in C_1 because $\mathbf{x}(c)$ was close to $\mathbf{x}(a)$.

$$d(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_i^2) + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_i||_2^2 \leq \frac{||\mathbf{x}(a) - \mathbf{x}(c)||_2^2}{1-\alpha} \leq \mathtt{t}^2(1+3\alpha\mathrm{r}),$$

and furthermore two samples from components i or 1 did not achieve minimum and hence, $2d\boldsymbol{\sigma}_i^2 \geq \mathtt{t}^2(1-\alpha)$ and $2d\boldsymbol{\sigma}_1^2 \geq \mathtt{t}^2(1-\alpha)$. Solving, we get $2d\boldsymbol{\sigma}_i^2 \leq \mathtt{t}^2(1+7\alpha\mathrm{r})$ and $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_i||_2^2 \leq 4\mathtt{t}^2\alpha\mathrm{r}$. Similarly, $2d\boldsymbol{\sigma}_j^2 \leq \mathtt{t}^2(1+7\alpha\mathrm{r})$ and $\left|\left|\boldsymbol{\mu}_l - \boldsymbol{\mu}_j\right|\right|_2^2 \leq 4\mathtt{t}^2\alpha\mathrm{r}$, for some $l \in \{1,2\}$. We now have all the inequalities necessary to show that $||\mathbf{x}(c) - \mathbf{x}(d)||_2^2 \leq \mathtt{t}^2(1+50\alpha\mathrm{r})$ and hence would not be discarded.

$$\begin{aligned}
||\mathbf{x}(c) - \mathbf{x}(d)||_2^2 &\overset{(a)}{\leq} \left(d(\boldsymbol{\sigma}_i^2 + \boldsymbol{\sigma}_j^2) + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2\right)(1+\alpha) \\
&\overset{(b)}{\leq} \mathtt{t}^2(1+7\alpha\mathrm{r}) + \left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_l + \boldsymbol{\mu}_l - \boldsymbol{\mu}_j\right|\right|_2^2 \\
&\overset{(c)}{\leq} \mathtt{t}^2(1+7\alpha\mathrm{r}) + 3(||\boldsymbol{\mu}_i - \boldsymbol{\mu}_1||_2^2 + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_l||_2^2 + \left|\left|\boldsymbol{\mu}_l - \boldsymbol{\mu}_j\right|\right|_2^2) \\
&\overset{(d)}{\leq} \mathtt{t}^2(1+50\alpha\mathrm{r}).
\end{aligned}$$

(a) follows from Equation (7.7). (b) follows from the bounds on $\boldsymbol{\sigma}_i^2$ and $\boldsymbol{\sigma}_j^2$. (c) follows from Cauchy-Schwarz inequality and (d) the bounds on the difference of means which we have shown implies (d). □

The above four lemmas immediately yields,

Lemma 7.32. After coarse clustering, the algorithm divides the samples into clusters such that with probability $\geq 99/100 - 2\delta$,

- There is no mis-clustering.
- For any pair of components i, j within a cluster with $\hat{w}_i, \hat{w}_j \geq 2/n$, the variances lie within a factor of $1 \pm 56\alpha r$ around $\mathtt{t}^2$ and $\left|\left|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right|\right|_2^2 \leq \mathcal{O}(\mathtt{t}^2\alpha^2)$.
- For every component i within a cluster C, $||\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}(C)||_2^2 + d\boldsymbol{\sigma}_i^2 \leq \mathcal{O}(\mathtt{t}^2)$.

It can be shown that once the conclusions in Lemma 7.32 holds, then the performance of recursive clustering algorithm would be same as Lemma 7.10 up to constants. The only modification is the computation of $\hat{\sigma}^2(C)$ which is given by

$$\hat{\sigma}^2(C) = \frac{1}{|C|(|C|-1)} \sum_{a,b \in C} \frac{1}{2d} \left|\left|\mathbf{x}(a) - \mathbf{x}(b)\right|\right|_2^2.$$

By Lemma 7.32, out of $\binom{|C|}{2}$ pairs at most $k|C|$ would have distances away from $\mathtt{t}^2$. It can be shown that this does not affect the analysis.

Finally for the exhaustive search, instead of just substituting a single σ', we try out all possible combinations of $\sigma'(C)$ for each cluster C, where

$$\sigma'(C) \in \hat{\sigma}^2(1 + i\epsilon/d\sqrt{128dk^2}), \forall -L' < i \leq L'\},$$

where $L' = \frac{32k\sqrt{\log n^2/\delta}}{\epsilon}$. Note that since we are searching over k different variances instead of just one, the number of candidate mixtures increases by and hence the time complexity. The time complexity for unequal variances can be shown to be

$$\mathcal{O}\left(n^2 d \log n + d\Big(\frac{k^7}{\epsilon^3}\log^2\frac{d}{\delta}\Big)^{\frac{k^2}{2}}\Big(\frac{k\sqrt{\log d/\delta}}{\epsilon}\Big)^k\right).$$

Note that even though our error probability is $1/100 + 2\delta$, and is not arbitrarily close to 0, we can repeat the entire algorithm $\mathcal{O}(\log\frac{1}{\delta'})$ times and run SCHEFFE on the resulting components to find the closest one. By the Chernoff bound, the error probability of this new estimator would be $\leq \delta'$.

7.8 Lower bounds

We first show a lower bound for a single Gaussian distribution and generalize it to mixtures.

7.8.1 Single Gaussian distribution

The proof is an application of the following version of Fano's inequality (see Lemma 2.15), which states that we cannot simultaneously estimate all distributions in a class using n samples if they satisfy certain conditions.

We consider $d-$dimensional spherical Gaussians with identity covariance matrix, with means along any coordinate restricted to $\pm\frac{c\epsilon}{\sqrt{d}}$. The KL divergence between two spherical Gaussians with identity covariance matrix is the squared distance between their means. Therefore, any two distributions we consider have KL distance at most

$$\beta = \sum_{i=1}^{d} \left(2\frac{c\epsilon}{\sqrt{d}}\right)^2 = 4c^2\epsilon^2,$$

We now consider a subset of these 2^d distributions to obtain a lower bound on α. By the Gilbert-Varshamov bound, there exists a binary code with $\geq 2^{d/8}$ codewords of length d and minimum distance $d/8$. Consider one such code. Now for each codeword, map $1 \to \frac{c\epsilon}{\sqrt{d}}$ and $0 \to -\frac{c\epsilon}{\sqrt{d}}$ to obtain a distribution in our class. We consider this subset of $\geq 2^{d/8}$ distributions as our f_i's.

Consider any two f_i's. Their means differ in at least $d/8$ coordinates. We show that the ℓ_1 distance between them is $\geq c\epsilon/4$. Without loss of generality, let the means differ in the first $d/8$ coordinates, and furthermore, one of the distributions has means $c\epsilon/\sqrt{d}$ and the other has $-c\epsilon/\sqrt{d}$ in the first $d/8$ coordinates. The sum of the first $d/8$ coordinates is $N(c\epsilon\sqrt{d}/8, d/8)$ and $N(-c\epsilon\sqrt{d}/8, d/8)$. The ℓ_1 distance between these normal random variables is a lower bound on the ℓ_1 distance of the original random variables. For small values of $c\epsilon$ the distance between the two Gaussians is at least $\geq c\epsilon/4$. This serves as our α.

Applying the Fano's Inequality, the ℓ_1 error on the worst distribution is at

least

$$\frac{c\epsilon}{8}\Big(1 - \frac{n4c^2\epsilon^2 + \log 2}{d/8}\Big),$$

which for $c = 16$ and $n < \frac{d}{2^{14}\epsilon^2}$ is at least ϵ. In other words, the smallest n to approximate all spherical normal distributions to ℓ_1 distance at most ϵ is $> \frac{d}{2^{14}\epsilon^2}$.

7.8.2 Mixtures of k Gaussians

We now provide a lower bound on the sample complexity of learning mixtures of k Gaussians in d dimensions. We extend the construction for learning a single spherical Gaussian to mixtures of k Gaussians and show a lower bound of $\Omega(kd/\epsilon^2)$ samples. We will again use Fano's inequality over a class of $2^{kd/64}$ distributions as described next.

To prove the lower bound on the sample complexity of learning spherical Gaussians, we designed a class of $2^{d/8}$ distributions around the origin. Let $\mathcal{P} \stackrel{\text{def}}{=} \{\mathbf{p}_1, \ldots, \mathbf{p}_T\}$, where $T = 2^{d/8}$, be this class. Recall that each $\mathbf{p}_i$ is a spherical Gaussian with unit variance. For a distribution $\mathbf{p}$ over $\mathbb{R}^d$ and $\boldsymbol{\mu} \in \mathbb{R}^d$, let $\mathbf{p} + \boldsymbol{\mu}$ be the distribution $\mathbf{p}$ shifted by $\boldsymbol{\mu}$.

We now choose $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$'s *extremely well-separated*. The class of distributions we consider will be a mixture of k components, where the jth component is a distribution from $\mathcal{P}$ shifted by $\boldsymbol{\mu}_j$. Since the $\boldsymbol{\mu}$'s will be well separated, we will use the results from last section over each component.

For $i \in [T]$, and $j \in [k]$, $\mathbf{p}_{ij} \stackrel{\text{def}}{=} \mathbf{p}_i + \boldsymbol{\mu}_j$. Each $(i_1, \ldots, i_k) \in [T]^k$ corresponds to the mixture

$$\frac{1}{k}(\mathbf{p}_{i_1 1} + \mathbf{p}_{i_2 2} + \ldots + \mathbf{p}_{i_k k})$$

of k spherical Gaussians. We consider this class of $T^k = 2^{kd/8}$ distributions. By the Gilbert-Varshamov bound, for any $T \geq 2$, there is a T-ary codes of length k, with minimum distance $\geq k/8$ and number of codewords $\geq 2^{k/8}$. This implies that among the $T^k = 2^{dk/8}$ distributions, there are $2^{kd/64}$ distributions such that any two tuples $(i_1, \ldots, i_k)$ and $(i'_1, \ldots, i'_k)$ corresponding to different distributions differ in at least $k/8$ locations.

If we choose the $\boldsymbol{\mu}$'s well separated, the components of any mixture distribution have very little overlap. For simplicity, we choose $\boldsymbol{\mu}_j$'s satisfying

$$\min_{j_1 \neq j_2} ||\boldsymbol{\mu}_{j_1} - \boldsymbol{\mu}_{j_2}||_2 \geq \left(\frac{2kd}{\epsilon}\right)^{100}.$$

This implies that for $j \neq l$, $||\mathbf{p}_{ij} - \mathbf{p}_{i'l}||_1 < (\epsilon/2dk)^{10}$. Therefore, for two different mixture distributions,

$$\begin{aligned}
&\left|\left|\frac{1}{k}(\mathbf{p}_{i_1 1} + \mathbf{p}_{i_2 2} + \ldots + \mathbf{p}_{i_k k}) - \frac{1}{k}(\mathbf{p}_{i'_1 1} + \mathbf{p}_{i'_2 2} + \ldots + \mathbf{p}_{i'_k k})\right|\right|_1 \\
&\overset{(a)}{\geq} \frac{1}{k} \sum_{j \in [k], i_j, i'_j \in [T]} |\mathbf{p}_{i_j j} - \mathbf{p}_{i'_j j}| - k^2(\epsilon/2dk)^{10} \\
&\overset{(b)}{\geq} \frac{1}{8}\frac{c\epsilon}{4} - k^2(\epsilon/2dk)^{10}.
\end{aligned}$$

where (a) follows form the fact that two mixtures have overlap only in the corresponding components, (b) uses the fact that at least in $k/8$ components $i_j \neq i'_j$, and then uses the lower bound from the previous section.

Therefore, the ℓ_1 distance between any two of the $2^{kd/64}$ distributions is $\geq c_1\epsilon/32$ for c_1 slightly smaller than c. We take this as α.

Now, to upper bound the KL divergence, we simply use the convexity, namely for any distributions $\mathbf{p}_1 \ldots \mathbf{p}_k$ and $\mathbf{q}_1 \ldots \mathbf{q}_k$, let $\bar{\mathbf{p}}$ and $\bar{\mathbf{q}}$ be the mean distributions. Then,

$$D(\bar{\mathbf{p}}||\bar{\mathbf{q}}) \leq \frac{1}{k}\sum_{i=1}^{k} D(\mathbf{p}_i||\mathbf{q}_i).$$

By the construction and from the previous section, for any j,

$$D(\mathbf{p}_{i_j j}||\mathbf{p}_{i'_j j}) = D(\mathbf{p}_i||\mathbf{p}_{i'}) \leq 4c^2\epsilon^2.$$

Therefore, we can take $\beta = 4c^2\epsilon^2$.

Therefore by the Fano's inequality, the ℓ_1 error on the worst distribution is at least

$$\frac{c_1\epsilon}{64}\Big(1 - \frac{n4c^2\epsilon^2 + \log 2}{dk/64}\Big),$$

which for $c_1 = 128, c = 128.1$ and $n < \frac{dk}{8^8\epsilon^2}$ is at least ϵ.

Acknowledgement

Chapter 7 is adapted from Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Near-optimal-sample estimators for spherical Gaussian mixtures", *Neural Information Processing Systems (NIPS),* 2014 [118].

Appendix A

Concentration inequalities

We use the following concentration inequalities for Poisson, Gaussian, Chi-Square, and sum of Bernoulli random variables in the dissertation.

Lemma A.1 (Poisson Chernoff bound)**.** If $X \sim \text{poi}(\lambda)$, then for $x \geq \lambda$,

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

and for $x < \lambda$,

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

Lemma A.2 (Variation of Bernstein's Inequality)**.** Let $X_1, X_2, \ldots X_n$ be n independent zero mean random variables such that with probability $\geq 1-\epsilon_i$, $|X_i| < M$, then

$$\Pr(|\sum_i X_i| \geq t) \leq 2\exp\left(-\frac{t^2}{\sum_i \mathbb{E}[X_i^2] + Mt/3}\right) + \sum_{i=1}^{n} \epsilon_i.$$

If $t = \sqrt{2\left(\sum_i \mathbb{E}[X_i^2]\right)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}$, then

$$\Pr\left(\left|\sum_i X_i\right| \geq \sqrt{2\left(\sum_i \mathbb{E}[X_i^2]\right)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}\right) \leq 2\delta + \sum_{i=1}^{n}\epsilon_i.$$

To prove the concentration of estimators, we bound the variance and show that with high probability the absolute value of each X_i is bounded by M and use Bernstein's inequality with $t = \sqrt{2\left(\sum_i \mathbb{E}[X_i^2]\right)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}$. For example,

Lemma A.3 (Binomial Chernoff bound)**.** If $X_1, X_2 \dots X_n$ are distributed according to Bernoulli p, then with probability $1 - \delta$,

$$\left| \frac{\sum_{i=1}^{n} X_i}{n} - p \right| \leq \sqrt{\frac{2p(1-p)}{n} \log \frac{2}{\delta}} + \frac{2}{3} \frac{\log \frac{2}{\delta}}{n}.$$

Lemma A.4 (Gaussian tail bound)**.** For a Gaussian random variable X with mean μ and variance σ^2,

$$\Pr(|X - \mu| \geq t\sigma) \leq e^{-t^2/2}.$$

Lemma A.5 ([119])**.** If $Y_1, Y_2, \dots Y_n$ be n *i.i.d.* Gaussian variables with mean 0 and variance σ^2, then

$$\Pr\left(\sum_{i=1}^{n} Y_i^2 - n\sigma^2 \geq 2(\sqrt{nt} + t)\sigma^2 \right) \leq e^{-t},$$

and

$$\Pr\left(\sum_{i=1}^{n} Y_i^2 - n\sigma^2 \leq -2\sqrt{nt}\sigma^2 \right) \leq e^{-t}.$$

Furthermore for a fixed vector $\mathbf{a}$,

$$\Pr\left(\left| \sum_{i=1}^{n} \mathbf{a}_i (Y_i^2 - 1) \right| \leq 2(||\mathbf{a}||_2 \sqrt{t} + ||\mathbf{a}||_\infty t)\sigma^2 \right) \leq 2e^{-t}.$$

A simple combination of the above two results proves the following.

Lemma A.6. If $\mathbf{X}$ is distributed according to $N(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_d)$ then,

$$\Pr\left(-2\sqrt{dt}\sigma^2 - 2\,||\boldsymbol{\mu}||_2\, t\sigma \geq ||\mathbf{X}||_2^2 - ||\boldsymbol{\mu}||_2^2 - d\sigma^2 \geq 2(\sqrt{dt} + t)\sigma^2 + 2\,||\boldsymbol{\mu}||_2\, t\sigma \right)$$

is at most $2e^{-t} + e^{-t^2/2}$.

We now state a non-asymptotic concentration inequality for random matrices that helps us bound errors in spectral algorithms.

Lemma A.7 ([114] Remark 5.51)**.** Let $\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n)$ be generated according to $N(0, \Sigma)$. For every $\epsilon \in (0, 1)$ and $t \geq 1$, if $n \geq c' d \left(\frac{t}{\epsilon}\right)^2$ for some constant c', then with probability $\geq 1 - 2e^{-t^2 n}$,

$$\left|\left| \sum_{i=1}^{n} \frac{1}{n} \mathbf{y}(i)\mathbf{y}^t(i) - \Sigma \right|\right| \leq \epsilon\, ||\Sigma||\,.$$

Bibliography

[1] Irving J. Good and Good H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[2] Gina Kolata. Shakespeare's new poem: An ode to statistics. *Science (New York, NY)*, 231(4736):335, 1986.

[3] Stanford. Stanford statistics department brochure, 1992. https://statistics.stanford.edu/sites/default/files/1992_StanfordStatisticsBrochure.pdf.

[4] Eric P. Xing, Michael I. Jordan, and Richard M. Karp. Feature selection for high-dimensional genomic microarray data. In *Proceedings of the 18th Annual International Conference on Machine Learning (ICML)*, pages 601–608, 2001.

[5] Inderjit S. Dhillon, Yuqiang Guan, and Jacob Kogan. Iterative clustering of high dimensional text data augmented by local search. In *Proceedings of the 2002Industrial Conference on Data Mining (ICDM)*, pages 131–138, 2002.

[6] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[7] D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley New York, 1985.

[8] Bruce G. Lindsay. *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Conference series in Probability and Statistics, Penn. State University, 1995.

[9] William A. Gale and Geoffrey Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[10] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 1996 Association for Computational Linguistics (ACL)*, pages 310–318, 1996.

[11] Liam Paninski. Variational minimax estimation of discrete distributions under KL loss. In *Proceedings of the 18th Annual Conference on Neural Information Processing (NIPS)*, pages 1033–1040, 2004.

[12] Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38, 1994.

[13] Fredrick Jelinek and Robert L. Mercer. Probability distribution estimation from sparse data. *IBM Tech. Disclosure Bull.*, 1984.

[14] Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[15] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[16] Raphail E. Krichevsky. The performance of universal encoding. *Transactions on Information Theory*, 44(1):296–303, January 1998.

[17] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1066–1100, 2015.

[18] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[19] David A. McAllester and Robert E. Schapire. On the convergence rate of Good-Turing estimators. In *Proceedings of the 14th Annual Conference on Learning Theory (COLT)*, pages 1–6, 2000.

[20] Evgeny Drukh and Yishay Mansour. Concentration bounds for unigrams language model. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 170–185, 2004.

[21] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 764–796, 2013.

[22] Alon Orlitsky, Narayana P. Santhanam, and Junan Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *Proceedings of the 44th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 179–188, 2003.

[23] Boris Yakovlevich Ryabko. Twice-universal coding. *Problemy Peredachi Informatsii*, 1984.

[24] Boris Yakovlevich Ryabko. Fast adaptive coding algorithm. *Problemy Peredachi Informatsii*, 26(4):24–37, 1990.

[25] Dominique Bontemps, Stéphane Boucheron, and Elisabeth Gassiat. About adaptive coding on countable alphabets. *IEEE Transactions on Information Theory*, 60(2):808–821, 2014.

[26] Stéphane Boucheron, Elisabeth Gassiat, and Mesrob I. Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *CoRR*, abs/1402.6305, 2014.

[27] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[28] Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, pages 584–653, 2006.

[29] Peter J Bickel, Chris A Klaassen, YA'Acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[30] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[31] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2004.

[32] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 47–68, 2011.

[33] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Theertha Suresh. Competitive classification and closeness testing. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 22.1–22.18, 2012.

[34] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. A competitive test for uniformity of monotone distributions. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 57–65, 2013.

[35] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2014.

[36] Gregory Valiant and Paul Valiant. Instance optimal learning. *CoRR*, abs/1504.05321, 2015.

[37] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good-Turing good. In *Proceedings of the 29th Annual Conference on Neural Information Processing (NIPS)*, pages 2134–2142, 2015.

[38] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

[39] Aaron B. Wagner, Pramod Viswanath, and Sanjeev R. Kulkarni. Strong consistency of the Good-Turing estimator. In *Proceedings of the 2006 IEEE International Symposium on Information Theory (ISIT)*, pages 2526–2530, 2006.

[40] Mesrob I. Ohannessian and Munther A. Dahleh. Rare probability estimation under regularly varying heavy tails. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 21.1–21.24, 2012.

[41] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[42] George G. Lorentz. *Bernstein polynomials*. Chelsea Publishing Company, Incorporated, 1986.

[43] Raphail E. Krichevsky and Victor K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

[44] Vladimir G. Vovk. A game of prediction with expert advice. In *Proceedings of the 9th Annual Conference on Learning Theory (COLT)*, pages 51–60, 1995.

[45] Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[46] Peter D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

[47] Nicolò Cesa-Bianchi and Gábor Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the 13th Annual Conference on Learning Theory (COLT)*, pages 12–18, 1999.

[48] Alon Orlitsky, Narayana P. Santhanam, and Juan Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469– 1481, 2004.

[49] Gil Shamir. A new redundancy bound for universal lossless compression of unknown alphabets. In *Proceedings of the Conference on Information Sciences and Systems (CISS)*, pages 1175–1179, 2004.

[50] Jayadev Acharya, Hirakendu Das, and Alon Orlitsky. Tight bounds on profile redundancy and distinguishability. In *Proceedings of the 26th Annual Conference on Neural Information Processing (NIPS)*, pages 3266–3274, 2012.

[51] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[52] Tugkan Batu. *Testing properties of distributions*. PhD thesis, Cornell University, 2001.

[53] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000.

[54] Ulisses Braga-Neto. Classification and error estimation for discrete data. *Pattern Recognition*, 10(7):446–462, 2009.

[55] Jayadev Acharya, Hirakendu Das, H. Mohimani, Alon Orlitsky, and Shengjun Pan. Exact calculation of pattern probabilities. In *Proceedings of the 2010 IEEE International Symposium on Information Theory (ISIT)*, pages 1498 –1502, 2010.

[56] Lucien LeCam. *Asymptotic methods in statistical decision theory*. Springer series in statistics. Springer, New York, 1986.

[57] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[58] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270, 1984.

[59] Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

[60] John Bunge and M Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[61] Robert K Colwell, Anne Chao, Nicholas J Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L Chazdon, and John T Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

[62] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[63] Ronald Thisted and Bradley Efron. Did shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

[64] Peter J Haas, Jeffrey F Naughton, S Seshadri, and Lynne Stokes. Sampling-based estimation of the number of distinct values of an attribute. In *VLDB*, volume 95, pages 311–322, 1995.

[65] Dinei Florencio and Cormac Herley. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web*, pages 657–666. ACM, 2007.

[66] Ian Kroes, Paul W Lepp, and David A Relman. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, 96(25):14547–14552, 1999.

[67] Bruce J Paster, Susan K Boches, Jamie L Galvin, Rebecca E Ericson, Carol N Lau, Valerie A Levanos, Ashish Sahasrabudhe, and Floyd E Dewhirst. Bacterial diversity in human subgingival plaque. *Journal of bacteriology*, 183(12):3770–3783, 2001.

[68] Jennifer B Hughes, Jessica J Hellmann, Taylor H Ricketts, and Brendan JM Bohannan. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and environmental microbiology*, 67(10):4399–4406, 2001.

[69] Zhan Gao, Chi-hong Tseng, Zhiheng Pei, and Martin J Blaser. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences*, 104(8):2927–2932, 2007.

[70] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of t-cell receptor β-chain diversity in $\alpha\beta$ t cells. *Blood*, 114(19):4099–4107, 2009.

[71] Iuliana Ionita-Laza, Christoph Lange, and Nan M Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13):5008–5013, 2009.

[72] Ronald Aylmer Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

[73] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Wiley-Interscience, New York, NY, 1964.

[74] Herbert E Robbins. Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics*, 39(1):256–257, 1968.

[75] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[76] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 685–694, 2011.

[77] Paul Valiant and Gregory Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.

[78] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *submitted to IEEE Transactions on Information Theory, arxiv:1407.0381*, Jul 2015.

[79] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

[80] T. Tony Cai and Mark G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[81] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[82] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *preprint arxiv:1504.01227*, Apr. 2015.

[83] Anne Chao. Species estimation and applications. *Encyclopedia of statistical sciences*, 2005.

[84] Eric P Smith and Gerald van Belle. Nonparametric estimation of species richness. *Biometrics*, pages 119–129, 1984.

[85] Tsung-Jen Shen, Anne Chao, and Chih-Feng Lin. Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3):798–804, 2003.

[86] United States Census Bureau. Frequently occuring surnames from the Census 2000, 2014. http://www.census.gov/topics/population/genealogy/data/2000_surnames.html.

[87] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Estimating the number of unseen species: A bird in the hand is worth log n in the bush. *arXiv preprint arXiv:1511.07428*, 2015.

[88] Andrew D. Barbour and Peter Hall. On the rate of poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 95:473–480, 5 1984.

[89] J. Michael Steele. An efron-stein inequality for nonsymmetric statistics. *The Annals of Statistics*, 14(2):753–758, 06 1986.

[90] Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, 8(4):745–764, 08 1980.

[91] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2), 1984.

[92] Jinwen Ma, Lei Xu, and Michael I. Jordan. Asymptotic convergence rate of the em algorithm for Gaussian mixtures. *Neural Computation*, 12(12), 2001.

[93] Sanjoy Dasgupta and Leonard J. Schulman. A two-round variant of EM for Gaussian mixtures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 152–159, 2000.

[94] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th Innovations in Theoretical Computer Science Conference (ITCS)*, pages 11–20, 2013.

[95] Martin Azizyan, Aarti Singh, and Larry A. Wasserman. Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Proceedings of the 27th Annual Conference on Neural Information Processing (NIPS)*, pages 2139–2147, 2013.

[96] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of Gaussians using the k-means algorithm. *CoRR*, abs/0912.0086, 2009.

[97] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 113–122, 2002.

[98] Sanjoy Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 634–644, 1999.

[99] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 458–469, 2005.

[100] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Proceedings of the 51st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 103–112, 2010.

[101] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 553–562, 2010.

[102] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the 51st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 93–102, 2010.

[103] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 1135–1164, 2014.

[104] Michael J. Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 273–282, 1994.

[105] Alon Orlitsky, Narayana P. Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 426–435, 2004.

[106] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-modal distributions via testing. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, pages 1371–1385, 2012.

[107] Siu On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, pages 1380–1394, 2013.

[108] Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. PAC learning axis-aligned mixtures of Gaussians with no separation assumption. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, pages 20–34, 2006.

[109] Yoav Freund and Yishay Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 13th Annual Conference on Learning Theory (COLT)*, pages 53–62, 1999.

[110] Jon Feldman, Ryan O'Donnell, and Rocco A. Servedio. Learning mixtures of product distributions over discrete domains. In *Proceedings of the 46th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 501–510, 2005.

[111] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 1183–1213, 2014.

[112] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

[113] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitksy, and Ananda Theertha Suresh. Sorting with adversarial comparators and application to density estimation. In *Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT)*, pages 1682–1686, 2014.

[114] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010.

[115] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[116] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[117] Siu On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 2014.

[118] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Near-optimal-sample estimators for spherical Gaussian mixtures. In *Proceedings of the 28th Annual Conference on Neural Information Processing (NIPS)*, pages 1395–1403, 2014.

[119] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):pp. 1302–1338, 2000.